DOCUMENT RESUME

ED 459 812 IR 058 348

TITLE	Proceeding of the ACM/IEEE-CS Joint Conference on Digital Libraries (1st, Roanoke, Virginia, June 24-28, 2001).
INSTITUTION	Association for Computing Machinery, New York, NY.; Institute of Electrical and Electronics Engineers, Inc., Washington, DC. Computer Society.
ISBN	ISBN-1-58113-345-6
PUB DATE	2001-06-00
NOTE	578p.; For individual conference papers, see IR 058 349-386. Some abstracts and papers are copyrighted by the authors and are not available from ERIC. Some figures may not reproduce well. Also sponsored by ACM Special Interest Group on Information Retrieval and Special Interest Group on Hypertext, Hypermedia, and the Web; IEEE-CS Technical Committee on Digital Libraries; Coalition for Networked Information; Lucent Technologies; Sun Microsystems; Virginia Tech; and VTLS.
AVAILABLE FROM	Association for Computing Machinery, 1515 Broadway, New York, NY 10036. Tel: 800-342-6626 (Toll Free); Tel: 212-626-0500; e-mail: acmhelp@acm.org. For full text: http://www1.acm.org/pubs/contents/proceedings/dl/379437/.
PUB TYPE	Collected Works - Proceedings (021)
EDRS PRICE	MF03/PC24 Plus Postage.
DESCRIPTORS	Access to Information; Archives; Computer Oriented Programs; Computer System Design; *Electronic Libraries; Information Services; *Information Technology; Internet; Library Collection Development; *Library Services; Online Systems; Technological Advancement

ABSTRACT

Papers in this Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (Roanoke, Virginia, June 24-28, 2001) discuss: automatic genre analysis; text categorization; automated name authority control; automatic event generation; linked active content; designing e-books for legal research; metadata harvesting; mapping the interoperability landscape for networked information retrieval; distributed resource discovery; enforcing interoperability with the open archives initiative (OAI) repository explorer; an OAI service provider for cross-archive searching; managing change on the Web; reputation of Web sites; personalized spiders; global digital library development; Web-based scholarship; multi-view intelligent editor for digital video libraries; digital libraries environments for scientific thinking; open virtual learning environment; automatic identification and organization of index terms; digital library collaborations; public use of digital community information systems; collaborative design with use case scenarios; automatic keyphrasing; legal deposit of digital publications; technology and values; digital strategy for the Library of Congress; multiple digital libraries; technical support workers; digital library recommendation services; digital libraries and data scholarship; metasearching; President's Information Technology Advisory Committee's February 2001 digital library report; searchable collections of enterprise speech data; transcript-free search of audio archives; audio watermarking techniques; music-notation searching and digital libraries; automatic classification of musical instrument sounds; adding content-based searching to a traditional music library catalog server; locating question difficulty through explorations in question space; browsing by phrases; improving video indexing in a digital library; access to a digital video library; digital library design; bucket architecture for the open video project; personalized search and summarization over multimedia healthcare information; automation and human mediation in libraries; preservation of digital information; trading networks of digital archives; cost-driven design for archival repositories; notification service for digital libraries; automated rating of reviewers; digital libraries supporting digital government; digital libraries for young children; preserving, restoring and analyzing damaged manuscripts; digital music libraries; content management for digital museum exhibitions; hierarchical document clustering of digital library retrieval results; interactive visualization of video metadata; categorizing hidden-Web resources; digital facsimile editions and online editing; and image semantics for picture libraries. (AEF)

Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (1st, Roanoke, Virginia, June 24-28, 2001)

JCDL 2001

following in the tradition of the ACM Digital Libraries & IEEE CS Advances in Digital Libraries Conferences

Sponsored by

The Association for Computing Machinery(ACM)
Special Interest Group on Information Retrieval (ACM SIGIR)
Special Interest Group on Hypertext, Hypermedia, and the Web (ACM SIGWEB)
The Institute for Electrical and Electronics Engineers Computer Society (IEEE Computer Society)
Technical Committee on Digital Libraries (TCDL)
and

The Coalition for Networked Information
Lucent Technologies
Sun Microsystems
Virginia Tech
VTLS

The Joint Conference on Digital Libraries is a major international forum focusing on digital libraries and associated technical, practical, and social issues. JCDL 2001 enhances the tradition of conference excellence already established by the ACM and IEEE-CS by combining the annual events that these professional societies have sponsored on an annual basis, the ACM Digital Libraries Conferences and the IEEE-CS Advances in Digital Libraries Conferences. Also, following JDCL will be an NSF PI meeting for the US Digital Libraries Initiative.

JCDL encompasses the many meanings of the term "digital libraries", including (but not limited to) new forms of information institutions; operational information systems with all manner of digital content; new means of selecting, collecting, organizing, and distributing digital content; and theoretical models of information media, including document genres and electronic publishing.

Digital libraries are distinguished from information retrieval systems because they include more types of media, provide additional functionality and services, and include other stages of the information life cycle, from creation through use. Digital libraries also can be viewed as a new form of information institution or as an extension of the services libraries currently provide.

WIRELESS NETWORK ACCESS FOR FREE AT JCDL!

REGISTRATION DATES

~~May 15~~	**Last Day** *to register at reduced rate*
~~June 15~~	Last Day to register on-line
June 24	Conference Starts - Registration in-place

The intended community for this conference includes those interested in such aspects of digital libraries as infrastructure; institutions; metadata; content; services; digital preservation; system design; implementation; interface design; human-computer interaction; evaluation of performance; evaluation of usability; collection development; intellectual property; privacy; electronic publishing; document genres; multimedia; social, institutional, and policy issues; user communities; and associated theoretical topics.

Participation is sought from all parts of the world and from the full range of disciplines and professions involved in digital library research and practice, including computer science, information science, librarianship, archival science and practice, museum studies and practice, technology, medicine, social sciences, and humanities. All domains - academe, government, industry, and others - are encouraged to participate as presenters or attendees.

The First
ACM – IEEE

Joint Conference on Digital Libraries

Final Program

General Information

Registration: Roanoke Foyer / North Entry
Sunday 7:30-21:00
Monday 7:30-20:00
Tuesday 7:30-17:30
Wednesday 7:30-17:30
Thursday 7:30-12:00

Conference Office: Bent Mountain Room
Sunday – Thursday, 8:00 – 18:00

Email / Net Access: Mill Mountain Room
Sunday – Thursday, 8:00 – 18:00

Help Desk: Roanoke Foyer / North Entry
Sunday – Thursday, 8:00 – 18:00

Complimentary lunches for all registered participants will be served daily in Roanoke Ballroom CDH.
Complimentary continental breakfast, morning and afternoon coffee breaks will be served daily in the Crystal Foyer.

Program Note: The JCDL 2001 program includes both long and short papers. Long papers are assigned a 30 minute period; short papers (indicated by an asterisk preceding the title) a 15 minute period.

Sunday, June 24, 2001 – Tutorials and Side Trips

7:30-9:00
Complimentary Continental Breakfast for Registered Tutorial Participants in Crystal Foyer

8:00-17:00: Tutorials

Practical Digital Libraries Overview: Part 1 & 2 – Buck Mountain Room
Edward A. Fox (Department of Computer Science, Virginia Tech)

Evaluating, Using, and Publishing eBooks: Part 1 & 2 – Harrison/Tyler Room
Gene Golovchinsky (FX Palo Alto Laboratory)
Cathy Marshall (Microsoft)
Elli Mylonas (Scholarly Technology Group, Brown University)

Thesauri and Ontologies in Digital Libraries: Part 1 & 2 – Monroe Room
Dagobert Soergel (College of Information Studies, University of Maryland, College Park)

How to Build a Digital Library Using Open-Source Software – Wilson Room (morning)
Ian H. Witten (Department of Computer Science, University of Waikato)

Hands-On Workshop: Build Your Own Digital Library – Wilson Room (afternoon)
Ian H. Witten (Department of Computer Science, University of Waikato)
David Bainbridge (Department of Computer Science, University of Waikato)

Building Interoperable Digital Libraries: A Practical Guide to Creating Open Archives – Crystal Ballroom E
Hussein Suleman (Department of Computer Science, Virginia Tech)

12:00-13:00
Complimentary Lunch for Registered Tutorial Participants in Roanoke GH

Side Trips

Virginia Wine Tour 9:30-16:00: Bus leaves from Conference Center North Entrance

Bottom Creek Gorge Hike 10:00-16:00: Buses leave from Conference Center North Entrance

Sunday Evening, June 24, 2001 – Main Conference Opening

18:00-20:00
Opening Reception (Roanoke Ballroom AB): Virginia ham and assorted hors d'oeuvres;
music by *No Strings Attached.*
Welcome by Robert C. Bates, Dean of Arts and Science, Virginia Tech.

Monday, June 25, 2001 – Main Conference

7:30-9:00
Complimentary Continental Breakfast in Crystal Foyer
Newcomers' Breakfast in Roanoke CD
Help yourselves to continental breakfast in the foyer and get together with other newcomers and friends.

Sessions 1 and 2: 9:00-10:30

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
	Conference Opening (1): Welcome by Edward A. Fox, Conference Chair, and Christine Borgman, Program Chair **Keynote Address (2):** *Public Access to Digital Materials* Brewster Kahle (President, Alexa Internet; Director, Internet Archive) Introduction by Edward A. Fox	

10:30-11:00
Break. Refreshments in Crystal Foyer.

Session 3: 11:00-12:30

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
Papers (3B): Digital Libraries for Education: Technology, Services, and User Studies **Moderator:** Dale Flecker (Harvard University) *Linked Active Content: A Service for Digital Libraries for Education* David Yaron, D. Jeff Milton, Rebecca Freeland (Carnegie Mellon University) *A Component Repository for Learning Objects: A Progress Report* Jean R. Laleuf, Anne Morgan Spalter (Brown University) *Designing e-Books for Legal Research* Catherine C. Marshall, Morgan N. Price, Gene Golovchinsky, Bill N. Schilit (FX Palo Alto Laboratory) **Honorable Mention, Vannevar Bush Award**	**Panel (3C): The Open Archives Initiative: Perspectives on Metadata Harvesting** **Moderator:** James B. Lloyd (University of Tennessee) **Panelists:** Tim Cole (Chair, Library Information Technology Committee, University of Illinois, Urbana-Champaign) Caroline Arms (Library of Congress) Donald Waters (Mellon Foundation) Simeon Warner (Los Alamos National Laboratory) Jeffrey Young (OCLC)	**Papers (3A): Methods for Classifying and Organizing Content in Digital Libraries** **Moderator:** Jose Luis Borbinha (Biblioteca Nacional, Portugal) *Integrating Automatic Genre Analysis into Digital Libraries* Andreas Rauber, Alexander Mueller-Koegler (Vienna University of Technology) *Text Categorization for Multi-page Documents: A Hybrid Naive Bayes HMM Approach* Paolo Frasconi, Giovanni Soda, Alessandro Vullo (University of Florence) **Automated Name Authority Control* James W. Warner, Elizabeth W. Brown (Johns Hopkins University) **Automatic Event Generation From Multi-lingual News Stories* Kin Hui, Wai Lam, Helen M. Meng (The Chinese University of Hong Kong)

12:30-13:30
Complimentary Lunch for all Registered Attendees in Roanoke CDH.

Monday, Session 4: 13:30-15:00

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
Papers (4A): Approaches to Interoperability Among Digital Libraries **Moderator:** James Frew (University of California, Santa Barbara) **Mapping the Interoperability Landscape for Networked Information Retrieval* William E. Moen (University of North Texas) **Distributed Resource Discovery: Using Z39.50 to Build Cross-Domain Information Servers* Ray R. Larson (University of California, Berkeley) *The Open Archives Initiative* Carl Lagoze, Herbert Van de Sompel (Cornell University) **Enforcing Interoperability with the Open Archives Initiative Repository Explorer* Hussein Suleman (Virginia Tech) **Arc-An OAI Service Provider for Cross-Archive Searching* Xiaoming Liu, Kurt Maly, Mohammad Zubair, Michael L. Nelson (Old Dominion University)	**Panel (4C): Different Cultures Meet – Lessons Learned in Global Digital Library Development** **Moderator:** Ching-chih Chen (Professor, Graduate School of Library and Information Science, Simmons College) **Panelists:** Hsueh-hua Chen (Chair, Department of Library and Information Science, National Taiwan University) Wen Gao (Deputy President, Graduate Schools, Chinese Academy of Sciences, Beijing) Von-Wun Soo (Professor of Computer Science, National Tsinghua University, Taipei) Li-Zhu Zhou (Chair, Department of Computer Science, Tsinghua University, Beijing)	**Papers (4B): Digital Libraries and the Web: Technology and Trust** **Moderator:** Jonathan Furner (University of California, Los Angeles) *Managing Change on the Web* Luis Francisco-Revilla, Frank Shipman, Richard Furuta, Unmil Karadkar, Avital Arora (Texas A&M University) **Measuring the Reputation of Web Sites: A Preliminary Exploration* Greg Keast, Joan Cherry, Elaine G. Toms (University of Toronto) *Personalized Spiders for Web Search and Analysis* Michael Chau, Daniel Zeng, Hsinchun Chen (University of Arizona, Tucson) **Salticus: Guided Crawling for Personal Digital Libraries* Robin Burke (California State University, Fullerton)

15:00-15:30
Break. Refreshments in Crystal Foyer.

Session 5: 15:30-17:00

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
Papers (5A): Tools for Constructing and Using Digital Libraries **Moderator:** Hsinchun Chen (University of Arizona, Tucson) *Power to the People: End-User Building of Digital Library Collections* Ian H. Witten, David Bainbridge, Stefan J. Boddie (University of Waikato) **Web-Based Scholarship: Annotating the Digital Library* Bruce Rosenstock, Michael Gertz (University of California, Davis) *A Multi-View Intelligent Editor for Digital Video Libraries* Brad A. Myers, Juan P. Casares, Scott Stevens, Laura Dabbish, Dan Yocum, Albert Corbett (Carnegie Mellon University) **VideoGraph: A New Tool for Video Mining and Classification* Jia-Yu Pan, Christos Faloutsos (Carnegie Mellon University)	**Panel (5C): Digital Library Collaborations in a World Community** **Moderator:** David Fulker (Unidata Program Center) **Panelists:** Sharon Dawes (SUNY Albany) Leonid Kalinichenko (Institute of Informatics Problems, Russian Academy of Science, Moscow State University) Tamara Sumner (Center for LifeLong Learning and Design, Dept. of Computer Science and the Institute of Cognitive Science, University of Colorado) Constantino Thanos (DELOS Director, Consiglio Nazionale delle Ricerche, Istituto di Elaborazione della Informazione) Alex Ushakov (ChemQuest Project, Department of Chemistry and Biochemistry, University of Northern Colorado)	**Papers (5B): Systems Design and Evaluation for Undergraduate Learning Environments** **Moderator:** Traugott Koch (Lund University & Technical Knowledge Center of Denmark) **The Alexandria Digital Earth Prototype System* Terence R. Smith, Greg Janee, James Frew, Anita Coleman (University of California, Santa Barbara) **Iscapes: Digital Libraries Environments for the Promotion of Scientific Thinking by Undergraduates in Geography* Anne J. Gilliland-Swetland, Gregory H. Leazer (University of California, Los Angeles) **Project ANGEL: An Open Virtual Learning Environment with Sophisticated Access Management* John MacColl (University of Edinburgh) **NBDL: A CIS Framework for NSDL* Joe Futrelle, C. Kevin Chang (University of Illinois, Urbana-Champaign) Su-Shing Chen (University of Missouri, Columbia) *Automatic Identification and Organization of Index Terms for Interactive Browsing* Nina Wacholder, David K. Evans, Judith L. Klavans (Columbia University)

Monday 19:00-21:00

Poster and Demo Session and Reception (Conference Center Foyer); Pasta bar and assorted hors d'oeuvres.

Demonstrations	Posters
Content Management for Digital Museum Exhibitions Jen-Shin Hong, Bai-Hsuen Chen, Jieh Hsiang (National Chi-Nan University) Tien-Yu Shu (National Museum of Natural Science) *Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results* Christopher R. Palmer, Jerome Pesenti, Raul E. Valdes-Perez, Michael G. Christel, Alex G. Hauptmann, Dorbin Ng, and Howard D. Wactlar (Carnegie Mellon University) *Indiana University Digital Music Library Project* Jon W. Dunn, Eric J. Isaacson (Indiana University, Bloomington) *Interactive Visualization of Video Metadata* Mark Derthick (Carnegie Mellon University) *PERSIVAL: Categorizing Hidden-Web Resources* Panagiotis G. Ipeirotis, Luis Gravano (Columbia University) Mehran Sahami (E.piphany, Inc.) *PERSIVAL; Personalized Search and Summarization over Multimedia Health-care Information* Noemie Elhadad, Min-Yen Kan, Simon Lok and Smaranda Muresan (Columbia University) *PERSIVAL: View Segmentation and Static/Dynamic Summary Generation for Echocardiogram Videos* Shahram Ebadollahi, Shih-Fu Chang (Columbia University) *Stanford Encyclopedia of Philosophy: A Dynamic Reference Work* Edward N. Zalta, Uri Nodelman (Stanford University) Colin Allen (Texas A&M University) *A System for Adding Content-Based Searching to a Traditional Music Library Catalogue Server* Matthew J. Dovey (Kings College, London) *Using the Repository Explorer to Achieve OAI Protocol Compliance* Hussein Suleman (Virginia Tech)	*An Atmospheric Visualization Collection for the NSDL* Keith Andrew (Eastern Illinois University) Christopher Klaus (Argonne National Laboratory) Gerald Mace (University of Utah) *Breaking the Metadata Generation Bottleneck: Preliminary Findings* Elizabeth D. Liddy (School of Information Studies, Syracuse University) Stuart Sutton (School of Library & Information Science, University of Washington) Woojin Paik (solutions-united.com, Syracuse, NY) Eileen Allen (School of Information Studies, Syracuse University) Sarah Harwell (solutions-united.com, Syracuse, NY) Michelle Monsour (School of Information Studies, Syracuse University) Anne Turner (School of Library & Information Science, University of Washington) Jennifer Liddy (School of Information Studies, Syracuse University) *Building the Physical Sciences Information Infrastructure, A Phased Approach* Judy C. Gilmore, Valerie S. Allen (U.S. Department of Energy) *Development of an Earth Environmental Digital Library System for Soil and Land-Atmospheric Data* Eiji Ikoma, Taikan Oki, Masaru Kitsuregawa (Institute of Industrial Science, Univ. of Tokyo) *Digital Facsimile Editions and On-Line Editing* Harry Plantinga (Computer Science, Calvin College) *Dspace at MIT: Meeting the Challenges* Michael Bass (Hewlett-Packard) Margret Branschofsky (Faculty Liaison, MIT) *Exploiting Image Semantics for Picture Libraries* Kobus Barbard, David Forsyth (University of California at Berkeley) *Feature Extraction for Content-Based Image Retrieval in DARWIN (Digital Analysis and Recognition of Whale Images on a Network)* Kelly R. Debure, Adam S. Russell (Eckerd College) *Guided Linking: Efficiently Making Image-to-Transcript Correspondence* Cheng Jiun Yuan, W. Brent Seales (University of Kentucky) *Integrating Digital Libraries by CORBA, XML and Servlet* Wing Hang Cheung, Michael R. Lyu, Kam Wing Ng (The Chinese University of Hong Kong) *A National Digital Library for Undergraduate Mathematics and Science Teacher Preparation and Professional Development* Kimberly S. Roempler (Eisenhower National Clearinghouse, The Ohio State University) *Print to Electronic: Measuring the Operational and Economic Implications of an Electronic Journal Collection* Carol Montgomery, Linda Marion (Hagerty Library, Drexel University) *Turbo Recognition: Decoding Page Layout* Taku A. Tokuyasu (University of California at Berkeley) *Using Markov Models and Innovation-Diffusion as a Tool for Predicting Digital Library Access and Distribution Rates* Bruce R. Barkstrom (NASA Langley Research Center) *A Versatile Facsimile and Transcription Service for Manuscripts and Rare Old Books at the Miguel de Cervantes Digital Library* Alejandro Bia (Miguel de Cervantes Digital Library, University of Alicante, Spain) *The Virtual Naval Hospital: The Digital Library as Knowledge Management Tool for Nomadic Patrons* Michael P. D'Alessandro, Donna M. D'Alessandro, Mary J. C. Hendrix (University of Iowa) Richard S. Bakalar (Naval Medical Information Management Center) Denis E. Ashley (US Navy Bureau of Medicine and Surgery)

Tuesday, June 26, 2001 – Main Conference

7:30-9:00
Complimentary Continental Breakfast in Crystal Foyer

Session 6: 9:00-10:30

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
Panel (6C): A Digital Strategy for the Library of Congress: Discussion of the LC21 Report and the Role of the Digital Library Community **Moderator:** Alan Inouye (Computer Science and Telecommunications Board, National Academy of Sciences) **Panelists:** Dale Flecker (Associate Director of Planning and Systems, Harvard University Library) Margaret Hedstrom (Associate Professor, School of Information, University of Michigan.) David Levy, Professor (Information School, University of Washington)	**Papers (6A): Studying the Users of Digital Libraries: Formative and Summative Evaluations** **Moderator:** John Ober (California Digital Library) *Public Use of Digital Community Information Systems: Findings from a Recent Study with Implications for System Design* Karen E. Pettigrew (University of Washington, Seattle) Joan C. Durrance (University of Michigan, Ann Arbor) **Evaluating the Distributed National Electronic Resource* Peter Brophy, Shelagh Fisher (The Manchester Metropolitan University) **Collaborative Design with Use Case Scenarios* Lynne Davis (University Corporation for Atmospheric Research) Melissa Dawe (University of Colorado, Boulder) *Human Evaluation of Kea, an Automatic Keyphrasing System* Steve Jones, Gordon W. Paynter (University of Waikato)	**Papers (6B): Digital Library Collections: Policies and Practices** **Moderator:** William Arms (Cornell University) *Community Design of DLESE's Collections Review Policy: A Technological Frames Analysis* Michael Khoo (University of Colorado, Boulder) *Legal Deposit of Digital Publications: A Review of Research and Development Activity* Adrienne Muir (Loughborough University) **Comprehensive Access to Printed Materials (CAPM)* G. Sayeed Choudhury, Mark Lorie, Erin Fitzpatrick, Ben Hobbs, Greg Chirikjian, Allison Okamura (Johns Hopkins University) Nick Flores (University of Colorado, Boulder) **Technology and Values: Lessons from Central and Eastern Europe* Nadia Caidi (University of Toronto)

10:30-11:00
Break. Refreshments in Crystal Foyer.

Session 7: 11:00-12:30

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
Panel (7C): The President's Information Technology Advisory Committee's February 2001 Digital Libraries Report and Its Impact **Moderator:** Sally E. Howe (National Coordination Office for Information Technology Research & Development) **Panelists:** David Nagel (President AT&T Labs, Chair PITAC Digital Library Panel) Ching-chih Chen (Professor, Graduate School of Library and Information Science, Simmons College, and PITAC) Stephen M. Griffin (NSF, Digital Libraries Initiative) *[CONTINUED ON NEXT PAGE]*	**Papers (7A): Studying the Users of Digital Libraries: Qualitative Approaches** **Moderator:** Cliff McKnight (Loughborough University, United Kingdom) *Use of Multiple Digital Libraries: A Case Study* Ann Blandford, Hanna Stelmaszewska (Middlesex University) Nick Bryan-Kinns (Icon MediaLab London) *An Ethnographic Study of Technical Support Workers: Why We Didn't Build a Tech Support Digital Library* Sally Jo Cunningham, Chris Knowles (University of Waikato) Nina Reeves (Cheltenham and Cloucestershire College of Higher Education) *[CONTINUED ON NEXT PAGE]*	**Papers (7B): Techniques for Managing Distributed Collections** **Moderator:** Ray Larson (University of California, Berkeley) **Overview of the Virtual Data Center Project and Software* Micah Altman, L. Andreev, M. Diggory, G. King, E. Kolster, A. Sone, S. Verba (Harvard University) D.L. Kiskis, M. Krott (University of Michigan, Ann Arbor) **Digital Libraries and Data Scholarship* Bruce R. Barkstrom (NASA Langley Research Center) *[CONTINUED ON NEXT PAGE]*

Tuesday, Session 7: 11:00-12:30

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
[CONTINUED]	*[CONTINUED]*	*[CONTINUED]*
James H. Lightbourne (National SMETE Digital Library (NSDL) Program, National Science Foundation) Walter L. Warnick (Department of Energy, Director Office of Scientific and Technical Information)	**Developing Recommendation Services for a Digital Library with Uncertain and Changing Data* Gary Geisler (University of North Carolina, Chapel Hill) David McArthur, Sarah Giersch (Eduprise) **Evaluation of DEFINDER: A System to Mine Definitions from Consumer-Oriented Medical Text* Judith L. Klavans, Smaranda Muresan (Columbia University)	*SDLIP + STARTS = SDARTS: A Protocol and Toolkit for Metasearching* Noah Green, Panagiotis G. Ipeirotis, Luis Gravano (Columbia University) *Database Selection for Processing k Nearest Neighbors Queries in Distributed Environments* Clement Yu, Prasoon Sharma, Yan Qin (University of Illinois, Chicago) Weiyi Meng (State University of New York, Binghamton)

12:30-13:30
Complimentary Lunch for all Registered Attendees in Roanoke CDH.

Session 8: 13:30-15:30

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
Panel (8C): The National SMETE Digital Library Program **Moderator:** Brandon Muramatsu (SMETE.ORG Project Director, University of California at Berkeley) **Panelists:** James H. Lightbourne (National SMETE Digital Library (NSDL) Program, National Science Foundation) Cathryn A. Manduca (University Corporation for Atmospheric Research) Marcia Mardis (Merit Networks, TeacherLib NSDL Project) Flora P. McMartin (University of Califomia at Berkeley)	**Papers (8A): The Sound of Digital Libraries: Audio, Music, and Speech** **Moderator:** Edie Rasmussen (University of Pittsburgh) *Building Searchable Collections of Enterprise Speech Data* James W. Cooper, Mahesh Viswanathan (IBM T J Watson Research Center) Donna Byron (University of Rochester) Margaret Chan (Columbia University) **Transcript-Free Search of Audio Archives at the National Gallery of the Spoken Word* J.H.L. Hansen (University of Colorado, Boulder) J.R. Deller, Jr., M.S. Seadle (Michigan State University, East Lansing) **Audio Watermarking Techniques for the National Gallery of the Spoken Word* J.R. Deller, Jr., A. Gurijala, M.S. Seadle (Michigan State University, East Lansing) *Music-Notation Searching and Digital Libraries* Donald Byrd (University of Massachusetts, Amherst) **Feature Selection for Automatic Classification of Musical Instrument Sounds* Mingchun Liu, Chunru Wan (Nanyang Technological University) **Adding Content-Based Searching to a Traditional Music Library Catalogue Server* Matthew J. Dovey (Kings College, London)	**Papers (8B): Information Search and Retrieval in Digital Libraries** **Moderator:** Nicholas Belkin (Rutgers University) **Locating Question Difficulty through Explorations in Question Space* Terry Sullivan (University of North Texas) **Browsing by Phrases: Terminological Information in Interactive Multilingual Text Retrieval* Anselmo Peñas, Julio Gonzalo, Felisa Verdejo (Universidad Nacional de Educación a Distancia) **Approximate Ad-hoc Query Engine for Simulation Data* Ghaleb Abdulla, Chuck Baldwin, Terence Critchlow, Roy Kamimura, Ida Lozares, Ron Musick, Nu Ai Tang (Lawrence Livermore National Laboratory) Byung S. Lee, Robert Snapp (University of Vermont, Burlington) **Extracting Taxonomic Relationships from On-Line Definitional Sources Using LEXING* Judith Klavans, Brian Whitman (Columbia University) *Hierarchical Indexing and Document Matching in BoW* Maayan Geffet, Dror G. Feitelson (The Hebrew University) *Scalable Integrated Region-based Image Retrieval using IRM and Statistical Clustering* James Z. Wang (Pennsylvania State University and Stanford University) Yanping Du (Pennsylvania State University)

Tuesday 15:30-16:00
Break. Refreshments in Crystal Foyer.

Session 9: 16:00-17:00

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
	Keynote Address: *Digital Rights Management: What Does This Mean For Libraries?* Pamela Samuelson (Professor of Information Management and of Law, University of California at Berkeley) Introduction by Christine L. Borgman (UCLA)	

18:00-22:00

Banquet and Reception at Virginia Museum of Transportation: Buffet dinner with Virginia wine bar; music by *The Celtibillies.* Presentation of Vannevar Bush Award around 21:00. Buses leave from Hotel Roanoke main entrance beginning at 18:00, and will be available for transport and return throughout the evening.

Directions for walkers: VMT is about five blocks from HRCC. Take the pedestrian bridge over the railroad tracks, then face back across the tracks toward the HRCC. Between you and the tracks is the beginning of Roanoke's interpretive Railwalk. Follow the Railwalk west (left when facing the tracks) to Warehouse Row. Pass warehouses on their left side (away from tracks), continue along Warehouse Row through the parking lot, and pass under the 2nd Street bridge at the Mercury / Redstone rocket. The Virginia Museum of Transportation is the long low brick building directly in front of you. Pass left of the building; the main entrance is about half way down the building.

Wednesday, June 27, 2001 – Main Conference

7:30-9:00
Complimentary Continental Breakfast in Crystal Foyer

Session 10: 9:00-10:00

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
	Keynote Address: *Interoperability: The Still-Unfulfilled Promise of Networked Information* Clifford Lynch (Executive Director, Coalition for Networked Information) Introduction by Erich J. Neuhold (GMD-IPSI)	

10:00-10:30
Break. Refreshments in Crystal Foyer.

Session 11: 10:30-12:30

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
Panel (11C): High Tech or High Touch: Automation and Human Mediation in Libraries **Moderator:** David Levy (Professor, Information School, University of Washington) **Panelists:** William Arms (Cornell University) Oren Etzioni (Associate Professor, Department of Computer Science and Engineering, University of Washington) Diane Nester Kresh (Director Public Services Collections, Library of Congress) Barbara Tillett (Library of Congress)	**Papers (11A): Digital Video Libraries: Design and Access** **Moderator:** Neil Rowe (Naval Postgraduate School) *Cumulating and Sharing End Users Knowledge to Improve Video Indexing in a Video Digital Library* Marc Nanard, Jocelyne Nanard (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) *XSLT for Tailored Access to a Digital Video Library* Michael Christel, Bryan Maher, Andrew Begun (Carnegie Mellon University) *Design of a Digital Library for Human Movement* Jezekiel Ben-Arie, Purvin Pandit, ShyamSundar Rajaram (University of Illinois, Chicago) **A Bucket Architecture for the Open Video Project* Michael L. Nelson (NASA Langley Research Center) Gary Marchionini, Gary Geisler, Meng Yang (University of North Carolina, Chapel Hill) **The Físchlár Digital Video System: A Digital Library of Broadcast TV Programmes* A.F. Smeaton, N. Murphy, N.E. O'Connor, S. Marlow, H. Lee, K. McDonald, P. Browne, J. Ye (Dublin City University)	**Papers (11B): Systems Design and Architecture for Digital Libraries** **Moderator:** Robin Williams (IBM Almaden Research Center) *Design Principles for the Information Architecture of a SMET Education Digital Library* Andy Dong, Alice Agogino (University of California, Berkeley) *Toward A Model of Self-Administering Data* B. Hoon Kang, Robert Wilensky (University of California, Berkeley) *PERSIVAL, a System for Personalized Search and Summarization over Multimedia Healthcare Information* Kathleen R. McKeown, Shih-Fu Chang, James Cimino, Steven K. Feiner, Carol Friedman, Luis Gravano, Vasileios Hatzivassiloglou, Steven Johnson, Desmond A. Jordan, Judith L. Klavans, Vimla Patel, Simone Teufel (Columbia University) Andre Kushniruk (York University) **An Approach to Search for Digital Libraries* Elaine G. Toms, Joan C. Bartlett (University of Toronto) **TilePic: A File Format for Tiled Hierarchical Data* J. Anderson-Lee, R. Wilensky (University of California, Berkeley)

Wednesday 12:30-13:30
Complimentary Lunch for all Registered Attendees in Roanoke CDH.

Session 12: 13:30-15:00

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
Panel (12C): Digital Libraries Supporting Digital Government **Moderator:** Gary Marchionini (Cary C. Boshamer Professor, School of Library and Information Science, University of North Carolina at Chapel Hill) **Panelists:** Lawrence E. Brandt (Program Manager for Digital Government Research, National Science Foundation) Hsinchun Chen (University of Arizona) Anne Craig (Illinois State Library) Judith Klavans (Columbia University)	**Papers (12A): Digital Preservation: Technology, Economics, and Policy** **Moderator:** Michael Seadle (Michigan State University, East Lansing) *Long Term Preservation of Digital Information* Raymond A. Lorie (IBM Almaden Research Center) *Creating Trading Networks of Digital Archives* Brian Cooper, Hector Garcia-Molina (Stanford University) *Cost-driven Design for Archival Repositories* Arturo Crespo, Hector Garcia-Molina (Stanford University)	**Papers (12B): Scholarly Communication and Digital Libraries** **Moderator:** Marianne Afifi (University of Southern California) *Hermes– A Notification Service for Digital Libraries* Daniel Faensen, Lukas Faulstich, Heinz Schweppe, Annika Hinze, Alexander Steidinger (Freie Universität Berlin) *An Algorithm for Automated Rating of Reviewers* Tracy Riggs, Robert Wilensky (University of California, Berkeley) *HeinOnline: An Online Archive of Law Journals* Richard J. Marisa (Cornell University)

Wednesday 15:00-15:30
Break. Refreshments in Crystal Foyer.

Session 13: 15:30-17:00

Crystal Ballroom ABC	Roanoke Ballroom AB	Roanoke Ballroom EFG
Panel (13C): Digital Music Libraries – Research and Development **Moderator:** Christine Brancolini (Director, Indiana University Digital Library Program) **Panelists:** David Bainbridge (University of Waikato) Mary Wallace Davidson (William and Gayle Cook Music Library, Indiana University) Andrew P. Dillon (School of Library and Information Science, Indiana University) Matthew Dovey (Libraries Automation Service, University of Oxford) Jon W. Dunn (Digital Library Program, Indiana University) Michael Fingerhut (Centre Pompidou) Ichiro Fujinaga (Peabody Conservatory of Music, Johns Hopkins University) Eric J. Isaacson (School of Music, Indiana University)	**Papers (13B): Applications of Digital Libraries in the Humanities** **Moderator:** Sally Howe (National Coordination Office for IT Research & Development) *Building a Hypertextual Digital Library in the Humanities: A Case Study on London* Gregory Crane, David A. Smith, Clifford E. Wulfman (Tufts University) **Winner, Vannevar Bush Award** **Document Quality Indicators and Corpus Editions* Jeffrey A. Rydberg-Cox (University of Missouri, Kansas City) Anne Mahoney, Gregory R. Crane (Tufts University) *Digital Atheneum: New Approaches for Preserving, Restoring, and Analyzing Damaged Manuscripts* Michael S. Brown, W. Brent Seales (University of Kentucky, Lexington) **Towards an Electronic Variorum Edition of* Don Quixote Richard Furuta, Shueh-Cheng Hu, Siddarth Kalasapur, Rajiv Kochumman, Eduardo Urbina, Ricardo Vivancos-Pérez (Texas A&M University)	**Papers (13A): Designing Digital Libraries for Education: Technology, Services and User Studies** **Moderator:** Joyce Ray (Institute of Museum and Library Services) *Designing a Digital Library for Young Children: An Intergenerational Partnership* Allison Druin, Ben Bederson, Juan Pablo Hourcade, Lisa Sherman, Glenda Revelle, Michele Platner, Stacy Weng (University of Maryland, College Park) *Dynamic Digital Libraries for Children* Yin Leng Theng, Norliza Mohd-Nasir, George Buchanan, Bob Fields, Harold Thimbleby (Middlesex University), Noel Cassidy (St. Albans School) *Looking at Digital Library Usability from a Reuse Perspective* Tamara Sumner, Melissa Dawe (University of Colorado, Boulder)

Thursday, June 28, 2001 – Workshops

7:30-9:00
Complimentary Continental Breakfast for Registered Workshop Participants in Crystal Foyer

8:00-17:00: Workshops

Visual Interfaces to Digital Libraries - Their Past, Present, and Future – Roanoke Ballroom B
Katy Börner (Assistant Professor, Information Science & Cognitive Science, Indiana University, SLIS)
Chaomei Chen (Reader, Department of Information Systems and Computing; Director, The VIVID Research Centre, Brunel University)

The Technology of Browsing Applications – Roanoke Ballroom GH
Nina Wacholder (Center for Research on Information Access, Columbia University}
Craig Nevill Manning (Computer Science, Rutgers University)

Classification Crosswalks: Bringing Communities Together – Roanoke Ballroom EF
Gail Hodge (Information International Associates, Inc./National Biologica Information Infrastructure)
Paul Thompson (West Group)
Diane Vizine-Goetz (Senior Research Scientist, OCLC Office of Research)
Marcia Lei Zeng (Associate Professor, School of Library and Information Science, Kent State University)

12:00-13:00
Complimentary Lunch for Registered Workshop Participants in Roanoke D.
Box Lunches for "The Technology of Browsing Applications" will be delivered to Roanoke GH.

Participants in the IMLS and NSF DLI2 Workshops, please pick up separate schedules when you register.

Emergency Contact Information

Most conference personnel can be found at the
JCDL Conference Office
Bent Mountain Room, HRCC

General Chair
Edward Fox
Dept. of Computer Science
Virginia Tech
Blacksburg, VA, USA
fox@vt.edu

Program Chair
Christine Borgman
Dept. of Information Studies
UCLA
Los Angeles, CA, USA
cborgman@ucla.edu

Treasurer
Neil Rowe
Computer Science Dept.
Naval Postgraduate School
Monterey, California, USA
rowe@cs.nps.navy.mil

A/V Coordinator
Luis Francisco-Revilla
Texas A&M University
College Station, Texas, USA
l0f0954@csdl.tamu.edu

Demos Chair
Ray Larson
University of California at Berkeley
Berkeley, California, USA
ray@sims.berkeley.edu

Networking Coordinator
Unmil Karadkar
Texas A&M University
College Station, Texas, USA
unmil@csdl.tamu.edu

New Attendees, Doctoral Students
Allison L. Powell
Dept. of Computer Science
University of Virginia
Charlottesville, Virginia, USA
alp4g@cs.virginia.edu

NSDL Support
Brandon Muramatsu
University of California at Berkeley
Berkeley, California, USA
mura@smete.org

Panels Chair
Gene Golovchinsky
FX Palo Alto Laboratory, Inc
Palo Alto, CA, USA
gene@pal.xerox.com

Posters Chair
Craig Nevill-Manning
Dept. of Computer Science
Rutgers University
Piscataway, NJ, USA
nevill@cs.rutgers.edu

European Liaison
Constantino Thanos
Instituto di Elaborazione della Informazione
Consiglio Nazionale delle Richerche
Italy
thanos@iei.pi.cnr.it

Local Arrangements
Robert France
Digital Library Research Laboratory
Virginia Tech
Blacksburg, VA, USA
france@vt.edu

Sponsoring and Exhibiting
Michael L. Nelson
School of Information and Library Science
University of North Carolina
Chapel Hill, NC, USA
mln@ils.unc.edu

Student Volunteers Chair
Ghaleb Abdulla
Lawrence Livermore National Lab
Livermore, CA, USA
abdulla1@llnl.gov

Tutorials Chair
Jonathan Furner
Information Studies
UCLA
Los Angeles, CA USA
jfurner@ucla.edu

Publicity
Edie Rasmussen
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA, USA
erasmus@mail.sis.pitt.edu

Registration
Jim French
Dept. of Computer Science
University of Virginia
Charlottesville, VA USA
french@cs.virginia.edu

Webmaster
Fernando Das-Neves
Dept. of Computer
Virginia Tech
Blacksburg, VA, USA
fdasneve@vt.edu

Workshops Chair
Marianne Afifi
Center for Scholarly Technology
Information Services Division
University of Southern California
Los Angeles, CA, USA
afifi@usc.edu

Papers

- ***Note:*** Long papers are assigned a *30 minute* period, while short papers (shown by an asterisk * preceding the title) have a *15 minute* period.

Paper Presentation Schedule (this is not the complete conference schedule)			
	Monday June 25	**Tuesday June 27**	**Wednesday June 28**
9:00am-10:30		6A 6B	
10:30-11:00	Break	Break	11A 11B
11:00-12:30pm	3A 3B	7A 7B	
12:30-1:30			
1:30-3:00	4A 4B	8A 8B	12A 12B
3:00-3:30			
3:30-5:00pm	5A 5B		13A 13B

MONDAY, June 25

Paper Session 3A (90min): Methods for Classifying and Organizing Content in Digital Libraries (11:00am-12:30pm)

Integrating Automatic Genre Analysis into Digital Libraries
Andreas Rauber, Alexander Mueller-Koegler (Vienna University of Technology)
Text Categorization for Multi-page Documents: A Hybrid Naive Bayes HMM Approach
Paolo Frasconi, Giovanni Soda, Alessandro Vullo (University of Florence)
***Automated Name Authority Control**
James W. Warner, Elizabeth W. Brown (Johns Hopkins University)
***Automatic Event Generation From Multi-lingual News Stories**
Kin Hui, Wai Lam, Helen M. Meng (The Chinese University of Hong Kong)

Paper Session 3B (90min): Digital Libraries for Education: Technology, Services, and User Studies (11:00am-12:30pm)

Linked Active Content: A Service for Digital Libraries for Education
David Yaron, D. Jeff Milton, Rebecca Freeland (Carnegie Mellon University)
A Component Repository for Learning Objects A Progress Report
Jean R. Laleuf, Anne Morgan Spalter (Brown University)
Designing e-Books for Legal Education
Catherine C. Marshall, Morgan N. Price, Gene Golovchinsky, Bill N. Schilit (FX Palo Alto Laboratory)

Paper Session 4A (90min): Approaches to Interoperability Among Digital Libraries (1:30pm-3:00)

***Mapping the Interoperability Landscape for Networked Information Retrieval**
William E. Moen, Teresa Lepchenske (University of North Texas)
***Distributed Resource Discovery: Using Z39.50 to Build Cross-Domain Information Servers**
Ray R. Larson (University of California, Berkeley)
The Open Archives Initiative
Carl Lagoze, Herbert Van de Sompel (Cornell University)
***Enforcing Interoperability with the Open Archives Initiative Repository Explorer**
Hussein Suleman (Virginia Tech)
***Arc-An OAI Service Provider**

Xiaoming Liu, Kurt Maly, Mohammad Zubair, Michael L. Nelson (Old Dominion University)

Paper Session 4B (90min): Digital Libraries and the Web: Technology and Trust (1:30pm-3:00)

Managing Change on the Web
Luis Francisco-Revilla, Frank Shipman, Richard Furuta, Unmil Karadkar, Avital Arora (Texas A&M University)

***Measuring the Reputation of Web Sites: A Preliminary Exploration**
Greg Keast, Joan Cherry, Elaine G. Toms (University of Toronto)

Personalized Spiders for Web Search and Analysis
Michael Chau, Daniel Zeng, Hsinchun Chen (University of Arizona, Tucson)

***Salticus: Guided Crawling for Personal Digital Libraries**
Robin Burke (California State University, Fullerton)

Paper Session 5A (90min): Tools for Constructing and Using Digital Libraries (3:30pm-5:0pm)

Power to the People: End-User Building of Digital Library Collections
Ian H. Witten, David Bainbridge, Stefan J. Boddie (University of Waikato)

***Web-Based Scholarship: Annotating the Digital Library**
Bruce Rosenstock, Michael Gertz (University of California, Davis)

A Multi-View Intelligent Editor for Digital Video Libraries
Brad A. Myers, Juan P. Casares, Scott Stevens, Laura Dabbish, Dan Yocum, Albert Corbett (Carnegie Mellon University)

***VideoGraph: A New Tool for Video Mining and Classification**
Jia-Yu Pan, Christos Faloutsos (Carnegie Mellon University)

Paper Session 5B (90min): Systems Design and Evaluation for Undergraduate Learning Environments (3:30pm-5:0pm)

***The Alexandria Digital Earth Prototype System**
T. R. Smith, G. Janee, J. Frew, A. Coleman, N. Faust (University of California, Santa Barbara)

***Iscapes: Digital Library Environments to Promote Scientific Thinking by Undergraduates in Geography**
Anne J. Gilliland-Swetland, Gregory H. Leazer (University of California, Los Angeles)

***Project ANGEL: an Open Virtual Learning Environment with Sophisticated Access Management**
John MacColl (University of Edinburgh)

***NBDL: A CIS Framework for NSDL**
Joe Futrelle, C. Kevin Chang (University of Illinois, Urbana-Champaign), Su-Shing Chen (University of Missouri, Columbia)

Automatic Identification and Organization of Index Terms for Interactive Browsing
Nina Wacholder, David K. Evans, Judith L. Klavans (Columbia University)

TUESDAY, June 26

Paper Session 6A (90min): Studying the Users of Digital Libraries: Formative and Summative Evaluations (9:00am-10:30)

Public Use of Digital Community Information Systems: Findings from a Recent Study with Implications for System Design
Karen E. Pettigrew (University of Washington, Seattle), Joan C. Durrance (University of Michigan, Ann Arbor)

***Evaluating the Distributed National Electronic Resource**
Peter Brophy, Shelagh Fisher (The Manchester Metropolitan University)

***Collaborative Design with Use Case Scenarios**
Lynne Davis (University Corporation for Atmospheric Research) Melissa Dawe (University of Colorado, Boulder)

Human Evaluation of Kea, an Automatic Keyphrasing System

Steve Jones, Gordon W. Paynter (University of Waikato)

Paper Session 6B (90min): Digital Library Collections: Policies and Practices (9:00am-10:30)

Community Design of DLESE's Collections Review Policy: A Technological Frames Analysis
Michael Khoo (University of Colorado, Boulder)
Legal Deposit of Digital Publications: a Review of Research and Development Activity
Adrienne Muir (Loughborough University)
*Comprehensive Access to Printed Materials (CAPM)
G. Sayeed Choudhury, Mark Lorie, Erin Fitzpatrick, Ben Hobbs, Greg Chirikjian, Allison Okamura (Johns Hopkins University) Nick Flores (University of Colorado, Boulder)
*Technology and Values: Lessons from Central and Eastern Europe
Nadia Caidi (University of Toronto)

Paper Session 7A (90min): Studying the Users of Digital Libraries: Qualitative Approaches (11:00am-12:30pm)

Use of Multiple Digital Libraries: a Case Study
Ann Blandford, Hanna Stelmaszewska (Middlesex University), Nick Bryan-Kinns (Icon MediaLab London)
An Ethnographic Study of Technical Support Workers: Why We Didn't Build a Tech Support Digital Library
Sally Jo Cunningham, Chris Knowles (University of Waikato), Nina Reeves (Cheltenham and Cloucestershire College of Higher Education)
*Developing Recommendation Services for a Digital Library with Uncertain and Changing Data
Gary Geisler (University of North Carolina, Chapel Hill), David McArthur, Sarah Giersch (Eduprise)
*Evaluation of DEFINDER: A System to Mine Definitions from Consumer-Oriented Medical Text
Judith L. Klavans, Smaranda Muresan (Columbia University)

Paper Session 7B (90min): Techniques for Managing Distributed Collections (11:00am-12:30pm)

*An Overview of the Virtual Data Center: A Complete, Open-Source, Digital Library for Distributed Collections of Quantitative Data
Micah Altman, L. Andreev, G. King, E. Kolster, A. Sone, S. Verba (Harvard University), D.L. Kiskis, M. Krott (University of Michigan, Ann Arbor)
*Digital Libraries and Data Scholarship
Bruce R. Barkstrom (NASA Langley Research Center)
SDLIP + STARTS = SDARTS: A Protocol and Toolkit for Metasearching
Noah Green, Panagiotis G. Ipeirotis, Luis Gravano (Columbia University)
Database Selection for Processing k Nearest Neighbors Queries in a Distributed Environment
Clement Yu, Prasoon Sharma, Yan Qin (University of Illinois, Chicago), Weiyi Meng (State University of New York, Binghamton)

Paper Session 8A (2hr): The Sound of Digital Libraries: Audio, Music, and Speech (1:30pm-3:30)

Building Searchable Collections of Enterprise Speech Data
James W. Cooper (IBM T J Watson Research Center), Donna Byron (University of Rochester), Margaret Chan (Columbia University)
*Transcript-Free Search of Audio Archives at the National Gallery of the Spoken Word
J.H.L. Hansen (University of Colorado, Boulder), J.R. Deller, Jr., M.S. Seadle (Michigan State University, Lansing)
*A Watermarking Strategy for Audio Archives at the National Gallery of the Spoken Word
J.R. Deller, Jr., A. Gurijala, M.S. Seadle (Michigan State University, Lansing)
Music-Notation Searching and Digital Libraries
Donald Byrd (University of Massachusetts, Amherst)
*Feature Selection for Automatic Classification of Musical Instrument Sounds
Mingchun Liu, Chunru Wan (Nanyang Technological University)
*Adding Content Based Searching to a Traditional Music Library Catalogue Server
Matthew J. Dovey (Kings College, London)

Paper Session 8B (2hr): Information Search and Retrieval in Digital Libraries (1:30pm-3:30)

***Locating Question Difficulty through Explorations in Question Space**
Terry Sullivan (University of North Texas)
***Browsing by Phrases: Terminological Information in Interactive Multilingual Text Retrieval**
Anselmo Peñas, Julio Gonzalo, Felisa Verdejo (Universidad Nacional de Educación a Distancia)
***Approximate Ad-hoc Query Engine for Simulation Data**
Ghaleb Abdulla, Chuck Baldwin, Terence Critchlow, Roy Kamimura, Ida Lozares, Ron Musick, Nu Ai Tang (Lawrence Livermore National Laboratory), Byung S. Lee, Robert Snapp (University of Vermont, Burlington)
***Extracting Taxonomic Relationships from On-Line Definitional Sources Using LEXING**
Judith Klavans, Brian Whitman (Columbia University)
Hierarchical Indexing and Document Matching in BoW
Maayan Geffet, Dror G. Feitelson (The Hebrew University)
Scalable Integrated Region-based Image Retrieval using IRM and Statistical Clustering
James Z. Wang, Yanping Du (Pennsylvania State University)

WEDNESDAY, June 27

Paper Session 11A (2hr): Digital Video Libraries: Design and Access (10:30am-12:30pm)

Cumulating and Sharing End Users Knowledge to Improve Video Indexing in a Video Digital Library
Marc Nanard, Jocelyne Nanard (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier)
XSLT for Tailored Access to a Digital Video Library
Michael Christel, Bryan Maher, Andrew Begun (Carnegie Mellon University)
Design of a Digital Library for Human Movement
Jezekiel Ben-Arie, Purvin Pandit, ShyamSundar Rajaram (University of Illinois, Chicago)
***A Bucket Architecture for the Open Video Project**
Michael L. Nelson (NASA Langley Research Center), Gary Marchionini, Gary Geisler, Meng Yang (University of North Carolina, Chapel Hill)
***The Físchlár Digital Video System: A Digital Library of Broadcast TV Programmes**
A.F. Smeaton, N. Murphy, N.E. O'Connor, S. Marlow, H. Lee, K. McDonald, P. Browne, J. Ye (Dublin City University)

Paper Session 11B (2hr): Systems Design and Architecture for Digital Libraries (10:30am-12:30pm)

Design Principles for the Information Architecture of a SMET Education Digital Library
Andy Dong, Alice Agogino (University of California, Berkeley)
Toward A Model of Self-administering Data
B. Hoon Kang, Robert Wilensky (University of California, Berkeley)
PERSIVAL, a System for Personalized Search and Summarization over Multimedia Healthcare Information
Kathleen R. McKeown, Shih-Fu Chang, James Cimino, Steven K. Feiner, Carol Friedman, Luis Gravano, Vasileios Hatzivassiloglou, Steven Johnson, Desmond A. Jordan, Judith L. Klavans, Andre Kushniruk, Vimla Patel, Simone Teufel (Columbia University)
***An Approach to Search for Digital Libraries**
Elaine G. Toms, Joan C. Bartlett (University of Toronto)
***TilePic: A File Format for Tiled Hierarchical Data**
J. Anderson-Lee, R. Wilensky (University of California, Berkeley)

Paper Session 12A (90min): Digital Preservation: Technology, Economics, and Policy (1:30pm-3:00)

Long Term Preservation of Digital Information
Raymond A. Lorie (IBM Almaden Research Center)
Creating Trading Networks of Digital Archives
Brian Cooper, Hector Garcia Molina (Stanford University)

Cost-driven Design for Archival Repositories
Arturo Crespo, Hector Garcia-Molina (Stanford University)

Paper Session 12B (90min): Scholarly Communication and Digital Libraries (1:30pm-3:00)

Hermes– A Notification Service for Digital Libraries
Daniel Faensen, Lukas Faulstich, Heinz Schweppe, Annika Hinze, Alexander Steidinger (Freie Universität Berlin)
An Algorithm for Automated Rating of Reviewers
Tracy Riggs, Robert Wilensky (University of California, Berkeley)
HeinOnline: An Online Archive of Law Journals
Richard J. Marisa (Cornell University)

Paper Session 13A (90min): Designing Digital Libraries for Education: Technology, Services and User Studies (3:30pm-5:00)

Designing a Digital Library for Young Children: An Intergenerational Partnership
Allison Druin, Ben Bederson, Juan Pablo Hourcade, Lisa Sherman, Glenda Revelle, Michele Platner, Stacy Weng (University of Maryland, College Park)
Dynamic Digital Libraries For Children
Yin-Leng Theng, Norliza Mohd-Nasir, George Buchanan, Bob Fields, Harold Thimbleby (Middlesex University)
Looking at Digital Library Usability from a Reuse Perspective
Tamara Sumner, Melissa Dawe (University of Colorado, Boulder)

Paper Session 13B (90min): Applications of Digital Libraries in the Humanities (3:30pm-5:00)

Building a Hypertextual Digital Library in the Humanities: A Case Study on London ***(Winner of JCDL 2001 Vannevar Bush paper Award)***
Gregory Crane, David A. Smith (Tufts University)
***Document Quality Indicators and Corpus Editions**
Jeffrey A. Rydberg-Cox (University of Missouri, Kansas City), Anne Mahoney, Gregory R. Crane (Tufts University)
Digital Atheneum: New Approaches for Preserving, Restoring, and Analyzing Damaged Manuscripts
Michael S. Brown, W. Brent Seales (University of Kentucky, Lexington)
***Towards an Electronic Variorum Edition of Don Quixote**
Richard Furuta, Shueh-Cheng Hu, Siddarth Kalasapur, Rajiv Kochumman, Eduardo Urbina, Ricardo Vivancos-Perez (Texas A&M University)

[go to top of page] ▲

library home | list alphabetically | list by SIG | search library | register DL | subscribe DL | feedback

ACM Digital Library

ACM/IEEE-CS Joint Conference on Digital Libraries

Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries
June 24 - 28, 2001, Roanoke, VA USA

access | related SIGs | related conferences

Table of Contents

For full text in PDF, use Adobe Acrobat Reader.

The open archives initiative: building a low-barrier interoperability framework
Carl Lagoze and Herbert Van de Sompel
Pages 54 - 62
metadata: abstract index terms
full text: PDF 349 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Enforcing interoperability with the open archives initiative repository explorer
Hussein Suleman
Pages 63 - 64
metadata: abstract index terms
full text: PDF 117 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Arc: an OAI service provider for cross-archive searching
Xiaoming Liu, Kurt Maly, Mohammad Zubair and Michael L. Nelson
Pages 65 - 66
metadata: abstract index terms
full text: PDF 131 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Managing change on the web
Luis Francisco-Revilla, Frank Shipman, Richard Furuta, Unmil Karadkar and Avital Arora
Pages 67 - 76
metadata: abstract index terms
full text: PDF 268 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Measuring the reputation of web sites: a preliminary exploration
Greg Keast, Elaine G. Toms and Joan Cherry
Pages 77 - 78
metadata: abstract index terms
full text: PDF 115 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Personalized spiders for web search and analysis
Michael Chau, Daniel Zeng and Hinchun Chen
Pages 79 - 87
metadata: abstract index terms
full text: PDF 656 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Salticus: guided crawling for personal digital libraries
Robin Burke
Pages 88 - 89
metadata: abstract
full text: PDF 104 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Different cultures meet (panel session): lessons learned in global digital library development
Ching Chen, Wen Gao, Hsueh-hua Chen, Li-Zhu Zhou and Von-Wun Soo
Pages 90 - 93
metadata: abstract
full text: PDF 138 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Power to the people: end-user building of digital library collections
Ian H. Witten, David Bainbridge and Stefan J. Boddie
Pages 94 - 103
metadata: abstract
full text: PDF 393 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Web-based scholarship: annotating the digital library
Bruce Rosenstock and Michael Gertz
Pages 104 - 105
metadata: abstract index terms
full text: PDF 82 KB
[Discuss this Article | Find Related Articles | Add to Binder]

A multi-view intelligent editor for digital video libraries
Brad A. Myers, Juan P. Casares, Scott Stevens, Laura Dabbish, Dan Yocum and Albert Corbett
Pages 106 - 115
metadata: abstract
full text: PDF 7549 KB
[Discuss this Article | Find Related Articles | Add to Binder]

The Alexandria digital earth prototype
Terence R. Smith, Greg Janee, James Frew and Anita Coleman
Pages 118 - 119
metadata: abstract
full text: PDF 138 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Iscapes: digital libraries environments for the promotion of scientific

thinking by undergraduates in geography

Anne J. Gilliland-Swetland and Gregory L. Leazer
Pages 120 - 121
metadata: abstract
full text: PDF 110 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Project ANGEL: an open virtual learning envoronment with sophisticated access management

John MacColl
Pages 122 - 123
metadata: abstract
full text: PDF 103 KB

[Discuss this Article | Find Related Articles | Add to Binder]

NBDL: a CIS framework for NSDL

Joe Futrelle, Su-Shing Chen and Kevin C. Chang
Pages 124 - 125
metadata: abstract index terms
full text: PDF 154 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Automatic identification and organization of index terms for interactive browsing

Nina Wacholder, Dvid K. Evans and Judith L. Klavans
Pages 126 - 134
metadata: abstract
full text: PDF 290 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Digital library collaborations in a world community

David Fulker, Sharon Dawes, Leonid Kalinichenko, Tamara Sumner, Constantino Thanos and Alex Ushakov
Page 135
metadata: abstract
full text: PDF 72 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Public use of digital community information sstems: findings from a recent study with implications for system design

Karen E. Pettigrew and Joan C. Durrance
Pages 136 - 143
metadata: abstract
full text: PDF 219 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Technology and values: lessons from central and eastern europe

Nadia Caidi
Pages 176 - 177
metadata: abstract
full text: PDF 107 KB
[Discuss this Article | Find Related Articles | Add to Binder]

A digital strategy for the library congress

Alan Inouye, Margaret Hedstrom, Dale Flecker and David Levy
Page 178
metadata: abstract
full text: PDF 53 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Use of multiple digital libraries: a case study

Ann Blandford, Hanna Stelmaszewska and Nick Bryan-Kinns
Pages 179 - 188
metadata: abstract
full text: PDF 173 KB
[Discuss this Article | Find Related Articles | Add to Binder]

An ethnographic study of technical support workers: why we didn't build a tech support digital library

Sally Jo Cunningham, Chris Knowles and Nina Reeves
Pages 189 - 198
metadata: abstract index terms
full text: PDF 222 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Developing recommendation services for a digital library with uncertain and changing data

Gary Geisler, David McArthur and Sarah Giersch
Pages 199 - 200
metadata: abstract
full text: PDF 115 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Evaluation of DEFINDER: a system to mine definitions from consumer-oriented medical text

Judith L. Klavans and Smaranda Muresan
Pages 201 - 202
metadata: abstract
full text: PDF 122 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Overview of the virtual data center project and software
Micah Altman, L. Andreev, M. Diggory, G. King, E. Kolster, A. Sone, S. Verba, Daniel Kiskis and M. Krot
Pages 203 - 204
metadata: abstract index terms
full text: PDF 149 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Digital libraries and data scholarship
Bruce R. Barkstrom
Pages 205 - 206
metadata: abstract
full text: PDF 96 KB
[Discuss this Article | Find Related Articles | Add to Binder]

SDLIP + STARTS = SDARTS a protocol and toolkit for metasearching
Noah Green, Panagiotis G. Ipeirotis and Luis Gravano
Pages 207 - 214
metadata: abstract index terms
full text: PDF 294 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Database selection for processing k nearest neighbors queries in distributed environments
Clement Yu, Prasoon Sharma, Weiyi Meng and Yan Qin
Pages 215 - 222
metadata: abstract
full text: PDF 193 KB
[Discuss this Article | Find Related Articles | Add to Binder]

The president's information technology advisory committee's february 2001 digital library report and its impact
Sally E. Howe, David C. Nagel, Ching-chih Chen, Stephen M. Griffin, James Lightbourne and Walter L. Warnick
Pages 223 - 225
metadata: abstract
full text: PDF 112 KB
[Discuss this Article | Find Related Articles | Add to Binder]

Building searchable collections of enterprise speech data
James W. Cooper, Mahesh Viswanathan, Donna Byron and Margaret Chan
Pages 226 - 234
metadata: abstract
full text: PDF 348 KB

[Discuss this Article | Find Related Articles | Add to Binder]

[Discuss this Article | Find Related Articles | Add to Binder]

Cumulating and sharing end users knowledge to improve video indexing in a video digital library
Marc Nanard and Jocelyne Nanard
Pages 282 - 289
metadata: abstract index terms
full text: PDF 244 KB

[Discuss this Article | Find Related Articles | Add to Binder]

XSLT for tailored access to a digtal video library
Michael G. Christel, Bryan Maher and Andrew Begun
Pages 290 - 299
metadata: abstract index terms
full text: PDF 871 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Design of a digital library for human movement
Jezekiel Ben-Arie, Purvin Pandit and ShyamSundar Rajaram
Pages 300 - 309
metadata: abstract index terms
full text: PDF 353 KB

[Discuss this Article | Find Related Articles | Add to Binder]

A bucket architecture for the open video project
Michael L. Nelson, Gary Marchionini, Gary Geisler and Meng Yang
Pages 310 - 311
metadata: abstract index terms
full text: PDF 780 KB

[Discuss this Article | Find Related Articles | Add to Binder]

The físchlár digital video system: a digital library of broadcast TV programmes
A. F. Smeaton, N. Murphy, N. E. O'Connor, S. Marlow, H. Lee, K. McDonald, P. Browne and J. Ye
Pages 312 - 313
metadata: abstract
full text: PDF 100 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Design principles for the information architecture of a SMET education digital library
Andy Dong and Alice M. Agogino
Pages 314 - 321
metadata: abstract

full text: PDF 435 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Toward a model of self-administering data

ByungHoon Kang and Robert Wilensky0

Pages 322 - 330

metadata: abstract

full text: PDF 301 KB

[Discuss this Article | Find Related Articles | Add to Binder]

PERSIVAL, a system for personalized search and summarization over multimedia healthcare information

Kathleen R. McKeown, Shih-Fu Chang, James Cimino, Steven Feiner, Carol Friedman, Luis Gravano, Vasileios Hatzivassiloglou, Steven Johnson, Desmond A. Jordan, Judith L. Klavans, André Kushniruk, Vimla Patel and Simone Teufel

Pages 331 - 340

metadata: abstract index terms

full text: PDF 360 KB

[Discuss this Article | Find Related Articles | Add to Binder]

An approach to search for the digital library

Elaine G. Toms and Joan C. Bartlett

Pages 341 - 342

metadata: abstract index terms

full text: PDF 122 KB

[Discuss this Article | Find Related Articles | Add to Binder]

TilePic: a file format for tiled hierarchical data

Jeff Anderson-Lee and Robert Wilensky

Pages 343 - 344

metadata: abstract

full text: PDF 101 KB

[Discuss this Article | Find Related Articles | Add to Binder]

High tech or high touch (panel session): automation and human mediation in libraries

David Levy, William Arms, Oren Etzioni, Diane Nester and Barbara Tillett

Page 345

metadata: abstract

full text: PDF 51 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Long term preservation of digital information

Raymond A. Lorie

Pages 346 - 352

Designing a digital library for young children

Allison Druin, Benjamin B. Bederson, Juan Pablo Hourcade, Lisa Sherman, Glenda Revelle, Michele Platner and Stacy Weng
Pages 398 - 405

metadata: abstract index terms

full text: PDF 1125 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Dynamic digital libraries for children

Yin Leng Theng, Norliza Mohd-Nasir, George Buchanan, Bob Fields, Harold Thimbleby, Noel Cassidy and Noel Cassidy
Pages 406 - 415

metadata: abstract

full text: PDF 1087 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Looking at digital library usability from a reuse perspective

Tamara Sumner and Melissa Dawe
Pages 416 - 425

metadata: abstract index terms

full text: PDF 1864 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Building a hypertextual digital library in the humanities: a case study on London

Gregory Crane, David A. Smith and Clifford E. Wulfman
Pages 426 - 434

metadata: abstract

full text: PDF 353 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Document quality indicators and corpus editions

Jeffrey A. Rydberg-Cox, Anne Mahoney and Gregory R. Crane
Pages 435 - 436

metadata: abstract

full text: PDF 112 KB

[Discuss this Article | Find Related Articles | Add to Binder]

The digital atheneum: new approaches for preserving, restoring and analyzing damaged manuscripts

Michael S. Brown and W. Brent
Pages 437 - 443

metadata: abstract

full text: PDF 2248 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Towards an electronic variorum dition of *Don Quixote*
Richard Furuta, Shueh-Cheng Hu, Siddarth Kalasapur, Rajiv Kochumman, Eduardo Urbina and Ricardo Vivancos
Pages 444 - 445
metadata: abstract
full text: PDF 699 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Digital music libraries - research and development
David Bainbridge, Gerry Bernbom, Mary Wallace, Andrew P. Dillon, Matthew Dovey, Jon W. Dunn, Michael Fingerhut, Ichiro Fujinaga and Eric J. Isaacson
Pages 446 - 448
metadata: abstract

[Discuss this Article | Find Related Articles | Add to Binder]

Content management for digital museum exhibitions
Jen-Shin Hong, Bai-Hsuen Chen, Jieh Hsiang and Tien-Yu Hsu
Page 450
metadata: abstract
full text: PDF 84 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Demonstration of hierarchical document clustering of digital library retrieval results
C. R. Palmer, J. Pesenti, R. E. Valdes-Perez, M. G. Christel, A. G. Hauptmann, D. Ng and H. D. Wactlar
Page 451
metadata: abstract index terms
full text: PDF 94 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Indiana university digital music library project
Jon W. Dunn and Eric J. Isaacson
Page 452
metadata: abstract index terms
full text: PDF 97 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Interactive visualization of video metadata
Mark Derthick
Page 453
metadata: abstract index terms
full text: PDF 115 KB

[Discuss this Article | Find Related Articles | Add to Binder]

PERSIVAL demo: categorizing hidden-web resources

Panagiotis G. Ipeirotis, Luis Gravano and Mehran Sahami
Page 454
metadata:
full text: PDF 99 KB

[Discuss this Article | Find Related Articles | Add to Binder]

PERSIVAL: personalized summarization over multimedia health-care information

Noemie Elhadad, Min-Yen Kan, Simon Lok and Smaranda Muresan
Page 455
metadata: abstract
full text: PDF 44 KB

[Discuss this Article | Find Related Articles | Add to Binder]

View segmentation and static/dynamic summary generation for echocardiogram videos

Shahram Ebadollahi and Shih-Fu Chang
Page 456
metadata:
full text: PDF 76 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Stanford encyclopedia of philosophy: a dynamic reference work

Edward N. Zalta, Colin Allen and Uri Nodelman
Page 457
metadata: abstract
full text: PDF 41 KB

[Discuss this Article | Find Related Articles | Add to Binder]

A system for adding content-based searching to a traditional music library catalogue server

Matthew J. Dovey
Page 458
metadata: abstract index terms
full text: PDF 101 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Using the repository explorer to achieve OAI protocol compliance

Hussein Suleman
Page 459
metadata:
full text: PDF 118 KB

Workshop 3: classification crosswalks

Paul Thompson, Traugott Koch, John Carter, Heike Neuroth, Ed O'Neill and Dagobert Soergel

Page 484

metadata: abstract

full text: PDF 84 KB

[Discuss this Article | Find Related Articles | Add to Binder]

Workshop 4: digital libraries in asian languages

Su-Shing Chen and Ching-chih Chen

Page 485

metadata:

[Discuss this Article | Find Related Articles | Add to Binder]

Workshop 5: information visualization for digital libraries: defining a research agenda for heterogeneous multimedia collections

Lucy Nowell and Elizabeth Hetzler

Page 486

metadata: abstract

[Discuss this Article | Find Related Articles | Add to Binder]

VideoGraph: a new tool for video mining and classification

Jia-Yu Pan and Christos Faloutsos

Page

metadata: abstract

[Discuss this Article | Find Related Articles | Add to Binder]

For full text in PDF, use Adobe Acrobat Reader.

library home | list alphabetically | list by SIG | search library | register DL | subscribe DL | feedback

MADE WITH CASCADING STYLE SHEETS | W3C CSS | W3C HTML 4.0

Integrating Automatic Genre Analysis into Digital Libraries

Andreas Rauber, Alexander Müller-Kögler
Department of Software Technology
Vienna University of Technology
Favoritenstr. 9 - 11 / 188
www.ifs.tuwien.ac.at/ifs

ABSTRACT

With the number and types of documents in digital library systems increasing, tools for automatically organizing and presenting the content have to be found. While many approaches focus on topic-based organization and structuring, hardly any system incorporates automatic structural analysis and representation. Yet, genre information (unconsciously) forms one of the most distinguishing features in conventional libraries and in information searches. In this paper we present an approach to automatically analyze the structure of documents and to integrate this information into an automatically created content-based organization. In the resulting visualization, documents on similar topics, yet representing different genres, are depicted as books in differing colors. This representation supports users intuitively in locating relevant information presented in a relevant form.

Keywords

Genre Analysis, Self-Organizing Map (SOM), SOMLib, Document Clustering, Visualization, Metaphor Graphics

1. INTRODUCTION

While the question of *what* a document is about has been recognized as being crucial for presenting relevant information to a user, the question of *how* a given piece of information is presented is largely neglected by most present electronic information systems. Yet, this type of information is – mostly unconsciously – used in almost any contact with information in everyday life. Personal letters are treated differently than mass-mailings, a short story is read on different occasions than long novels, popular science literature addresses a different readership than dissertations or scientific papers, both of which themselves will provide highly similar information at differing levels of detail for different audiences. Specific sub-genres, such as for example, executive summaries or technical reports, were even specifically designed to satisfy the same information need, i.e. to provide information about a given topic, in different ways. Whenever looking for information, these issues are taken into account, and they form one of the most important distinguishing features in conventional libraries, together with other non-content based information such as the age of a document, the fact whether it looks like it is being used frequently or remains untouched for long periods of time, and many others.

As long as a digital library can be cared for in a way similar to how conventional libraries are organized, this type of information is carefully captured in the form of metadata descriptions, and provided to the user, albeit mostly in rather inconvenient, not intuitive textual form. Yet, with this information available, ways for more intuitive representations can be devised. A different situation is encountered in many less-controlled digital library settings, where pieces of information from different sources are integrated, or the mere amount of information added to a repository effectively prevents it from being manually indexed and described. For these settings an automatic analysis of the structure of a given piece of information is essential to allow the user to quickly find the correct document, not only in terms of the content provided, but also with respect to the way this content is presented.

In this paper we present a way to provide automatic analysis of the structure of text documents. This analysis is based on a combination of various surface level features of texts, such as word statistics, punctuation information, the occurrences of special characters and keywords, as well as mark-up tags capturing image, equation, hyperlink and similar information. Based on these structural descriptions of documents, the self-organizing map (SOM) [12], a popular unsupervised neural network, is used to cluster documents according to their structural similarities. This information is incorporated into the SOMLib digital library system [17] which provides an automatic, topic-based organization of documents using again the self-organizing map to group documents according to their content. The libViewer, a metaphor-graphical interface to the SOMLib system depicts the documents in a digital library as hardcover or paperback books, binders, or papers, sorted by content into various bookshelves, labeled by automatically extracted content descriptors using the LabelSOM technique. Integrating the results of the structural analysis of documents allows us to color the documents, which are sorted by subject into the various shelves, according to their structural similarities, making e.g. complex descriptions stand apart

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

from summaries or legal explanations on the same subject. Similarly, interviews on a given topic are depicted different from reports, as are numerical tables or result listings. We demonstrate the benefits of an automatic structural analysis of documents in combination with content-based classification using a collection of news articles from several Austrian daily, weekly and monthly news magazines.

The reminder of this paper is structured as follows. Section 2 provides a brief introduction into the principles of genre analysis and presents a review of related work in this area. We proceed by presenting the architecture and training procedure of the self-organizing map in Section 3. The application of the SOM to content-based document classification is presented in Section 4, including an overview of the various modules of the SOMLib digital library system with a special focus on the metaphor-graphics based libViewer interface. We then present our approach to structural classification of documents and its integration with the content-based representation provided by the SOMLib system in Section 5. Section 6 presents our experimental results using a collection of newspaper articles, reporting on content-based organization, structural classification, and their integration. Some conclusions as well as future work are listed in Section 7.

2. GENRE ANALYSIS

Genre analysis has a long history in linguistic literature. Conventionally, *genre* is associated with terms such as short stories, science fiction, novels of the 17th or 18th century, fiction, reports, satire, and many others. Still, the definition of genre is somewhat vague. According to Webster's Dictionary of English Language, genre is defined as *a category of artistic, musical, or literary composition characterized by a particular style, form, or content.* Although differing definitions may be found, the main goal of genre analysis is to identify certain subgroups within a set of given objects that share a common form of transmission, purpose, and discourse properties. Basically, the term *genre* can be applied to most forms of communications, although it is frequently restricted to non-interactive, and, for the scope of this paper, literary information, excluding music or film genres. While the common interpretation of genre refers to literary styles, such as *fiction, novel, letter, manuals*, etc., automatic analysis of genres takes a slightly different approach, focusing on structural analysis using surface level cues as the main structural similarity between documents, from which genre-style information is deducted..

Several approaches have been taken to evaluate the structure or readability of text documents, resulting in numerous different measures for grading texts automatically based on surface features. Many of these features are readily available in various implementations of the Unix *STYLE* command [6]. Among the measures included in this package are the *Kincaid Formula*, which is targeted towards technical material, having been developed for Navy training manuals. The *Flesh reading easy formula* stems from 1948 and is based on English school texts covering grades 3 to 12. A similar measure is the *SMOG-Grading*, or the *WSTF Index*, which has been developed specifically for German texts. All these measures basically compute their score by combining information about the number of words and syllables per sentence as well as characters per word statistics, weighted by various constants, to obtain the according grades.

More complex stylistic analyses can be found in the seminal work of Biber [1, 2]. He uses metrics such as pronoun counts and general text statistics to cluster texts in order to find underlying dimensions of variation and to detect general properties of genres.

More recently, classification of text documents by genre has been analyzed by Karlgren et al. [10]. Again, a number of different features are used to describe the structural characteristics of documents. However, additionally to the standard surface cues, additional features requiring syntactic parsing and tagged texts, such as required for noun counts, present participle count etc., were included. Discriminant analysis is used to obtain a set of discriminant functions based on a pre-categorized training set. This line of research is continued in [8], reporting in detail on the various features used for stylistic analysis. The stylistic variations of documents are further visualized as scatter plots based on combinations of two features. Specific areas are then (manually) assigned special genre-type descriptors to help users with analyzing the clusters of documents found in the scatter plot.

Recognizing the importance of integrating genre analysis into a content-based information retrieval process, the DropJaw interface [3, 9] incorporates genre-based classification using C4.5 based decision trees into content-based clustering using a hierarchical agglomerative group-average clustering algorithm. Documents are then represented in a two-dimensional matrix, with the rows representing the topical clusters found in the document set, whereas the columns organize these documents according to a number of genres the decision tree was trained to recognize.

A different approach describing documents by a number of facets rather than directly assigning a genre is reported in [11]. A facet is a property which distinguishes a class of texts that answers to certain practical interests, and which is associated with a characteristic set of computable structural or linguistic properties. Three principal categorical facets are analyzed. *Brow* characterizes a text with respect to the intellectual background required to understand a text, subdivided into *popular, middle, upper-middle*, and *high.* A binary *narrative* facet decides whether a text is written in a *narrative style*, and the third facet, *genre*, classifies a text either as *reportage, editorial, scitech, legal, nonfiction*, or *fiction.* A set of 55 lexical, character-level and derivative cues are used to describe the documents, and logistic regression is used to create a classifier based on a training set of 402 manually classified texts.

In [21], Ries applies genre classification to spontaneous spoken conversations, including features such as pauses in the conversation as well as histograms of a number of keywords, using a backpropagation-type neural network for the subsequent analysis.

It is interesting to note, that, although unsupervised methods are frequently used for content-based analysis of information, most of current research work turns to supervised models when it comes to the analysis of genre. This might be due to the case, that people tend to think in terms of well-defined genres, rather than in terms of structurally similar documents. Still, we find documents to frequently exhibit characteristics of several different genres to differing degrees. This is the more so as for hardly any genre there is a strict and well-defined, non-overlapping set of criteria by which it can be described, making strict classification as impossible a task as strict content-based classification. Similar as for content-based document organization, unsupervised cluster analysis of genre-oriented document descriptions should be able to capture the structural similarities accordingly.

3. THE SELF-ORGANIZING MAP

The self-organizing map [13] provides cluster analysis by producing a mapping of high-dimensional input data $x, x \in \Re^n$, onto a usually 2-dimensional output space while preserving the topological relationships between the input data items as faithfully as possible. This model consists of a set of units, which are arranged in some topology where the most common choice is a two-dimensional grid. Each of the units i is assigned a weight vector m_i of the same dimension as the input data, $m_i \in \Re^n$, initialized with random values.

During each learning step, the unit c with the highest activity level, i.e. the *winner* c with respect to a randomly selected input pattern x, is adapted in a way that it will exhibit an even higher activity level at future presentations of that specific input pattern. Commonly, the activity level of a unit is based on the Euclidean distance between the input pattern and that unit's weight vector. The unit showing the lowest Euclidean distance between it's weight vector and the presented input vector is selected as the winner. Hence, the selection of the winner c may be written as given in Expression (1).

$$c : ||x - m_c|| = \min_i\{||x - m_i||\} \qquad (1)$$

Adaptation takes place at each learning iteration and is performed as a gradual reduction of the difference between the respective components of the input vector and the weight vector. The amount of adaptation is guided by a learning rate α that is gradually decreasing in the course of time. This decreasing nature of adaptation strength ensures large adaptation steps in the beginning of the learning process where the weight vectors have to be tuned from their random initialization towards the actual requirements of the input space. The ever smaller adaptation steps towards the end of the learning process enable a fine-tuned input space representation.

As an extension to standard competitive learning, units in a time-varying and gradually decreasing neighborhood around the winner are adapted, too. Pragmatically speaking, during the learning steps of the self-organizing map a set of units around the winner is tuned towards the currently presented input pattern enabling a spatial arrangement of the input patterns such that alike inputs are mapped onto regions close to each other in the grid of output units. Thus, the training process of the self-organizing map results in a topological ordering of the input patterns.

The neighborhood of units around the winner may be described implicitly by means of a (Gaussian) neighborhood-kernel h_{ci} taking into account the distance—in terms of the output space—between unit i under consideration and unit c, the winner of the current learning iteration. This neighborhood-kernel assigns scalars in the range of [0, 1] that are used to determine the amount of adaptation ensuring that nearby units are adapted more strongly than units farther away from the winner.

It is common practice that in the beginning of the learning process the neighborhood-kernel is selected large enough to cover a wide area of the output space. The spatial width of the neighborhood-kernel is reduced gradually during the learning process such that towards the end of the learning process just the winner itself is adapted. This strategy enables the formation of large clusters in the beginning and fine-grained input discrimination towards the end of the learning process.

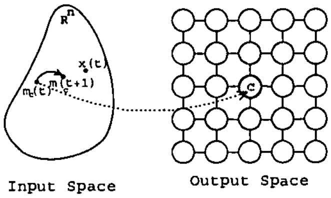


Figure 1: SOM architecture and training process

In combining these principles of self-organizing map training, we may write the learning rule as given in Expression (2). Please note that we make use of a discrete time notation with t denoting the current learning iteration. The other parts of this expression are α representing the time-varying learning rate, h_{ci} representing the time-varying neighborhood-kernel, x representing the currently presented input pattern, and m_i denoting the weight vector assigned to unit i.

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)] \qquad (2)$$

A simple graphical representation of a self-organizing map's architecture and its learning process is provided in Figure 1. In this figure the output space consists of a square of 25 units, depicted as circles. One input vector $x(t)$ is randomly chosen and mapped onto the grid of output units. In the next step of the learning process, the winner c showing the highest activation is selected. Consider the winner being the unit depicted as the black unit in the figure. The weight vector of the winner, $m_c(t)$, is now moved towards the current input vector. This movement is symbolized in the input space in Figure 1. As a consequence of the adaptation, unit c will produce an even higher activation with respect to input pattern x at the next learning iteration, $t + 1$, because the unit's weight vector, $m_c(t + 1)$, is now nearer to the input pattern x in terms of the input space. Apart from the winner, adaptation is performed with neighboring units, too. Units that are subject to adaptation are depicted as shaded units in the figure. The shading of the various units corresponds to the amount of adaptation and thus, to the spatial width of the neighborhood-kernel. Generally, units in close vicinity of the winner are adapted more strongly and consequently, they are depicted with a darker shade in the figure.

4. THE SOMLIB SYSTEM

4.1 Feature extraction

In order to utilize the SOM for organizing documents by their topic a vector-based description of the content of the documents needs to be created. While manually or semi-automatically extracted content descriptors may be used, research results have shown that a rather simple word frequency based description is sufficient to provide the necessary information in a very stable way [4, 14, 15, 19]. For this word frequency based representation a vector structure is created consisting of all words appearing in the document collection. This list of words is usually cleaned from so-called stop-words, i.e. words that do not contribute to content representation and topic discrimination between documents. Again, while manually crafted stop-word lists may be used, simple statistics allow the removal of most stop-words in a very convenient and language- or subject-independent way. On the one hand, words appearing in too many documents, e.g. in more than half of all documents, can be removed without the risk of loosing content information, as the content conveyed by these words is too general. On the other hand, words appearing in less than a minimum number of, say, 5 to 10 documents can be omitted for content-based classification, as the resulting sub-topic granularity would be too small to form a topic cluster of its own. Note, that the situation is different in the information retrieval domain, where rather specific terms need to be indexed to facilitate retrieval of a very specific subset of documents. In this respect, content-based organization and browsing of documents constitutes a conceptually different approach to accessing and interacting with document archives by browsing topical hierarchies. This obviously has to be supplemented by various searching facilities, including information retrieval capabilities as they are currently realized in many systems.

The documents are described within the resulting feature space of commonly between 5.000 and 15.000 dimensions, i.e. distinct terms, by the words they are made up of. While a basic binary indexing may be used to describe the content of a document by simply stating whether a word appears in the document or not, more sophisticated schemes, such as *tf* × *idf*, i.e. term frequency times inverse document frequency [22], provide a better content representation. This weighting scheme assigns higher values to terms that appear frequently within a document, i.e. have a high term frequency, yet rarely within the complete collection, i.e. have a low document frequency. Usually, the document vectors are normalized to unit length to make up for length differences of the various documents.

4.2 Topic-based organization

The resulting vector representations are fed into a standard self-organizing map for cluster analysis. As a result, documents on similar topics are located on neighboring units in the two-dimensional map display. In the simplest form, a document collection may then be represented as a rectangular table with similar documents being mapped onto the same cells. Using this model, users find a document collection to be automatically structured by content in a way similar to how documents are organized into shelves in conventional libraries. Due to its capabilities of automatically structuring a document collection by subject, we have chosen the SOM as the basic building block of our SOMLib digital library system [19]. Enhanced models of the SOM, such as the growing hierarchical self-organizing map (GHSOM), further allow the automatic detection of topical hierarchies by creating a layered structure of independent SOMs that adapt their size accordingly [20].

4.3 Labeling

While the SOM found wide appreciation in the field of text classification, its application had been limited by the fact that the topics of the various cluster were not evident from the resulting mapping. In order to find out which topics are covered in certain areas of the map, the actual articles had to be read to find descriptive keywords for a cluster. To counter this problem, we developed the LabelSOM method, which analyses the trained SOM to automatically extract a set of attributes, i.e. keywords, that are most descriptive for a unit [16]. Basically, the attributes showing a low quantization error value and a high weight vector value, comparable to a low variance and a high mean among all input vectors mapped onto a specific unit, are selected as labels. Thus, the various units are characterized by keywords describing the topics of the documents mapped onto them.

4.4 Visualization

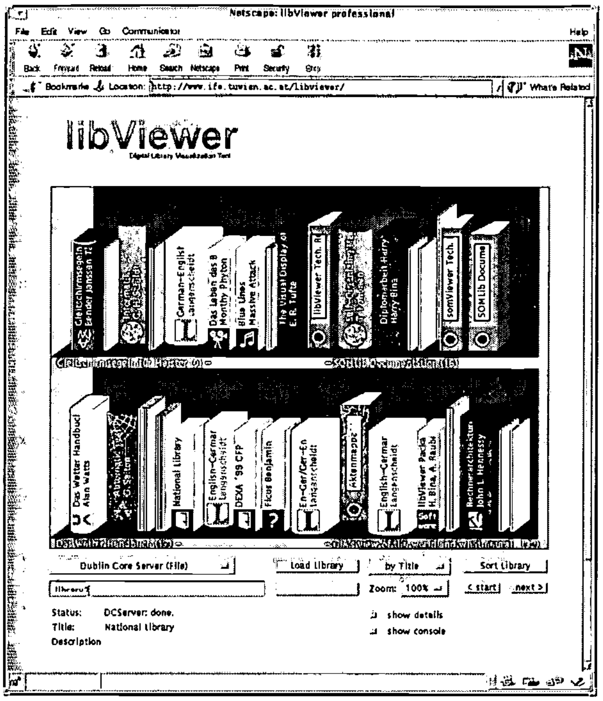
Last, but not least, while the spatial organization of documents on the 2-dimensional map in combination with the automatically extracted concept labels supports orientation in and understanding of an unknown document repository, much information on the documents cannot be told from the resulting representation. Information like the size of the underlying document, its type, the date it was created, when it was accessed for the last time and how often it has been accessed at all, its language etc. is not provided in an intuitively interpretable way. Rather, users are required to read and abstract from textual descriptions, inferring the amount or recent-ness of information provided by a given document by comparing size and date information.

We thus developed the libViewer, a metaphor-graphics based interface to a digital library [18]. Documents are no longer represented as textual listings, but as graphical objects of different *representation types* such as binders, papers, hardcover books, paperbacks etc, with further metadata information being conveyed by additional metaphors such as *spine width, logos, well-thumbed spines*, different degrees of *dustiness, highlighting glares, position in the shelf* and others. Based on these metaphors we can define a set of mappings of metadata attributes to be visualized, allowing the easy understanding of documents, similar to the usage of Chernoff faces for multidimensional space representation [5]. Figure 2 provides an example of the visualization of documents in a digital library using the libViewer interface. Documents are depicted using different document type representations, with additional metadata being conveyed by their position, color, spine width, the logos and textual information depicted on the spine, dust or highlighting glares, well-thumbed bindings and others. Metadata information about documents thus can be easily interpreted and compared across a collection, with a larger amount of information being represented as compared to standard textual descriptions.

5. STRUCTURE AND GENRE ANALYSIS

5.1 Structural features

Although the content-based organization and metaphor-graphical visualization of documents provided by the SOMLib system greatly supports the user in interacting with a digital library, all meta-information about documents has to be created and provided manually. While the size of


Netscape: libViewer professional
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: http://www.ifs.tuwien.ac.at/libviewer/
What's Related
libViewer
Digital Library Visualization Tool
Gleitschirmsegeln Bender Janssen T2
German-Englist Langenscheidt
Das Leben das B Monthy Phyton
Blue Lines Massive Attack
The Visual Display of E. R. Tufte
libViewer Tech. R
Diplomarbeit Harry Harry Bina
somViewer Tech.
SOMLib Docume
Das Wetter Handbuch Alan Watts
National Library
English-German Langenscheidt
DEXA '99 CFP
Ficus Benjamin
En-Ger/Ger-En Langenscheidt
Aktenmappe
English-German Langenscheidt
libViewer Packa H. Bina, A. Rauber
Rechnerarchitektur John L. Hennessy
Dublin Core Server (File)
Load Library
by Title
Sort Library
library
Zoom: 100%
&lt; start
next &gt;
show details
show console

Status: DCServer: done.
Title: National library
Description

Figure 2: libViewer visualization of a digital library

a document, its date of creation, the date of last access, or the author are usually available, hardly any consistent genre-information is provided in most electronic document collections that have too high a volume of documents added frequently to allow manual classification. Yet, this type of information is important for a user to be able to pick the most appropriate document.

Similar to the task of content-based organization, we would like to have a way to automatically organize and categorize documents by their structure and stylistic features into basic types of genre, and to fuse this information with the content-based organization provided by the SOMLib system and its libViewer interface. Yet, we do not want to use supervised models, both because of the limitations introduced by the supervised process and because of the tedious task of having to produce an accurate training set manually. We thus propose to follow an approach similar to the one chosen for content-based document organization creating a feature vector representation of documents capturing the stylistic characteristics of documents. Clustering documents according to their similarity conveyed by their structural features should then reveal basic types of documents, or genres.

5.1.1 Surface level cues

For the set of features we have to restrict ourselves to those features that can be automatically extracted from documents with acceptable computational costs. Furthermore, similar to content-based classification, we want the process to be as language- and domain independent as possible to allow flexible application of the system in different settings. Basically, four distinct types of features can be distinguished, which are (1) text complexity information and text statistics, (2) special character and punctuation counts, (3) characteristic keywords, and (4) format-specific markups.

5.1.2 Text complexity measures

Text complexity measures are based on word statistics such as the average number of words per sentence, the average word length, number of sentences and paragraphs, and similar formatting information. While being rather simple metrics, the complexity of certain constructs turns out to be captured quite well by these measures, especially if combined with other characteristics such as punctuation marks. More extensive measures analyzing the nesting depth of sentences may be considered, although we did not integrate them into the experiments presented below. Furthermore, instead of using the basic measures, derived measures like any of the existing readability grade methods, such as Kincaid, Coleman-Liau, Wheeler-Smith Index, Flesh Index, may be used as more condensed representations. Still, these formulas are all based on the above-mentioned basic measures anyway, using various transformations to obtain graded representations according to stylistic evaluation parameters. We thus refrained from using those in our initial experiments, although they may be considered for follow-up evaluations.

5.1.3 Special character and punctuation counts

A wealth of stylistic information can be obtained from specific character counts. As most prominent among these we consider punctuation marks, as some of them are rather characteristic for certain text genres. For example, the presence of exclamation mark (*!*) is highly indicative of more emotional texts as opposed to pure fact reports in a news magazine setting. Interviews exhibit a rather high count of question marks and colons, whereas high counts for semicolons or commas are indicative of more complex sentence structures, especially if co-occurring with rather high sentence lengths. Otherwise, if co-occurring with rather short sentences they might rather be attributed to listings and enumerations of information items. Similarly, we have to analyze the occurrence of quotation marks, hyphens, periods, apostrophe, slash marks, various brackets and others. More complex punctuation information might be worth considering if it can be feasibly extracted from the given texts, such as ellipsis points, differences between single and double quotation marks, or regarding the usage of dashes versus hyphens, especially if the semantic context can be deducted. However, as these more complex features could not be extracted from the given text base integrating different sources we had to refrain from representing this level of structural information for the given experiments.

Other characters worth incorporating into stylistic analysis are financial symbols like *\$, £, euro*. Furthermore, numbers as well as mathematical symbols, copyright and paragraph signs are worth including as they do hint at special categories of information, be it, for example, technical discussions, price listings, legal information, or simply examples to clarify and expand on a given topic.

5.1.4 Stopwords and keywords

Contrary to the principles of content representation, a lot of genre-information is conveyed by stop-words such as pronouns or adverbs. Thus, a list of characteristic words is added to the list of features, containing words such as *I, you,* *we, me, us, mine, yours* or *much, little, large, very, highly, probably, mostly, certainly, which, that, where* and others. Please note that this list is language dependent, yet may be easily adapted for most languages. (These adaptions may be rather complex for specific languages using word inflections rather than specific keywords for some characteristic sentence structures.) Depending on the specific document collection to be considered and the desired focus of genre analysis, additional keywords may be added to that list to facilitate, for example, the recognition of fact-reporting versus opinion-modifying articles, or the separation of speculative articles.

5.1.5 Mark-up tags

The fourth group of features is formed by specific mark-up tags that are used to extract information about the document. Using such mark-up tags, information such as the amount of images present in a given document, the number of tables and equations, links, references, etc. can be extracted and included in the genre analysis. These obviously have to be adapted to the actual formatting of the source documents in such a way that they are consistently mapped, such as, e.g. mapping the *\begin{figure}* and the *IMGSRC* tags onto the image count feature for LaTeX and HTML documents.

Overall, the parser currently recognizes almost 200 different features, some of which are specifically geared towards a specific file format, such as HTML documents, whereas others are generally applicable. Depending on the goal of the structure analysis, only a subset of the available attributes may be selected for further analysis.

The resulting document descriptions are further fed into a self-organizing map for training. As a result of the training process documents with similar stylistic characteristics are located close to each other on neighboring units of the SOM. We may thus find longer documents with rather complex sentences in one area of the map, whereas in another area interviews, characterized by shorter sentences with a high count of colons and quotation marks might be located. While this map may now serve as a kind of genre-based access to the document archive, it needs to be integrated into the content-based library representation to support users in their information finding process.

5.2 Integrating content and structure based classification

Two possibilities offer themselves for the integrated representation of content- and structure-based organization within the SOMLib system. We can either chose a semi-automatic approach by assigning a specific document representation type to a certain region on the genre-map, such as, for example, assigning all documents in the upper right area of this map the representation type *binder*, whereas all documents in the area on the lower left part of the map may be depicted as *hardcover* documents. This approach allows a very intuitive representation of documents in the content-based representation if a sensible assignment of the genres identified by the map to the available representation types can be defined. However, this approach has shown to have several deficiencies when applied to large and unknown document collections. Firstly, the number of available representation types is rather limited as opposed to the number of different document types present in any collection. While the total number of 4 representation types available in the libViewer system so far may be supplemented with additional object types showing, for example, additional bindings, the total possible number still will be rather limited. Furthermore, as the available representation types are strictly distinct from each other, no gradual shift from one type of document to the other can be conveyed, thus actually forcing the documents to be definitely of one genre or the other. Providing more subtle information about its general structure, which might be well in between two specific genres, would be more appropriate and highly preferable.

Secondly, a rather high manual effort is required to analyze the actual genres identified by the SOM to provide a sensible mapping of genre areas on the map onto the respective graphical metaphors. Yet, the precise mapping is crucial for the usability of the classification result as users will associate a specific type of information with a certain representation metaphor. The assignment of a wrong representation template may thus turn out to be highly counterproductive for information location.

We thus favour an automatic approach for integrating the information provided by the genre map into the content-based visualization using the color-metaphor. Documents of similar structure shall be assigned a similar color to allow intuitive recognition and interpretation of structural similarities. This metaphor turns out to be almost perfectly suited for conveying the desired information as it does not transport any specific meaning in the given setting by itself, as opposed to the realistic document representation types, which are intuitively associated with a certain kind of information.

A rather straight-forward mapping of the position on the genre-map onto a specific color is realized by mapping the rectangular map area onto a plane of the RGB color cube, similar to the color-coding technique for cluster identification in SOMs [7]. Thus, documents mapped onto neighboring units on the genre map will be depicted in similar colors, allowing easy recognition of mutal similarity in style as well as depicting even gradual transitions between the various structural clusters. On the other hand, documents in different regions on the genre map, exhibiting a clearly distinct structure, are thus depicted in different colors on the content-based libViewer visualization.

6. EXPERIMENTS

6.1 Data set

Various series of experiments have been performed in different settings, including technical documents and web site analysis. For the experiments presented below we created a collection of news reports by downloading the web-editions of 14 daily, weekly, or monthly Austrian newspapers and print magazines. This setting exhibits several characteristica typical for digital libraries that cannot be tendered to manually as carefully as necessary. Information from different sources having different internal classification schemata is integrated. As the majority of documents stems from daily newspapers, the number of articles to be organized is too large to allow manual classification. Furthermore, while the topics covered by the various sources overlap, the perspectives from which these issues are presented differ. To a large degeree this can be attributed to the general genre of a source, such as newspapers and magazines specializing on economic issues, but also, to a large degree, to the style of report chosen. In many situations we will find the same topic to be covered by a news report as well as by an interview or a column, or we find the same issue covered both in the general news as well as in, say, the economic section of a paper.

A cleansing procedure was implemented for each data source to automatically remove characteristic formatting structures of the various sources such as banners, footers, or navigation bars, as these would unduely interfere with the stylistic analysis. Furthermore, different HTML encodings for special characters were converted to a uniform representation. The results reported below are based on a subset of the entire collection consisting of 1.000 articles from March 2000. To keep the system as flexible and generally applicable as possible, no language- or domain-specific optimizations, such as stemming or the use of specific stop-word lists, were performed. The articles were parsed to create both a content-based and a structure-based description of the documents, which were further fed into two separate self-organizing maps for cluster analysis and representation. Due to space restrictions we cannot provide detailed representations of the according maps. Rather, we have selected representative clusters for detailed discussion.

6.2 Content-based organization

A 5 × 10 SOM was used for topical organization of the articles based on a 1.975-dimensional feature vector representation. The main topical clusters identified in the collection are, on the one hand, economic articles, which consist of several subclusters, such as a rather dominant group of articles relating to the telecom business, or the privatization of Austria's state-owned enterprises. This cluster is located in the lower left corner of the resulting SOM. On the opposite, upper right corner we find mostly articles covering political issues, such as the discussions concerning the formation of the new government following the 1999 elections. This political cluster basically covers the whole left area of the content-based SOM, moving from the initial elections-based discussions to the various political topics. Another prominent cluster is formed by sports reports covering soccer, formula 1, and horse races, to name a few. Other, smaller clusters address different areas of science, with two of the more prominent sub-clusters among these being devoted to medicine, and internet technologies.

Using the LabelSOM method appropriate labels were automatically extracted, describing the various topical clusters. (The keywords have been translated into English for discussion in the following sections.) We find, for example, one of the clusters representing articles on Austria's Freedom Party to be labelled with *fp, joerg, haider, haiders*, listing the parties abbreviation as well as the name of its political leader. The labels for this unit also demonstrate one of the weaknesses of the crude indexing approach chosen. As we do not apply any language-specific stemming techniques, the trailing genitiv-*s* causes the term *haider* to appear in two forms. Yet, this impreciseness does not cause distortions to the resulting content representation and organization, although language-specific adaptions would further improve the resulting classification, albeit sacrificing the language and domain independence.

This unit is located next to another unit labelled *minister of defence, fpoe, fp, westenthaler, klestil* in the bottom right corner, listing again the freedom party, another one of its leading polititians, Peter Westenthaler, as well as the name of Austria's president, Thomas Klestil. This co-location of similar, yet not identical topics, is one of the most important characteristics of SOMs making them particularly suitable for the organization of document collections for interactive browsing.

Shifting to another topic we find units from the economic cluster in the lower right corner to be labelled with, for example, *austria, stock-exchange, fonds manager, telecom* for the previously mentioned telecom-cluster, or *enterprise, state, va-tech, steel, oeiag, grasser, leitl* for the cluster on the privatization of Austria's state-own steel enterprise VA-Tech, and two of the leading polititians involved in the privatization process, Karl-Heinz Grasser and Christoph Leitl. Above this unit we find a similar topic, namely the privatization of the Austrian postal services, labelled with *privatization, psk, contracts*, nicely showing the topological ordering in the map.

As an example for labels from the sports cluster we might mention the cluster on Formula 1 with labels *races, bmw, williams, jaguar, wm.* Since we do not specify a manually designed stop-word list, some stop-words remain in the list of index terms and actually show up as labels as they form a prominent common feature of articles in a cluster. We thus also find labels such as *friday* and *both* as labels for the sports cluster. Again, a hand–crafted or semi-automatic approach may provide a better removal for stop-words, yet sacrificing domain and language independence to some degree, whereas the current approach can be applied to any given document collection. The cluster representing documents on soccer is labelled with *goal, real, madrid, muenchen, bavaria, rome, group*, listing important soccer clubs playing against each other in a given group of the tournament. Neighboring this unit we find another unit on soccer, this time labelled *champions league, cup, barcelona, madrid, porto.*

This map serves as a content-based index to the digital library, allowing users to find, by reading the labels, which topics are covered in which section of the library. The documents can be represented as located in an HTML table with the labels provided as text in the table cells. They may also be transformed into a graphical libViewer representation, with the according source of the document, i.e. the magazine's title etc. being provided as a logo on the spine. Yet, we do not have any information from the resulting representation whether a given document represents, say, an interview, a result listing etc., as this information is not provided as metatags within the articles.

6.3 Incorporating genre information

6.3.1 A genre map

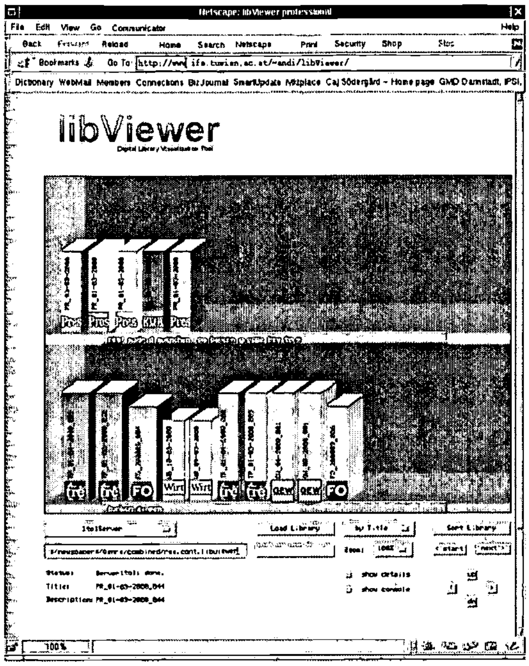
While the content-based SOM provides an organization of articles by their subject, the genre SOM analyzes the structural features of the documents and groups the documents accordingly.

We find, for example, a rather dominant cluster representing various forms of interviews, moving from reports with several quotations in them to long interviews on a given subject. The labels extracted by the LabelSOM technique help us to identify the most distinctive characteristics of a given cluster. For the interviews we find the characteristic attributes to be the number of opening and closing quotes as well as the colon. Further distinctions can be made by the average length of sentences as well as frequent line-breaks, setting interviews apart from articles with longer citations. Another cluster of documents having a high colon count, yet a completely different sentence length structure plus several other special characters occurring frequently in the text, such as opening and closing breaks or slashes, is formed by sports articles providing only result listings. These documents obviously also exhibit an unproportionally high count of numbers. While short reports separate themselves from longer articles due to a number of text length parameters, we find a further distinction within the short reports cluster having a higher count of numbers, yet less than for sports results. These articles are mostly reports or announcements for radio or TV shows, which may be either special documentaries or sports transmissions. Another large cluster of documents consists of legal documents, which set themselves apart by the frequent usage of the paragraph character (§). Internet articles are characterized by the *at-sign* (@).

6.3.2 Integrating genre information into the topical organization

While the labels extracted by the LabelSOM technique help the user interpreting and understanding what the SOM has learned, they are not sufficient to allow the user to intuitively tell which cluster of documents corresponds to which specific genre. This is because the extracted low-level features do not correspond to what the casual user will attribute to characteristic for, say, an interview or sports results listings. Instead of assigning every unit to a specific genre, we map the genre SOM into an RGB-color space, such that documents located in the upper left corner are colored black, whereas the upper right corner is assigned green, the lower left corner red, and the lower right corner yellow. The units inbetween are automatically assigned intermediate colors. According to its structure, each document is now assigned a color based on its location on the genre SOM. This color is used for representing the document in the content-based libViewer representation. We thus may now expect to find documents on the same topic, which are located on the same shelf on the content-based SOM, to be coloured differently if they exhibit a different structure.

Figure 3 shows one shelf from the upper midle area of the SOM representing articles on soccer from the sports section. The general topic of this unit is given as shelf labels. As can be told from the logos, the documents on this unit stem mostly from the daily newspaper *Die Presse*, with only one article being from *Kurier*, another daily newspaper. Also, the average length of all articles on this unit is rather homogenous, with all documents being short articles, thus depicted with only the minimum spine width. Still, we can see from the differing coloring that several distinct types of documents are located on this unit. The first and the last article on this shelf, colored orange, both represent short result reports, listing only the outcomes of the various matches. The second document, colored in bright yellow, contains a rather emotional report, looking more like a transcript from a live broadcast, but not reporting explicitly on results, next to a green document providing a rather factual report on the same match in a rather complicated style. The one-but-last document, colored dark-red, contains a somewhat longer report on several matches, listing not only results, but also short descriptions of various sections of the matches. More important, however, is the fact that it also contains a report on the financial situation of one of the soccer clubs, thus being colored entirely different than the other soccer result reports. This difference in structure, and its partial membership in the economic articles genre, can be attributed to the frequent occurrence of the Euro currency symbol.


Netscape: libViewer professional
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Go To: http://www.ifs.tuwien.ac.at/~andi/libViewer/
Dictionary WebMail Members Connections BizJournal SmartUpdate Mktplace Caj Södergård – Home page GMD Darmstadt, IPSI,
libViewer
Digital Library Visualization Tool
FO
Wirt
OEW
Load Library
by Title
Sort Library
Zoom: 100%
< start
next >
show details
show console
up
dn
100%

Status: Server(tol): done.

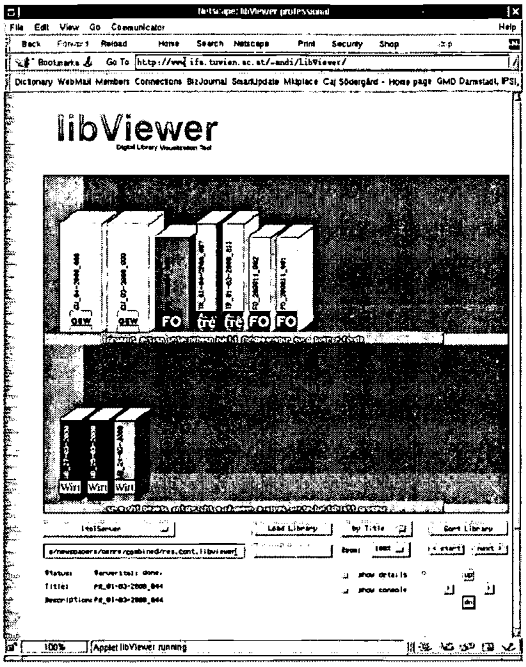
Title: PR_01-03-2000_044

Description: PR_01-03-2000_044

Figure 3: libViewer: sports and economy sections

In the shelf beneath the soccer reports we find documents reporting on financial issues, or more precisely, on interest rates, representing articles from 4 different publications, one of which is a daily newspaper, 1 weekly, and 2 monthly magazines. Except for the weekly magazine *Format*, which is a general news magazine, all publications on this shelf have a strong focus on economic issues. The according labels are given as *banks, interest rates.* Two distinct types of articles can be distinguished on this unit, colored dark brown (the first three documents) and various shadings of green, respectively. When taking a look at the according documents, we find the dark brown ones to be rather complicated, extensive reports on interest rates issues, whereas the green documents are written in a rather informal style. These are mainly made up of short sentences and rethoric questions, and list the most important issues in a tabular form rather than by complex explanations. Please note, that the more complex articles also are longer, as can be seen from the wider spine width.

Figure 4 depicts the next two shelves down the row, continuing in the economic section of the map. Here we find a number of articles on the stock exchange in the upper shelf, whereas the lower shelf contains reports on the economic data from the print magazine *Die Wirtschaftswoche.* The first two documents on the lower shelf are colored in black, and they provide detailed percentual listings of the magazines subscriber structured, their average income etc. The third, green document reports on the same issue, yet rather represents a short overview article describing the general


Netscape: libViewer professional
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop
Bookmarks Go To http://www.ifs.tuwien.ac.at/~andi/LibViewer/
Dictionary WebMail Members Connections BizJournal SmartUpdate Mktplace Caj Södergård - Home page GMD Darmstadt, IPSI,
libViewer
Digital Library Visualization Tool
GEW
GEW
FO
tre
tre
FO
FO
Wirt
Wirt
Wirt
ToolServer
Load Library
by Title
Sort Library
s/newspapers/genre/combined/res.cont.libviewer
Items: 100%
|< start
next >
show details
show console
up
dn
100%
Applet libViewer running

Status: Serverstatus: done.
Title: PR_01-03-2000_044
Description: PR_01-03-2000_044

Figure 4: libViewer: economy-section (cont.)

results from the market study in simple terms. It thus is very similar in structure to the green documents discussed previously.

We also find a number of green documents in the upper shelf, providing short descriptions and buying recomendations for fonds and insurance policies based on fonds. The third, dark-brown document again describes a series of issues related to stock market transaction in a more complex structure, as can be expected from documents exhibiting this color. The first two documents, apart from being rather long, and thus depicted with rather broad spines, differ from the other reports by describing several companies and their performance by citing experts. While being rather detailed, we find many short sentences, enclosed by quotes, as the characteristic features of these documents, thus separating them from the neighboring darker lengthy description.

6.4 Evaluation

Unfortunately, the actual genre of a document cannot be intuitively told by the color it is represented in, nor could we find a way of how the different structural characteristics of documents could be automatically translated into a small set of distinct genres, which could then be represented by more intuitive metaphors (such as papers for interviews, harcover books for lengthy reports and paperbacks for shorter, simpler depictions). Although such a mapping would be possible in principle in a semi-automatic way by assigning different representation metaphors to different areas of the genre SOM, we prefer the automatic mapping of the structural position of a document on the genre SOM into a simple color space. While this metapher needs to be learned to be interpreted correctly, the actually effort required to understand the structure intuitively, rather than explicitly, has shown to be rather small and straight-forward. Furthermore, the chosen approach allows gradual changes between various genres.

No large-scale usability study has yet been performed, although first tests with a small set of users, mostly students, have turned out encouraging results. After visiting a few documents on the respective areas of interest, most people had a feeling of what to expect from a document in a specific color, although they obviously were not able to describe it in terms of the low-level features used for classification by the map. Still, users know what to expect from, say, yellow to ochre documents (interviews, from black (numerical listings), greenish (short, simple articles), or others.

Obviously, the proposed approach to structural analysis will not be able to provide a full-scale genre analysis, capturing the fine differences between certain types of information representation, especially if they involve high-level linguistic analysis. To provide a rather far-fetched example, the presented system will definitely fail to separate a satire from factual information. It thus does not perform genre analysis in the strict sense.

Yet by capturing structural characteristics and similarities clearly is able to uncover specific genre information in given settings. While this might lead to misunderstandings in some situations, such as the impossibility of telling factual from fictional information (genre-wise), it should provide considerable support to the user trying to satisfy an information need. This is especially true as the utilization of the self-organizing map to produce a topology-preserving mapping allows to capture gradual differences between various structural concepts in a straight-forward manner.

7. CONCLUSIONS AND FUTURE WORK

Providing structural information about documents is essential to help users decide about the relevance of documents available in a digital library. Most document collections thus try to convey this information by using carefully designed metadata describing the genre of a resource. However, in many cases this uniform description of documents cannot be provided manually. This is especially true for digital libraries integrating documents from different sources, or where the number of documents to be described effectively prohibits manual classification.

In this paper we presented an automated approach to the structural analysis of text documents. Characteristic features such as the average length of a sentence, counts of punctuation marks and other special characters, as well as specific words such as pronouns etc. can be used to describe the structural characteristics of a document. The self-organizing map, a popular unsupervised neural network, is used to cluster the documents according to their similarity. Documents are then colored according to their location on the resulting two-dimensional map, such that structurally similar documents are colored similarly.

The result of the structural analysis is further incorporated into the content-based organization and representation of a digital library provided by the SOMLib system. Documents on the same topic, yet providing a different perspective of the same subject, such as reports and interviews, complex analyses, or short descriptions, are thus shown as books of different colors in the resulting graphical representation provided by the libViewer interface.

Initial experiments have shown encouraging results. In the next steps we would like to refine the mapping from the position on the structural SOM onto an appropriate color in such a way that the mutual similarity or distance of two units is reflected in the perceived distance of the colors assigned to the according documents. This would allow the structural similarities between, say, different types of interviews, to be more evident by assigning somewhat more similar colors to documents that are part of a larger cluster consisting of several sub-clusters. Such an improved mapping can be achieved by using distance information provided by the weight vectors of the units in the self-organizing map. Furthermore, a more suitable color space in terms of human perception may be chosen to further increase the perceived similarities and dissimilarities.

Secondly, we will take a closer look at additional features that offer themselves for genre analysis. First experiments indicate that, for example, a distinction between fact-reporting articles versus opinion-forming articles is possible by including additional keywords in the list of features.

8. REFERENCES

[1] D. Biber. *Variations across Speech and Writing.* Cambridge University Press, UK, 1988.

[2] D. Biber. A typology of english texts. *Linguistics*, 27:3 – 43, 1989.

[3] I. Bretan, J. Dewe, A. Hallberg, N. Wolkert, and J. Karlgren. Web-specific genre visualization. In *Proc of WebNet '98*, Orlando, FL, November 1998. http://www.stacken.kth.se/~dewe/.

[4] H. Chen, C. Schuffels, and R. Orwig. Internet categorization and search: A self-organizing approach. *Journal of Visual Communication and Image Representation*, 7(1):88–102, 1996. http://ai.BPA.arizona.edu/papers/.

[5] H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal American Statistical Association*, 68:361–368, 1973.

[6] L. Cherra and W. Vesterman. Writing tools: The STYLE and DICTION programs. Technical Report 91, Bell Laboratories, Murray Hill, NJ, 1981. Republished as part 4.4BSD User's Supplementary Documents by O'Reilly.

[7] J. Himberg. A SOM based cluster visualization and its application for false coloring. In *Proc Int'l Joint Conf on Neural Networks (IJCNN 2000)*, Como, Italy, July 24. – 27. 2000. IEEE Computer Society.

[8] J. Karlgren. Stylistic experiments in information retrieval. In T. Strzalkowski, editor, *Natural Language Information Retrieval.* Kluwer, 1999. http://www.sics.se/~jussi/Artiklar/.

[9] J. Karlgren, I. Bretan, J. Dewe, A. Hallberg, and N. Wolkert. Iterative information retrieval using fast clustering and usage-specific genres. In *Proc Eighth DELOS Workshop on User Interfaces in Digital Libraries*, pages 85–92, Stockholm, Sweden, October 1998. http://www.stacken.kth.se/~dewe/.

[10] J. Karlgren and D. Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proc 15. Int'l Conf on Computational Linguistics (COLING '94)*, Kyoto, Japan, 1994. http://www.sics.se/~jussi/Artiklar/.

[11] B. Kessler, G. Nunberg, and H. Schütze. Automatic detection of text genre. In *Proc 8. Conf Europ. Chapter of the Association for Computational Linguistics (ACL/EACL97)*, pages 32–38, Madrid, Spain, 1997. http://spell.psychology.wayne.edu/~bkessler/.

[12] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 1982.

[13] T. Kohonen. *Self-organizing maps.* Springer-Verlag, Berlin, 1995.

[14] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self-organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585, May 2000. http://ieeexplore.ieee.org/.

[15] D. Merkl and A. Rauber. Document classification with unsupervised neural networks. In F. Crestani and G. Pasi, editors, *Soft Computing in Information Retrieval*, pages 102–121. Physica Verlag, 2000. http://www.ifs.tuwien.ac.at/~andi/LoP.html.

[16] A. Rauber. LabelSOM: On the labeling of self-organizing maps. In *Proc Int'l Joint Conf on Neural Networks (IJCNN'99)*, Washington, DC, July 10. - 16. 1999. http://www.ifs.tuwien.ac.at/~andi/LoP.html.

[17] A. Rauber. SOMLib: A digital library system based on neural networks. In E. Fox and N. Rowe, editors, *Proc ACM Conf on Digital Libraries (ACMDL'99)*, pages 240–241, Berkeley, CA, August 11. - 14. 1999. ACM. http://www.acm.org/dl.

[18] A. Rauber and H. Bina. Visualizing electronic document repositories: Drawing books and papers in a digital library. In *Advances in Visual Database Systems: Proc IFIP TC2 Working Conf on Visual Database Systems*, pages 95 – 114, Fukuoka, Japan, May, 10.- 12. 2000. Kluwer Academic Publishers. http://www.ifs.tuwien.ac.at/~andi/LoP.html.

[19] A. Rauber and D. Merkl. The SOMLib Digital Library System. In*Proc 3. Europ. Conf on Research and Advanced Technology for Digital Libraries (ECDL99)*, LNCS 1696, pages 323–342, Paris, France, September 22. - 24. 1999. Springer. http://www.ifs.tuwien.ac.at/~andi/LoP.html.

[20] A. Rauber, M. Dittenbach, and D. Merkl. Automatically detecting and organizing documents into topic hierarchies: A neural-network based approach to bookshelf creation and arrangement. In *Proc 4. Europ. Conf on Research and Advanced Technologies for Digital Libraries (ECDL2000)*, LNCS 1923, pages 348–351, Lisboa, Portugal, September 18. - 20. 2000. Springer. http://www.ifs.tuwien.ac.at/~andi/LoP.html.

[21] K. Ries. Towards the detection and description of textual meaning indicators in spontaneous conversations. In *Proc Europ. Conf on Speech Communication and Technology (EUROSPEECH99)*, Budapest, Hungary, September 5-9 1999.

[22] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, Reading, MA, 1989.

Text Categorization for Multi-page Documents: A Hybrid Naive Bayes HMM Approach

Paolo Frasconi
Department of Systems and Computer Science
University of Florence
50139 Firenze, Italy
paolo@dsi.unifi.it

Giovanni Soda
Department of Systems and Computer Science
University of Florence
50139 Firenze, Italy
giovanni@dsi.unifi.it

Alessandro Vullo
Department of Systems and Computer Science
University of Florence
50139 Firenze, Italy
alex@mcculloch.ing.unifi.it

ABSTRACT

Text categorization is typically formulated as a concept learning problem where each instance is a single isolated document. In this paper we are interested in a more general formulation where documents are organized as page sequences, as naturally occurring in digital libraries of scanned books and magazines. We describe a method for classifying pages of sequential OCR text documents into one of several assigned categories and suggest that taking into account contextual information provided by the whole page sequence can significantly improve classification accuracy. The proposed architecture relies on hidden Markov models whose emissions are bag-of-words according to a multinomial word event model, as in the generative portion of the Naive Bayes classifier. Our results on a collection of scanned journals from the Making of America project confirm the importance of using whole page sequences. Empirical evaluation indicates that the error rate (as obtained by running a plain Naive Bayes classifier on isolated page) can be roughly reduced by half if contextual information is incorporated.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*Learning*; H.3.7 [**Information Systems**]: Information Storage and Retrieval—*Digital Libraries*; I.7.m [**Computing Methodologies**]: Document and Text Processing

General Terms

Algorithms, Performance

Keywords

Text categorization, Hidden Markov Models, Naive Bayes classifier, Multi-page documents

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

1. INTRODUCTION

Text categorization is the problem of grouping textual documents into different fixed classes or categories. The task is related to the ability of an intelligent system to automatically perform tasks such as personalized e-mail or news filtering, document indexing, metadata extraction. These problems are of great and increasing importance, mainly because of the recent explosive increase of online textual information. Text categorization is generally formulated in the machine learning framework. In this setting, a learning algorithm takes as input a set of labeled examples (where the label indicates which category the example document belongs to) and attempts to infer a function that will map new documents into their categories. Several algorithms have been proposed within this framework, including regression models [29], inductive logic programming [6], probabilistic classifiers [17, 21, 16], decision trees [18], neural networks [22], and more recently support vector machines [12].

Research on text categorization has been mainly focused on non-structured documents. In the typical approach, inherited from information retrieval, each document is represented by a sequence of words, and the sequence itself is normally flattened down to a simplified representation called *bag of words* (BOW). This is like representing each document as a feature-vector, where features are words in the vocabulary and components of the feature-vector are statistics such as word counts in the document. Although such a simplified representation is appropriate for relatively flat documents (such as email and news messages), other types of documents are internally structured and this structure should be exploited in the representation to better inform the learner.

In this paper we are interested in the domain of digital libraries and, in particular, collections of digitized books or magazines, with text extracted by an Optical Character Recognition (OCR) system. One important challenge for digital conversion projects is the management of structural and descriptive metadata. Currently, metadata management involves a large amount of keying work carried out by human operators. Automating the extraction of metadata from digitized documents could greatly improve efficiency and productivity [1]. This automation, however, is not a trivial task and involves recognition of the ordering of text divisions, such as chapters and sub-chapters, the identification of layout elements, such as headlines, footnotes, graphs, and captions, and the linking of articles within a periodical. Automatic recognition of these elements can be a hard problem, especially without any prior knowledge about the type of elements that are expected to be present within a given document page. Hence, page classification can represent a useful preliminary step to guide the subsequent extraction process. Moreover, extracting metadata related to the semantic contents of document parts (such as chapters or articles) can require the ability of recognizing the topic or the category of these parts. The solution to these problems can be helped by a classifier that assigns a category to each page of the document.

Unlike email or news articles, books and periodicals are *multi-page* documents and the simplest level of structure that can be exploited is the serial order relation defined among single pages. The task we consider is the automatic categorization of each page according to its (semantic) contents[1]. Exploiting the serial order relation among pages within a single document can be expected to improve classification accuracy when compared to a strategy that simply classifies each page separately. This is because the sequence of pages in documents such as books or magazines often follows regularities such as those implied by typographical and editorial conventions. Consider for example the domain of books and suppose categories of interest include `title-page`, `dedication-page`, `preface-page`, `index-page`, `table-of-contents`, `regular-page`, and so on. Even in this very simplified case we can expect constraints about the valid sequences of page categories in a book. For example, `title-page` is very unlikely to follow `index-page` and, similarly, `dedication-page` is more likely to follow `title-page` than `preface-page`. Constraint of this type can be captured and modeled using a stochastic grammar. Thus, information about the category of a given page can be gathered not only by examining the contents of that page, but also by examining the contents of other pages in the sequence. Since contextual information can significantly help to disambiguate between page categories, we expect that classification accuracy will improve if the learner has access to whole sequences instead that single-page documents.

In this paper we combine several algorithmic ideas to solve the problem of text categorization in the domain of multi-page documents. First, we use an algorithm similar to those described in [28] and [20] for inducing a stochastic regular grammar over sequences of page categories. Second, we introduce a hidden Markov model (HMM) that can deal with sequences of BOWs. Each state in the HMM is associated with a unique page category. Emissions are modeled by a multinomial distribution over word events, like in the generative component of the Naive Bayes classifier. The HMM is trained from (partially) labeled page sequences, i.e. state variables are partially observed in the training set. Unobserved states (which is the common setting in most classic applications of HMMs) arise here when document pages are partially unlabeled, like in the framework described in [23] and [13]. Finally, we solve the categorization problem by running the Viterbi algorithm on the trained HMM, yielding a sequence of page categories associated with new (unseen) documents. This is somewhat related to recent applications of HMMs to information extraction [9, 20] but the output labeling in our case is associated with the entire stream of text contained into a page, while in [9, 20] the HMM is used to attach labels to single words of shorter portions of text.

Our approach is validated on a real dataset consisting of 95 issues of the journal *American Missionary*, which is part of the "Making of America" collection [26]. In spite of text noise due to optical recognition, our system achieves about 85% page classification accuracy when training on 10 issues (year 1884) and testing on issues from 1885 to 1893. More importantly, we show that incorporating contextual information significantly reduces classification error, both in the case of completely labeled example documents and when unlabeled documents are included in the training set.

2. BACKGROUND

Let d be a generic multi-page document, and let d_t denote the t-th page within the document. The categorization task consists of learning from examples a function $f : d_t \to \{c^1, \cdots, c^K\}$ that maps each page d_t into one out of K classes.

2.1 The Naive Bayes classifier

The above task can also be reformulated in probabilistic terms as the estimation of the conditional probability $P(C_t = c^k|d_t)$, C_t being a multinomial class variable with realizations in $\{c^1, \cdots, c^K\}$. In so doing, f can be computed using Bayes' decision rule, i.e. $f(d)$ is the class with higher posterior probability. The Naive Bayes classifier computes this probability as

$$P(C_t = c^k|d_t) \propto P(d_t|C_t = c^k)P(C_t = c^k). \qquad (1)$$

What characterizes the model is the so-called Naive Bayes assumption, prescribing that word events (each occurrence of a given word in the page corresponds to one event) are conditionally independent *given* the page category. As a result, the class conditional probabilities can be factorized as

$$P(d_t|C_t = c^k) = \prod_{i=1}^{|d_t|} P(w_t^i|C_t = c^k) \qquad (2)$$

where $|d_t|$ denotes the length of page d_t and w_t^i is the i-th word in the page. This conditional independence assumption is graphically represented by the Bayesian network[2] shown in Figure 1.

Although the basic assumption is clearly false in the real world, the model works well in practice since classification requires finding a good separation surface, not necessarily a very accurate model of the involved probability distributions. Training consists of estimating model's parameters from a dataset $\mathcal{D}$ of labeled documents (see, e.g. [21]).

2.2 Hidden Markov models

HMMs have been introduced several years ago as a tool for probabilistic sequence modeling. The interest in this area developed particularly in the Seventies, within the speech

[1] A related formulation would consist of assigning a global category to a whole multi-page document, but this formulation is not considered in this paper.

[2] A Bayesian network is an annotated graph in which nodes represent random variables and *missing* edges encode conditional independence statements amongst these variables. Given a particular state of knowledge, the semantics of a Bayesian networks determine whether collecting evidence about a set of variables does modify one's belief about some other set of variables [24, 11].

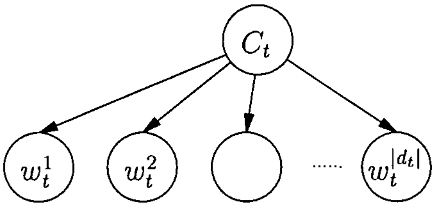


Figure 1: Bayesian network for the Naive Bayes classifier.

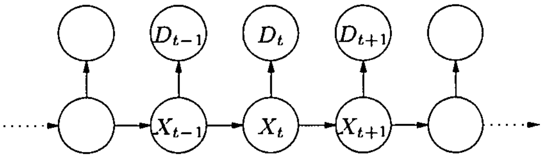


Figure 2: Bayesian networks for standard HMMs.

recognition research community [25]. During the last years a large number of variants and improvements over the standard HMM have been proposed and applied. Undoubtedly, Markovian modeling is now regarded as one of the most significant state-of-the-art approaches for sequence learning. Besides several applications in pattern recognition and molecular biology, HMMs have been also applied to several text related tasks, including natural language modeling [5] and, more recently, information retrieval and extraction [9, 20]. The recent view of the HMM as a particular case of Bayesian networks [2, 19, 27] has helped the theoretical understanding and the ability to conceive extensions to the standard model in a sound and formally elegant framework.

An HMM describes two related discrete-time stochastic processes. The first process pertains to hidden discrete state variables, denoted X_t, forming a first-order Markov chain and taking realizations on a finite alphabet $\{x^1, \cdots, x^N\}$. The second process pertains to observed variables or *emissions*, denoted D_t. Starting from a given state at time 0 (or given an initial state distribution) the model probabilistically transitions to a new state X_1 and correspondingly emits observation D_1. The process is repeated recursively until an end state is reached. Note that, as this form of computation may suggest, HMMs are closely related to stochastic regular grammars [5]. The Markov property prescribes that X_{t+1} is conditionally independent of $X_1, \ldots, X_{t-1}$ given X_t. Furthermore, it is assumed that D_t is independent of the rest given X_t. These two conditional independence assumptions are graphically depicted using the Bayesian network of Figure 2. As a result, an HMM is fully specified by the following conditional probability distributions[3]:

$$\begin{array}{ll} P(X_t|X_{t-1}) & \text{(transition distribution)} \\ P(D_t|X_t) & \text{(emission distribution)} \end{array} \quad (3)$$

Since the process is stationary, the transition distribution can be represented as a square probability matrix whose entries are transition probabilities $P(X_t = x^i|X_{t-1} = x^j)$, abbreviated as $P(x^i|x^j)$ in the following. In the classic literature, emissions are restricted to symbols in a finite alphabet or multivariate continuous variables [25]. As explained in the next section, our model allows emissions to be bag-of-words.

3. THE MULTI-PAGE CLASSIFIER

We now turn to the description of our classifier for multi-page documents. This section presents the architecture and the algorithms for grammar extraction, training, and classification.

3.1 Architecture

In our case, HMM emissions are associated with entire pages of the document. Thus the realizations of the observation D_t are bag-of-words representing the text in the t-th page of the document. Within our framework, states are related to pages categories by a a deterministic function ϕ that maps state realizations into page categories. We assume that ϕ is a surjection but not a bijection, i.e. that there are more state realizations than categories. This enriches the expressive power of the model, allowing different transition behaviors for pages of the same class, depending on where the page is actually encountered within the sequence. However, if the page *contents* depends on the category but not on the context of the category within the sequence[4], the use of multiple states per category may introduce too many free parameters and it may be convenient to assume that

$$P(D_t|x^i) = P(D_t|x^j) = P(D_t|c^k) \text{ if } \phi(x^i) = \phi(x^j) = c^k. \quad (4)$$

This assumption constrains emission parameters to be the same for all the HMM states labeled by the same page category, a form of parameters sharing that may help to reduce overfitting. The emission distribution is then defined as for the Naive Bayes classifier, i.e. for every observed page d_t

$$P(d_t|c^k) = \prod_{i=1}^{|d_t|} P(w_t^i|c^k) \quad (5)$$

Therefore, the architecture can be graphically described as the merging of the Bayesian networks for HMMs and Naive Bayes, as shown in Figure 3. We remark that the state (and hence the category) at page t depends not only on the contents of the page, but also on the contents of other pages in the document. This probabilistic dependency implements

[3] We adopt the standard convention of denoting variables by uppercase letters and realizations by the corresponding lowercase letters. Moreover, we use the table notation for probabilities as in [11]; for example $P(X)$ is a shorthand for the table $[P(X{=}x^1), \ldots, P(X{=}x^r)]$ and $P(X, y|Z)$ denotes the two-dimensional table with entries $P(X{=}x^i, Y{=}y|Z{=}z^k)$.

[4] Of course this does not mean that the *category* is independent on the context.

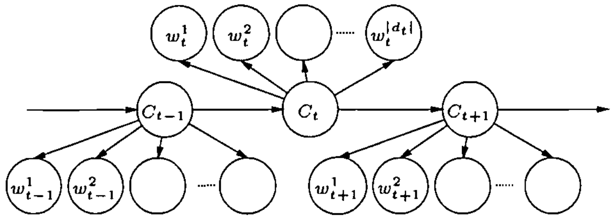


Figure 3: Bayesian network for the hybrid HMM Naive Bayes architecture.

the mechanism for taking contextual information into account.

The algorithms used in this paper are derived from the literature on Markov models [25], inference and learning in Bayesian networks [24, 11, 10], and classification with Naive Bayes [17, 15]. In the following we sketch the main issues related to the integration of all these methods.

3.2 Induction of HMM topology

The *structure* or topology of an HMM is a representation of the allowable transitions between hidden states. More precisely, the topology is described by a directed graph whose vertices are state realizations $\{x^1, \ldots, x^N\}$, and whose edges are the pairs (x^j, x^i) such that $P(x^i|x^j) \neq 0$. An HMM is said to be *ergodic* if its transition graph is fully-connected. However, in almost all interesting application domains, less connected structures are better suited for capturing the observed properties of the sequences being modeled, since they convey domain prior knowledge. Thus, starting from the right structure is an important problem in practical hidden Markov modeling. As an example, consider Figure 4, showing a (very simplified) graph that describes transitions between the parts of a hypothetical set of books. Possible state realizations are[5] {start, title, dedication, preface, toc, regular, index, end }. The structure indicates, among other things, that only dedication, preface, or table of contents can follow the title page. Self-loops indicate that a given category can be repeated for several consecutive pages. While a structure of this kind could be hand-crafted by a domain expert, it is may be more advantageous to learn it automatically from data.

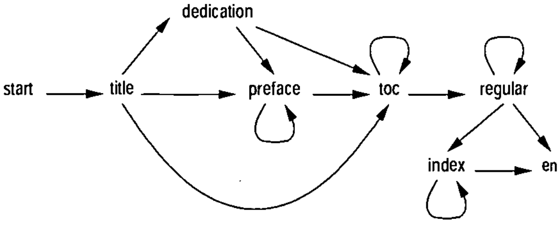


Figure 4: Example of HMM transition graph.

We now briefly describe the solution adopted to automatically infer HMM transition graphs from sample multi-page documents. Let us assume that all the pages of the available training documents are labeled with the class they belong to. One can then imagine to take advantage of the observable distribution of data to search for an effective structure in the space of HMMs topologies. Our approach is based on the application of an algorithm for data-driven model induction adapted from previous works in Bayesian HMM induction [28] and construction of HMMs of text phrases for information extraction [20]. The algorithms starts by building a structure that is capable only to "explain" the available training sequences (a maximally specific model). The initial structure includes as many paths (from the initial state to the final one) as there are training sequences. Every path is associated with one sequence of pages, i.e. a distinct state is created for every page in the training set. Each state x is labeled by $\phi(x)$, the category of the corresponding page in the document. Note that, unlike the example shown in Figure 4, several states are generated for the same category. The algorithm then iteratively applies merging heuristics that collapse states so as to augment generalization capabilities over unseen sequences. The first heuristic, called neighbor-merging, collapse two states x and x' if they are neighbors in the graph and $\phi(x) = \phi(x')$. The second heuristic, called V-merging, collapses two states x and x' if $\phi(x) = \phi(x')$ and they share a transition from or to a common state, thus reducing the branching factor of the structure.

3.3 Inference and learning

Given the HMM topology extracted by the algorithm described above, the learning problem consists of determining transition and emission parameters. One important distinction that need to be made when training Bayesian network is whether or not all the variables are observed. Assuming complete data (all variables observed), maximum likelihood estimation of the parameters could be solved using a one-step algorithm that collects sufficient statistics for each parameter [10]. In our case, data are complete if and only if the following two conditions are met:

1. there is a one-to-one mapping between HMM states and page categories (i.e. $N = K$ and for $k = 1, \ldots, N$, $\phi(x^k) = c^k$), and
2. the category is known for each page in the training documents, i.e. the dataset consists of sequences of pairs $(\{d_1, c_1^*\}, \ldots, \{d_T, c_T^*\})$, c_t^* being the (known) category of page t and T being the number of pages in the document.

Under these assumptions, estimation of transition parameters is straightforward and can be accomplished as follows:

$$P(x^i|x^j) = \frac{N(c^i, c^j)}{\sum_{\ell=1}^{N} N(c^\ell, c^j)} \tag{6}$$

where $N(c^i, c^j)$ is the number of times a page of class c^i follows a page of class c^j in the training set. Similarly, estimation of emission parameters in this case would be accomplished exactly like in the case of the Naive Bayes classifier (see, e.g. [21]):

$$P(w^\ell|c^k) = \frac{1 + N(w^\ell, c^k)}{|V| + \sum_{j=1}^{|V|} N(w^j, c^k)} \tag{7}$$

[5] Note that in this simplified example ϕ is a one-to-one mapping.

where $N(w^\ell, c^k)$ is the number of occurrences of word w^ℓ in pages of class c^k and $|V|$ is the vocabulary size ($1/|V|$ corresponds to a Dirichlet prior over the parameters and plays a regularization role for whose words which are very rare within a class).

Conditions 1 and 2 above, however, are normally not satisfied. First, in order to model more accurately different contexts in which a category may occur, it may be convenient to have multiple distinct HMM states for the same page category. Second, labeling pages in the training set is a time consuming process that needs to be performed by hand and it may be important to use also unlabeled documents for training [13, 23]. This means that label c_t^* may be not available for some t. If assumption 2 is satisfied but assumption 1 is not, we can derive the following approximated estimation formula for transition parameters:

$$P(x^i|x^j) = \frac{N(x^i, x^j)}{\sum_{\ell=1}^{N} N(x^\ell, x^j)} \tag{8}$$

where $N(x^i, x^j)$ counts how many times state x^i follows x^j during the state merge procedure described in Section 3.2. However, in general, the presence of hidden variables requires an *iterative* maximum likelihood estimation algorithm, such as gradient ascent or expectation-maximization (EM). Our implementation uses the EM algorithm, originally formulated in [7] and usable for any Bayesian network with local conditional probability distributions belonging to the exponential family [10]. Here the EM algorithm essentially reduces to the Baum-Welch form [25] with the only modification that some evidence is entered into state variables. State evidence is taken into account in the E-step by changing forward propagation as follows:

$$\alpha_t(j) = \begin{cases} 0 & \text{if } \phi(x^j) \neq c_t^* \\ \sum_{i=1}^{N} \alpha_{t-1}(i) P(x^j|x^i) P(d_t|x^j) & \text{otherwise} \end{cases} \tag{9}$$

where $\alpha_t(i) = P(d_1 d_2 \cdots d_t, X_t = x^i)$ is the forward variable in the Baum-Welch algorithm.

The M-step is performed in the standard way for transition parameters, by replacing counts in Equation 6 with their expectations given all the observed variables. Emission probabilities are also estimated using expected word counts. If parameters are shared as indicated in Equation 4, these counts should be summed over states having the same label. Thus in the case of incomplete data, Equation 7 is replaced by

$$P(w^\ell|c^k) = \frac{S + \sum_p \sum_t N(w^\ell, c^k) \sum_{i:\phi(x^i)=c^k} P(x^i|d_t)}{S|V| + \sum_{j=1}^{|V|} \sum_p \sum_t N(w^j, c^k) \sum_{i:\phi(x^i)=c^k} P(x^i|d_t)}$$

where S is the number of training sequences, $N(w^\ell, c^k)$ is the number of occurrences of word w^ℓ in pages of class c^k, and $P(x^i|d_t)$ is computed by the Baum-Welch procedure during the E-step. The sum on p extends over training sequences, while the sum on t extends over pages of the p-th document in the training set. The E- and M-steps are iterated until a local maximum of the (incomplete) data likelihood is reached.

It is interesting to point out a related application of the EM algorithm for learning from labeled and unlabeled documents [23]. In that paper the only concern was to allow the learner to take advantage of unlabeled documents in the training set. As a major difference, the method in [23] assumes flat single-page documents and, if applied to multipage documents, would be equivalent to a zero-order Markov model that cannot take into account contextual information.

3.4 Page classification

Given a document of T pages, classification is performed by first computing the sequence of states $\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_T$ that was most likely to have generated the observed sequence of pages, and then mapping each state to the corresponding category $\phi(\hat{x}_t)$. The most likely state sequence can be obtained by running the an adapted version of Viterbi's algorithm, whose more general form is the max-propagation algorithm for Bayesian networks described in [11].

3.5 Feature selection

Text pages should be first preprocessed with common information retrieval techniques, including stemming and stop words removal. Still, the bag-of-words representation of pages can lead to a very high-dimensional feature space corresponding to the vocabulary extracted from training documents. A high-dimensional feature space, especially in this case where features are noisy because of OCR errors, may lead to the overfitting phenomenon: the learner has very high accuracy on the training set but generalization to new examples is poor. Feature selection is a technique for limiting overfitting by removing non-informative words from documents. In our experiments we performed feature selection using information gain [30]. This criterion is often employed in different machine learning contexts. It measures the average number of bits of information about the category that are gained by including a word in a document. For each dictionary term w, the gain is defined as

$$\begin{aligned} G(w) &= -\sum_{k=1}^{K} P(c^k) \log_2 P(c^k) \\ &+ P(w) \sum_{k=1}^{K} P(c^k|w) \log_2 P(c^k|w) \\ &+ P(\overline{w}) \sum_{k=1}^{K} P(c^k|\overline{w}) \log_2 P(c^k|\overline{w}) \end{aligned}$$

where $\overline{w}$ denotes the absence of word w. Feature selection is performed by retaining only the words having the highest average mutual information with the class variable. OCR errors, however, can produce very noisy features which may be responsible of poor performance even if feature selection is performed. For this reason, it may be convenient to prune from the dictionary (before applying the information gain criterion) all the words occurring in the training set with a frequency below a given threshold h.

3.6 Learning with labeled and unlabeled pages

Creating a training set for text categorization involves hand labeling in order to assign a category to each document. Since this is an expensive human activity, it is interesting to evaluate a classification system when only a fraction of the training documents pages are labeled, while other documents are used without a category label. Clearly, unlabeled documents are available at very low cost. In the case of isolated page classification, previous research has demonstrated that learners such as Naive Bayes and support vector machines can take advantage of the inclusion in the training set of documents whose class is unknown [13, 23]. In particular, the method presented in [23] uses EM to deal with unobserved labels.

In the case of multi-page documents, the presence of missing labels means that some pages of the training document sequences have no assigned category. The architecture introduced in this paper (see Figure 3) can easily handle the presence of unlabeled pages in the training set. Basically, evidence is entered into the states of the HMM chain only for those pages for which a label is known, while other state variables are left unobserved. The belief propagation algorithm is in charge of computing probabilities for these hidden variables.

However, the structure learning algorithm presented in Section 3.2 cannot be applied in the case of partially labeled documents. Instead, it is possible to use ergodic (fully connected) HMMs and deriving a transition structure by pruning, after the learning phase, those transitions having small probabilities with respect to an assigned threshold. In this way, we let EM derive a specific structure for the model (note that the only alternative in the case of partially labeled documents would be to obtain a transition graph from a domain expert).

4. EXPERIMENTAL RESULTS

A preliminary evaluation of our system has been conducted in a digital library domain where data are naturally organized in the form of page sequences. The main purpose of our experiments was to make a comparison between our multi-page classification approach and a traditional isolated page classification system.

4.1 Data Set

We have chosen to evaluate the model over a subset of the Making of America (MOA) collection, a joined project between the University of Michigan and Cornell University (see `moa.umdl.umich.edu/about.html` and [26]) for collecting and making available digitized books and periodicals about history and evolution processes of the American society between the XIX and XX century. Presently, the whole archive contains electronic versions of important magazines of the XIX century. In our experiments, we selected a subset of the journal *American Missionary* (AMis), a sociological magazine with strong Christian guidelines. The task consists of correctly classifying pages of previously unseen documents into one of the ten categories described in Table 1. Most of these categories are related to the topic of the articles, but some are related to the parts of the journal (i.e. Contents, Receipts, and Advertisements). The dataset we selected contains 95 issues from 1884 to 1893, for a total of 3222 OCR text pages. Special issues and final report issues (typically November and December issues) have been removed from the dataset as they contain categories not found in the rest. The first year was selected as training set (10 training sequences, 342 pages). The remaining documents (from 1885 to 1993, for a total of 2880 pages) were used as a test set. The ten categories are temporally stable over the 1883–1893 time period.

Name	Description
1. Contents	Cover and index of surveys
2. Editorial	Editorial articles
3. The South	Afro-Americans' survey
4. The Indians	American Indians' survey
5. The Chinese	Reports from China missions
6. Bureau of Women's Work	Female conditions
7. Children's Page	Education and childhood
8. Communications	Magazine informations
9. Receipts	Lists of founders
10. Advertisements	contents is mostly graphic

Table 1: Categories in the *American Missionary* domain.

Category labels were obtained semi-automatically, starting from the MOA XML files supplied with the documents collection. The assigned category was then manually checked. In the case of pages containing the end and the beginning of two articles belonging to different categories, the page was assigned the category of the ending article.

Each page within a document is represented as a Bag-of-Words, counting the number of word occurrences within the page. It is worth remarking that in this application instances are text documents obtained by an OCR process. Imperfections of recognition algorithm and the presence of images in some pages yields noisy text, containing misspelled or nonexistent words, and trash characters (see [3] for a report of OCR accuracy in the MOA digital library). Although these errors may negatively affect the learning process and subsequent results in the evaluation phase, we made no attempts to correct and filter out misspelled words, except for the feature selection process described above. However, since OCR extracted documents preserve the text layout found in the original image, it was necessary to rejoin words that had been hyphenated due to line breaking.

4.2 Feature selection and isolated page classification

The purpose of the experiments in this section is to investigate the effects of feature selection and to assess the baseline prediction accuracy that can be attained using the Naive Bayes classifier on isolated pages. In a set of preliminary evaluations we have found that best performance are achieved by pruning words with less than $h = 10$ occurrences and then selecting an optimal set of informative words. We performed several tests by changing the information gain threshold that determines if a word is sufficiently informative (see Section 3.5), resulting in different vocabulary sizes with different accuracy of prediction. For each reduced vocabulary size we ran the Naive Bayes classifier on isolated pages. Results are shown in Figure 5. Vocabulary size ranges from 15635 words (no feature selection), yielding 65.07% classification accuracy, to 25 words, yielding 53.16% accuracy. The optimal vocabulary size is 297 words, obtained with a threshold gain of 0.089, yielding the best test-set accuracy of 72.57%. This result (72.57%) was considered as the base measure for performance comparison between our model and the Naive Bayes classifier.

4.3 Sequential page classification

Using the hybrid model presented in Section 3, documents can be organized into ordered sequences of pages. The training set contains 10 sequences (monthly issues) of the same

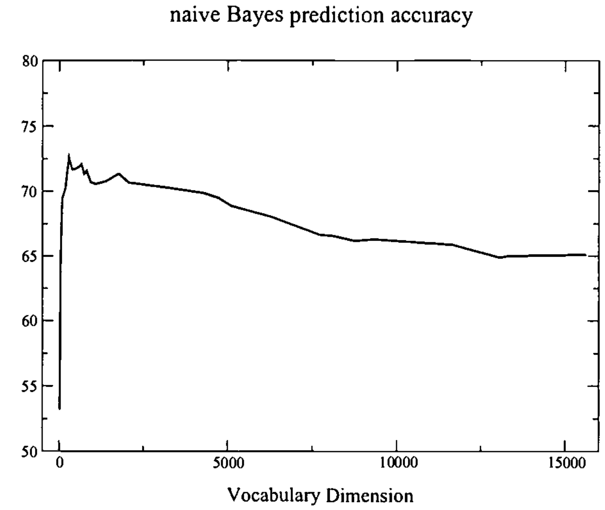


Figure 5: Naive Bayes accuracy as a function of vocabulary size (information gain criterion). Optimal vocabulary size is 297 words.

Category	Sequential	Isolated	Error red.
Contents	**100**	100	0%
Editorial	**80.9**	63.11	48.2%
The South	**90.81**	71.84	67.4%
The Indians	**61.07**	44.3	30.1%
The Chinese	**69.93**	60.78	23.3%
Bureau W.W.	**74.73**	66.3	25.0%
Children's Page	**78.26**	45.65	60.0%
Communications	**93.55**	92.47	14.3%
Receipts	**98.31**	98.31	0%
Advertisements	**90.7**	62.79	75.0%
Total Accuracy	**85.28**	72.57	46.3%

Table 2: Isolated classification (using the best Naive Bayes) vs. sequential classification (using the hybrid HMM with model merging).

342 documents for year 1884, while test set is organized into 85 sequences for a total of 2880 documents from year 1885 to 1893. The bag-of-words representation of pages fed into the HMM classifier was identical to that previously used with Naive Bayes (including preprocessing and feature selection with a vocabulary of 297 words). We have considered two settings for validating the system. In the first setting, it is assumed that category labels c_t^* are available for all the pages in the training set. In the second setting, some category labels are held out and training uses labeled and unlabeled pages.

4.3.1 Completely labeled documents

In the case of completely labeled documents it is possible to run the structure learning algorithm presented in Section 3.2. Figure 6 reports the structure learned from the 10 training issues. Each vertex in the transition graph is associated with one HMM state and is labeled with the corresponding category (see Table 1). Edges are labeled with the transition probability from source to target state, computed by counting state transitions during the state merging procedure (see Equation 8). The associated stochastic

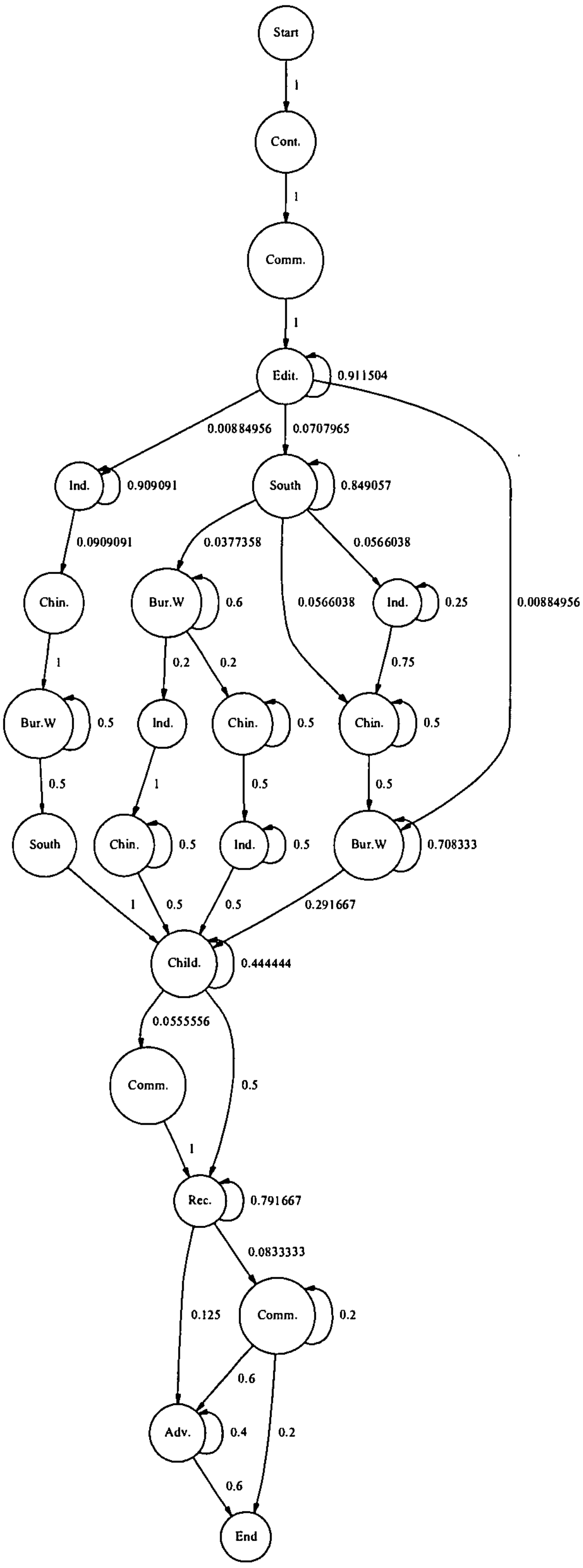


Figure 6: Data induced HMM topology for American Missionary, year 1884.

grammar implies that valid AMis sequences ought to start with the index page (class "Contents"), followed by a page of general communications. Next state is associated with a page of an editorial article. Self transition here has a value of 0.91, meaning that with high probability the next page will belong to the editorial too. With lower probability (0.07) next page is one of the "The South" survey or (prob. 0.008) "The Indians" or "Bureau of Women's work". Continuing this way we can associate a probability to each string of page categories. Since our purpose is to predict the correct string of categories, a good grammar helps filtering out classification hypothesis which generate low (or zero) probability strings. Note that under the parameter sharing assumption (see Equation 4), once the HMM structure is given, an estimate of the emission probabilities can be obtained using Equation 7. These values can be plugged in as initial emission parameters for the EM algorithm. Classification is finally performed by computing the most likely state sequence.

Table 2 summarizes classification results on test set documents sequences, after a training phase applied both to Naive Bayes and our hybrid model. We report accuracy of prediction on single classes and average accuracy over the total of text documents. Comparison is made with respect to the best isolated-page classifier. The hybrid HMM classifier (performing sequential classification) achieves 85.28% accuracy and consistently outperforms the plain Naive Bayes classifier working on isolated pages. The relative error reduction is about 46%, i.e. roughly half of the errors are recovered thanks to contextual information. In particular, it is interesting to note the large error reduction for the category "Advertisements." Pages in this category typically contain several images and few words of text. The isolated page classifier is subject to prediction errors in this case since parameter estimation for rarely occurring words can be poor. On the other hand, the constraints imposed by the grammar allow to recover many prediction errors since advertisements normally occur near the end of each issue.

In Figure 7 we report classification performances of the hybrid model on single issues of the journal. The graph is to be interpreted as the classifier temporal trend from 1885 to 1894. Negative accuracy peaks correspond to test issues with more than 70 pages, a significant deviation from the average number of pages per issue (about 32). Values range from a minimum of 50% to a maximum of 97.09% with 10.41 as standard deviation. To visualize a smoother trend, we calculated a running average over a temporal window of 10 months, showing a clear superior trend over standard naive Bayes.

4.3.2 *Partially labeled documents*

We have performed six different experiments, for different percentages of labeled documents. In this case the structure learning algorithm cannot be applied and we used ergodic HMMs with ten states (one per class). After training, transition with probabilities $< 10^{-3}$ were pruned. In one of the six experiments we used all the available page labels with an ergodic HMM. This experiment is useful to provide a basis for evaluating the benefits of the structure learning algorithm presented in Section 3.2.

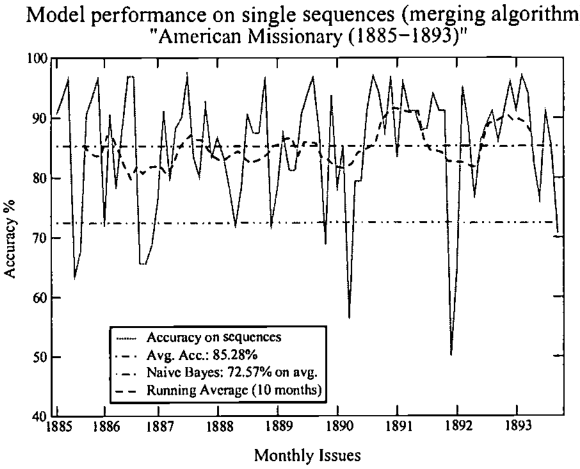


Figure 7: Performance of the hybrid model on single sequences (merging algorithm).

	% of labeled documents					
Category	0	30	50	70	90	100
Contents	0	100	100	100	100	100
Editorial	20.76	21.12	59.67	58.6	67.62	71.41
South	1.51	83.58	69.73	84.94	84.34	84.19
Indians	10.07	0	55.03	51.68	50.34	58.39
Chinese	0	27.45	83.66	76.47	75.82	75.16
Bur.W.W.	0	43.22	63.74	63	64.84	65.93
Child. P.	4.35	78.26	73.91	58.7	78.27	76.09
Commun.	0	91.4	91.4	93.55	93.55	93.55
Receipts	0	89.27	98.68	97.36	98.31	98.31
Advert.	81.4	69.77	93.02	90.7	90.7	90.7
Total Accuracy	8.23	55.66	73.54	75.66	78.7	80.24

Table 3: Results achieved by the model trained by Expectation-Maximization, varying percentage of labeled documents.

Table 3 shows detailed results of the experiments. Classification accuracy is shown for single classes and for the the entire test set. As we can see, EM being completed uninformed (0% evidence) is worse than the random guess (8.23% accuracy). With 50% of labeled documents, the model outperforms Naive Bayes (73.54% against 72.57%). This is a positive result, because the Naive Bayes training phase (in the standard formulation) need the knowledge of all document labels, while in this setting we simulate the knowledge of only a half of them. With greater percentages of labeled documents, performances begin to saturate reaching a maximum of 80.24% when all the labels are known. This result is worse compared to the 85.28% obtained with the first strategy (see Section *4.3.1*). The main difference is that in this case we started training from an ergodic model and we used one state per class. This confirms that in the case of completely labeled documents it is advantageous to use more states per class and to use the data-driven algorithm for structure selection.

5. CONCLUSIONS

We have presented a text categorization system for multi-page documents which is capable of effectively taking into account contextual information to improve accuracy with respect to traditional isolated page classifiers. Our method can smoothly deal with unlabeled pages within a document, although we have found that learning the HMM structure further improves performance compared to starting from an ergodic structure. The system uses OCR extracted words as features. Clearly, richer page descriptions could be integrated in order to further improve performance. For example, optical recognizer output information about the font, size, and position of text, that may be important to help discriminating between classes. Moreover, OCR text is noisy and another direction for improvement is to include more sophisticated feature selection methods, like morphological analysis or the use of n-grams [4, 14].

Another aspect is the granularity of document structure being exploited. Working at the level of pages is straightforward since page boundaries are readily available. However, actual category boundaries may not coincide with page boundaries and some pages contains portions of text related to different categories. Although this is not very critical for single-column journals such as the *American Missionary*, the case of documents typeset in two or three columns certainly deserves attention. A further direction of investigation is therefore related to the development of algorithms capable of performing automatic segmentation of a continuous stream of text, without necessarily relying on page boundaries.

The categorization method presented in this paper is targeted to textual information. However, the same hybrid HMM methodology could be applied for classification of pages based on layout information, provided an adequate emission model is available. A suitable generative model for document layout is presented in [8].

Finally, categorization algorithms that includes contextual information may be very useful for other types of documents natively available in electronic form. For example, the categorization of web pages may take advantage of the contents in neighbor pages (as defined by the hyperlink structure of the web).

6. ACKNOWLEDGMENTS

We thank Oya Rieger and the Cornell University Library for providing the data collected within the Making of America project. This research was partially supported by EC grant # IST-1999-20021 under METAe project.

7. REFERENCES

[1] The metadata engine project. `http://meta-e.uibk.ac.at`, 2001.

[2] Y. Bengio and P. Frasconi. An input output HMM architecture. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 427–434. MIT Press, 1995.

[3] D. A. Bicknese. Measuring the accuracy of the OCR in the Making of America. Report available at `moa.umdl.umich.edu/moaocr.html`, 1998.

[4] W. Cavnar and J. Trenkle. N-Gram based text categorization. In *Prof. of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, NV, 1994.

[5] E. Charniak. *Statistical Language Learning.* MIT Press, 1993.

[6] W. W. Cohen. Text categorization and relational learning. In *Proceedings of the Twelfth International Conference on Machine Learning*, Lake Tahoe, California, 1995.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, 39:1–38, 1977.

[8] M. Diligenti, P. Frasconi, and M. Gori. Image document categorization using hidden tree-Markov models and structured representations. In S. Singh, N. Murshed, and W. Kropatsch, editors, *Second Int. Conf. on Advances in Pattern Recognition*, volume 2013 of *LNCS*. Springer, 2001.

[9] D. Freitag and A. McCallum. Information extraction with hmm structures learned by stochastic optimization. In *Proc. 12th AAAI Conference*, Austin, TX, 2000.

[10] D. Heckerman. Bayesian networks dor data mining. *Data Mining and Knowledge Discovery*, 1(1):79–120, 1997.

[11] F. Jensen. *An Introduction to Bayesian Networks.* Springer Verlag, New York, 1996.

[12] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning.* Springer, 1998.

[13] T. Joachims. Transductive inference for text classification using support vector machines. In *Int. Conf. on Machine Learning*, 1999.

[14] M. Junker and R. Hoch. Evaluating OCR and non-OCR text representations for learning document classifiers. In *Prof. ICDAR 97*, 1997.

[15] T. Kalt. A new probabilistic model of text classification and retrieval. CIIR TR98-18, University of Massachusetts, 1996. url: `ciir.cs.umass.edu/publications/`.

[16] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proc. Fourteenth Int. Conf. on Machine Learning*, 1997.

[17] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *SIGIR-94*, 1994.

[18] D. Lewis and M. Ringuette. Comparison of two learning algorithms for text categorization. In *Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994.

[19] H. Lucke. Bayesian belief networks as a tool for stochastic parsing. *Speech Communication*, 16:89–118, 1995.

[20] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, 3:127–163, 2000.

[21] T. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

[22] H. Ng, W. Goh, and K. Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *Proc. of the 20th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 67–73, 1997.

[23] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[24] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[25] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[26] E. Shaw and S. Blumson. Online searching and page presentation at the University of Michigan. *D-Lib Magazine*, July/August 1997. url: `www.dlib.org/dlib/july97/america/07shaw.html`.

[27] P. Smyth, D. Heckerman, and M. I. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9(2):227–269, 1997.

[28] A. Stolcke and S. Omohundro. Hidden Markov Model induction by bayesian model merging. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 11–18. Morgan Kaufmann, San Mateo, CA, 1993.

[29] Y. Yang and C. Chute. An example-based mapping method for text classification and retrieval. *ACM Transactions on Information Systems*, 12(3):252–277, 1994.

[30] Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997.

Automated Name Authority Control

James W. Warner
Digital Knowledge Center, MSE Library
Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218
jwarner@jhu.edu

Elizabeth W. Brown
Cataloging Dept., MSE Library
Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218
ebrown@jhu.edu

ABSTRACT

This paper describes a system for the automated assignment of authorized names. A collaboration between a computer scientist and a librarian, the system provides for enhanced end-user searching of digital libraries without increasing drastically the cost and effort of creating a digital library. It is a part of the workflow management system of the Levy Sheet Music Project.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval—*Digital Libraries*

Keywords

Name Authority Control, automation, indexing, metadata, workflow management

1. INTRODUCTION

The Levy Sheet Music Collection[1] includes a rich online index created during digitization. This index, designed specifically to meet the needs of Levy users and based in part on Music Library Association guidelines, exists outside the context of the traditional library cataloging "MARC record" environment. Although it includes transcribed information about the composers of music and artists responsible for covers, the index does not provide for traditional library "authority control"—disambiguating and establishing the "authoritative" version of a person's name and providing the ability to search on "variants" or "cross-references."

Thus, users searching the collection are not able to retrieve easily the varying forms of names of persons; for instance, a search for "Stephen Foster" will not retrieve all the instances of Stephen Foster as "Stephen Collins Foster." More importantly, a user does not necessarily know he or she has not found all that the Levy Collection metadata has to offer on a particular artist or composer. An extreme case found in Levy is the search for "Alice Hawthorne," which does not retrieve the occurrences of "Septimus Winner." The Library of Congress name authority files, however, document these names, along with "Sep Winner," as names for the same person.

2. BACKGROUND

With an expanding audience and the upcoming addition of sound and full-text lyric searching, enhancing the searching capabilities with some kind of authority control is desirable. Bringing "retrospective authority control" to a collection of over 29,000 titles can be quite a daunting task. In the Levy II framework of developing a workflow management system, however, such a project poses interesting possibilities for devising authority control tools for digitized sheet music or text collections in general.

Levy II does not involve hand-transcribing the names from the statements of responsibility or manually searching a local or national authority file but establishing a way to automate these tasks. The project seeks to apply this "retrospective" process to creating a truly useful tool to provide for authority control in the "metadata" or "digital library environment" outside the context of traditional cataloging tools. Thus, indexers will be able to use the tool as part of a digitization project workflow. With OCLC's CORC emerging and developing during the Levy II process, we can assess how the Levy authority control tool compares to tools created in other metadata and cataloging initiatives.

The task involves first extracting the names of the composers, lyricists and arrangers, as well as engravers, lithographers and artists, from the transcribed statements in the metadata, then inventing an efficient method for creating authoritative forms of names. Adhering to the Library of Congress name authority file, a standard authority file, eliminates "reinventing the wheel" to a certain extent and provides for interoperability with other sheet music catalogs or indexes.

3. METHODOLOGY

In order to test the feasibility of the project, we have restricted our work so far to composers, lyricists, and arrangers. Since the names and their roles are not explicitly entered into the Levy metadata, we have developed an automated system to extract this information. The task is not difficult, since the names are entered directly from the title page of the piece. For example, a piece might say "Composed by John Smith. Arranged for the piano by Bill Jones." Using a dictionary as well as a list of first and last names,

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

we have extracted the names with simple pattern matching techniques. We achieved 97.2% precision and 98.7% recall. Recall was favored, since the disambiguation system itself should help recognize invalid names.

In order to retrieve authority records efficiently, we have loaded the personal names subset of the Library of Congress Name Authority File into a MySQL database. The database allows us to query for all records that match a given name. It also allows us to retrieve context from notes fields in the authority records and from author-title authority records. We will link names from the Levy Collection to their records in the authority file through the unique Library of Congress Control Numbers.

We started the process of name disambiguation by breaking up the names extracted from the Levy collection into four randomly selected groups. We have a seed group, two training groups, and a held-out test group. Each group is then clustered, putting similar names together. This clustering can be done strictly, loosely, or not at all depending on the demands of the collection. Clustering names results in having to disambiguate fewer names, but if names are often repeated within the collection, results could be adversely affected. Clustering and thus gathering varying forms of names for the same persons, however, prepares for bringing the names not found in the LC authority file under authority control. We disambiguate the seed group manually to train the automated disambiguator. From this seed, we can gather statistics about how names in the Levy collection match their records in the LC Authority File.

Since no automated system can be 100% accurate, the system must be able to express its level of confidence. We intend to use a Bayesian approach to achieve this goal. For example, if N represents a name in the Levy collection, R is an authority record in the authority file, and $FN(X)$ represents the first name associated with X, we say $P[N = R|FN(N) = FN(R)] = \frac{P[FN(N)=FN(R)|N=R]*P[N=R]}{P[FN(N)=FN(R)]}$. $P[FN(N) = FN(R)|N = R]$ can be calculated from the seed data. $P[FN(N) = FN(R)]$ can be calculated using first name frequencies and $P[N = R]$ is the prior probability. Each piece of evidence can be handled in this Bayesian manner with the reasonable assumption that each piece of evidence is conditionally independent.

Many types of evidence can be used to disambiguate the names. One such piece of evidence, as shown above, is how common a name is. For example, finding the name "John Smith" in the LC database is much less significant than finding the name "Irving Berlin." Additionally, if the notes fields of a record discuss music in some way, the likelihood of being a match increases. Thus, we use vector similarities with the seed notes to determine if a given record seems to be that of a musician. Another very important piece of evidence is publication date. A person could not have authored a document that was published before his or her birth, and, within the Levy Collection, most pieces were published during the author's lifetime. The probabilistic method allows a flexible weighting of all of these pieces of evidence. For example, if a name in the Levy collection does not match well with any of the name variants in the Authority File, the system could still make the connection if the contextual evidence is very strong.

Using the seed data, the system disambiguates the first training set. Unsupervised learning techniques can be employed to improve results.[2] A human can also intervene when confidence is low to help the program train. This whole process can then be repeated with the second training set. Finally, the accuracy can be measured on the held-out test set. This iterative process helps the scalability of the system by requiring only a small number of records to be disambiguated manually initially.

4. FUTURE WORK

Based on promising initial results, we are applying the methodology to the Levy Collection. We plan to quantify the results fully and compare different methods and parameters. Such parameters include the size of the manually assigned set and the prior probabilities for each piece of evidence. We will then test the system further by bringing the engravers, lithographers, and artists under authority control.

We believe the system could be extended beyond the domain of sheet music. Any large collection of digital documents could potentially benefit from the system, since doing the work manually is costly and time consuming. The system is also advantageous because the iterative nature requires that only a very small percentage of the documents be processed manually.

5. CONCLUSIONS

We believe that name authority control is beneficial to users of digital libraries. However, understanding that the manual process is expensive and time consuming, we have developed a system to automate authority control. We believe that the system can be applied generally to a variety of digital libraries.

6. ACKNOWLEDGMENTS

We would like to thank Sayeed Choudhury, Brian Harrington, and Tim DiLauro of the Digital Knowledge Center. The Levy Project is funded through NSF's DLI-2 initiative (Award #9817430), an IMLS National Leadership Grant, and support from the Levy Family.

7. REFERENCES

[1] G. S. Choudhury, C. Requardt, I. Fujinaga, T. DiLauro, E. W. Brown, J. W. Warner, and B. Harrington. Digital Workflow Management: The Lester S. Levy Digitized Collection of Sheet Music. *First Monday*, 5(6), June 2000.

[2] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, 1995.

Automatic Event Generation from Multi-lingual News Stories

Kin Hui
The Chinese University of Hong Kong
Shatin, Hong Kong, PRC
khui@se.cuhk.edu.hk

Wai Lam
The Chinese University of Hong Kong
Shatin, Hong Kong, PRC
wlam@se.cuhk.edu.hk

Helen M. Meng
The Chinese University of Hong Kong
Shatin, Hong Kong, PRC
hmmeng@se.cuhk.edu.hk

ABSTRACT

We propose a novel approach for automatic generation of topically-related events from multi-lingual news sources. Named entity terms are extracted automatically from the news content. Together with the content terms, they constitute the basis of representing the story. We employ transformation-based linguistic tagging approach for named entity extraction. Two methods of gross translation on Chinese story representation into English have been implemented. The first approach uses only a bilingual dictionary. The second method makes use of a parallel corpus as an additional resource. Unsupervised learning is employed to discover the events.

Keywords

Event Discovery, Event Detection, Multi-lingual Text Processing

1. INTRODUCTION

Interests in the event generation have grown rapidly in recent years. While most existing studies have concentrated on handling relevant specific queries, there is a need to explore techniques for classifying relevant issues from a continuous stream of data. Moreover, the rapid growth of the Internet and wide availability of electronic media allows the use of electronic means to access the newswire stories from diverse sources. Thus, the capability of identifying whether multi-lingual news are discussing the same event is one of the challenges of the event discovery task. In this paper, we particularly focus on dealing with English and Chinese news stories.

An event is defined as a particular activity or incident happening in certain place at a certain time as well as any follow-up progress. In the event generation problem, news stories are arriving from different sources in chronological order. Some stories are in Chinese and some are in English. Event generation aims at identifying whether the incoming stories belong to a new event or an existing event detected in previous stories. The detection can contribute to the construction of structured guidelines for story navigation of the whole news collection.

Traditionally, stories are represented solely on raw terms appeared in their content. In our approach, we augment the representation by named entity components, namely person names, geographical location names and organization names. This story representation is able to capture terms conveying useful context for event generation purpose. We look into two named entity extraction approaches and investigate their impacts on event generation. The first approach utilizes the named entity extraction module provided in a commercial text mining product. The second method makes use of the transformation-based linguistic tagger and a collection of transformation rules to extract named entity terms. The transformation rules are learned from a training corpus.

Another feature of our approach lies in the online gross translation. We have developed two methods to translate Chinese terms into English. The first one is the basic translation, which involves looking up each term in a bilingual dictionary and replaces it with possible term translation. Another approach utilizes the easily available resources like a parallel corpus to perform translation. Discovery of new events is tackled by unsupervised learning. Our unsupervised learning is based on a modified agglomerative clustering algorithm.

2. STORY REPRESENTATION

As raw stories contain a sequence of words, we need to identify sentence boundaries for subsequent processing. Lexical clues, such as punctuation marks, are used to locate the boundaries. To address the absence of word boundaries for Chinese stories, we employ the dynamic programming technique based on the tool provided by Linguistic Data Consortium (LDC) [1] to locate it.

We augment the story representation by a two-dimensional semantic expression, comprising of named entity feature representation and content term representation. Named entity feature representation is constituted by people name component, geographical location name component and organization name component. Each component is represented by a set of terms with corresponding weights expressed as a vector.

3. NAMED ENTITY EXTRACTION

We look into two approaches for extracting named entity terms. The first method is taken from the Feature Extraction Tool in the Intelligent Miner for Text Analysis package from IBM. Presently, this tool can only be applied for processing English newswire stories. The second method is based on transformation-based linguistic tagging (TEL) approach derived from Brill [2]. This approach can deal with English and Chinese as well as newswire and broadcast stories.

For Chinese stories, unknown words may cause incorrect segmentations of named entities which degrade the results of tagging. Therefore, we apply unknown word identification [3] to cope with this problem. Together with the word candidates

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

produced from the unknown word identifier and the named entity word tokens tagged by tagger [4], we associate a weight to each term to represent the story for the event discovery task.

4. GROSS TRANSLATION

In order to determine whether multi-lingual news stories are topically-related, we conduct translation on the story representation of Chinese stories to English so that we can perform unsupervised learning on a uniform data representation. Two approaches of translation have been developed, namely the basic translation and the enhanced translation.

4.1 Basic Translation

We make use of a bilingual dictionary and a pin-yin file provided by Linguistic Data Consortium (LDC) in finding the English translation terms. First, we look up the Chinese term in the bilingual dictionary and retrieve the corresponding set of English terms. If there is no dictionary entry for a term in the people name component or the geographical location name component, we obtain the translation by means of the pin-yin of the Chinese term. The English translation terms will replace the original Chinese term to represent the story with the assigned weights based on the original Chinese term weight.

4.2 Enhanced Translation

In addition to the bilingual dictionary, we utilize a parallel corpus collection on Chinese and English news items as an additional resource to improve the accuracy of translation. An automated procedure is developed to transform a parallel corpus to passage-aligned setting. For each Chinese story term, we use a similarity-based retrieval engine to retrieve the top-ranked Chinese passages in the parallel corpus that are relevant to the Chinese term by posing the Chinese term as a query. Then we retrieve the corresponding English passages of those Chinese passages. Finally, we adjust the weight of the translated terms by assigning a higher weight for terms with better translation quality based on the statistics obtained from the retrieval English passages.

5. EXPERIMENTS AND RESULTS

In order to examine the performance of our event generation system, we evaluate our method by the corpus used in the latest Topic Detection and Tracking 2000 (TDT2000) project organized by DARPA [5] . The corpus consists of text and transcribed speech news data, distributed by the National Institute of Standards and Technology, in Chinese and English spanning from January 1, 1998 to June 30, 1998. We followed the topic detection evaluation methodology designed in the TDT2000 project. The detection performance is measured by the metric *$(C_{DET})_{Norm}$*. Smaller values correspond to better performance.

We have conducted two sets of experiments to investigate the performance of our event generation approach. In the first set of experiments, we have processed the Chinese stories by the original Chinese terms, both the basic and enhanced translation are conducted on the stories under various settings to compare the effects of translation. Table 5.1 reveals the detection performance, in terms of *$(C_{DET})_{Norm}$*, under various similarity thresholds, *q*, from 0.05 to 0.15.

Table 5.1 Performance on Translated Chinese Stories Detection Under Different Similarity Threshold

	Similarity Threshold				
	0.05	0.08	0.1	0.13	0.15
Original Chinese Terms	0.4950	0.4820	0.4725	0.4876	0.5116
Enhanced Translation	0.6362	0.5961	0.5036	0.5214	0.4705
Basic Translation	0.6307	0.5224	0.5198	0.5184	0.4952

The result shows that, the detection performance of the enhanced translation method is more encouraging than that of the basic translation method. We also observe that the best result achieves a detection cost of 0.4705. If the number of content terms is increased to 25, it can even achieve 0.4470.

Table 5.2 summarizes the detection performance under different similarity threshold, *q*, by varying from 0.05 to 0.13 for event generation of all English stories and the multi-lingual stories.

Table 5.2 Performance on English Stories and Multi-lingual Stories Detection Under Different Similarity Threshold

	Similarity Threshold								
	0.05	0.06	0.07	0.08	0.09	0.1	0.11	0.12	0.13
ENG IMT	0.4308	0.4279	0.4379	0.4365	0.4346	0.4276	0.4349	0.4181	0.4376
ENG TAG	0.4132	0.3801	0.3864	0.3986	0.3854	0.3809	0.3802	0.3897	0.3881
MUL IMT	0.4545	0.4387	0.4223	0.5057	0.4127	0.4288	0.4291	0.4449	0.4252
MUL TAG	0.4082	0.3861	0.3882	0.3777	0.3913	0.3924	0.3766	0.3777	0.3744

From the result, it depicts that the detection cost of the stories with the named entity term extracted by the transformation-based linguistic approach (TAG) is better than those extracted by the commercial product (IMT) for both multi-lingual (MUL) and English (ENG) stories. The best results obtained from the TAG approach from both data stream are 0.3744 and 0.3801 respectively compared with 0.4127 and 0.4181, the best costs produced by IMT method.

6. FUTURE WORKS

We will conduct more experiments to optimize the parameter settings. Moreover, we will focus on investigations on unsupervised learning module and explore other techniques on clustering to further improve the performance.

7. REFERENCES

[1] http://morph.ldc.upenn.edu/TDT Linguistic Data Consortium

[2] E. Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. In 1995 Association for Computational Linguistics, Volume 21, Number 4, pages 543-565, 1995.

[3] C.W. Ip. Transformational Tagging for Topic Tracking in Natural Language. In Dept. of Systems Engg. & Engg Mngt of CUHK Master Thesis, June 2000.

[4] H. Meng and C.W. Ip. An Analytical Study of Transformational Tagging for Chinese Text. In *Proceedings of Research on Computational Linguistics Conference (ROCLING XII)*, Taipei, Taiwan, ROC,pages 101-122, 1999.

[5] The Year 2000 Topic Detection and Tracking (TDT2000) Task Definition and Evaluation Plan.

Linked Active Content: A Service for Digital Libraries for Education

David Yaron
Department of Chemistry
Carnegie Mellon University
Pittsburgh, PA 15213
412-268-1351
yaron@chem.cmu.edu

D. Jeff Milton
Department of Chemistry
Carnegie Mellon University
Pittsburgh, PA 15213
412-268-1065
milton@chem.cmu.edu

Rebecca Freeland
Department of Chemistry
Carnegie Mellon University
Pittsburgh, PA 15213
412-268-7981
rf51@andrew.cmu.edu

ABSTRACT

A service is described to help enable digital libraries for education, such as the NSDL, to serve as collaboration spaces for the creation, modification and use of active learning experiences. The goal is to redefine the line between those activities that fall within the domain of computer programming and those that fall within the domain of content authoring. The current location of this line, as defined by web technologies, is such that far too much of the design and development process is in the domain of software creation. This paper explores the definition and use of "linked active content", which builds on the hypertext paradigm by extending it to support active content. This concept has community development advantages, since it provides an authoring paradigm that supports contributions from a more diverse audience, including especially those who have substantial classroom and pedagogical expertise but lack programming expertise. It also promotes the extraction of content from software so that collections may be better organized and more easily repurposed to meet the needs of a diverse audience of educators and students.

Categories and Subject Descriptors

K.3.1. [**Computers in Education**]: Computer Uses in Education, *computer assisted instruction, computer managed instruction.*

General Terms

Human Factors, Experimentation

Keywords

Education, Active learning, Web authoring

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

1. INTRODUCTION

A large body of educational research has shown the benefits of active learning, whereby students do not sit passively while they are told information, but rather participate actively in the learning process [13]. Digital libraries for education, such as the NSDL, can help catalyze the shift towards more active, exploratory, inquiry-based learning activities. The digital library's potential lies not only in the ability to bring active learning experiences to an unprecedented and large audience of students, but also in the potential to serve as a collaboration space to catalyze the creation, modification and assessment of student activities. First, a grass-roots development approach, which remains in intimate touch with the needs of teachers and students, may reduce concerns that can prevent adoption of useful teaching innovations. Second, and perhaps more difficult to achieve, if the library is to be populated with a large amount of high-quality material, it will be important to engage a large community of developers and early-adopters, with a wide range of interests, expertise and approaches.

The goal of our research is to provide a technical infrastructure that will help digital libraries, such as the NSDL, overcome some of the main challenges associated with collaborative creation of engaging learning activities. From a community building perspective, creation of these activities requires both technical and pedagogical expertise, a combination of expertise that few members of the NSDL community can be expected to possess. From a collections perspective, the level of interactivity required for engaging activities typically leads to monolithic chunks of software that are difficult to subdivide into components that promote adaptation and reuse. Our premise is that these challenges are intimately coupled, since they both relate to where one draws the line between *software creation* and *content authoring*. The current location of this line, as defined by available web technologies, is such that far too much of the design and development process is in the domain of software creation. By pushing this line to allow for more powerful and flexible content authoring, we can better utilize the expertise of those members of the community with curriculum development and classroom experience. Redefining the line between programmer and author also promotes the extraction of content from software in a manner that leads to better organized collections that may be repurposed to meet the needs of a diverse audience of educators and students.

2. RELATED WORK

Over the last 40 years, there have been thousands of individual efforts to create science and engineering teaching and learning materials in digital formats, ranging from case study depositories to instructional software. Some have been quite effective, such as the Physics Academic Software (PAS) library [4]. It is now widely recognized, however, that developing configurable learning objects that will be reused is essential to vitalizing online instruction. The efforts of groups including the Instructional Management Systems group started by EDUCAUSE [7], the IEEE Learning Technology Standards Committee [11], the ARIADNE project [3], the Advanced Distributed Learning Initiative [5], and others to create standards for learning objects is partly motivated by the reality that reuse is rare. Most of the standards currently being developed revolve around textual content (metadata descriptions for content, standards for questions and answers, and curriculum structure standards), and not around software development and reuse.

It has long been hoped that instructional software developers would contribute to libraries of instructional software that would give less technically oriented educators access to the benefits of software simulations and interactive exercises. There are a number of significant new efforts at creating architectures for developing and sharing educational software components [2, 6, 8, 9, 10, 12], most of them based on the Java language. The ESCOT [9] project is a testbed that seeks to encourage development and re-use of learning objects, with a current emphasis on middle-school mathematics. The NEEDS [14] project to create a digital library for engineering education includes educational software on their site. The National Science Foundation Computer Courseware Repository [15] promises inclusion of instructional software but posts little to date

Current tools to invite participation of instructors with little programming expertise include general authoring environments, such as Macromedia Director and Authorware or Macromedia's Web-based CourseBuilder. While these are powerful and useful tools, they do not have a smooth integration path for coupling to repositories of Java objects, and so are not sufficient to meet the goals addressed here. They also present users with a steep learning curve. Another approach to the construction of tools for this audience starts with general component assembly models such as the JavaBeans model and simplifies and/or provides support to make this approach accessible to a broader audience. Development of such tools is an active research area, for instance at the ESCOT project. As discussed further below, our approach is related to this one but, rather than starting with component assembly and reducing its complexity, we start with Web content creation and extend its abilities. This leads to tools that complement those of other projects, and invite a much broader authoring audience.

3. LINKED ACTIVE CONTENT

One approach to redefining the line between educational software creation and content authoring is to view the creation of curricular materials as primarily *programming* and simplify this activity so that it is accessible to more instructors. Our approach starts from a very different perspective, that of curriculum creation as *Web authoring*. We believe that this approach has advantages in learnability, organizing and reusing content, and supporting domain-specific software components. Most instructors are already familiar with the hypertext link and image map functionalities of Web authoring. Grounded in this model, we extend it to include links to and between simulations and other active learning objects such as tutorials or animations. This builds on the familiar paradigm of hypertext linking: it is intuitively simple to go from creating a hypertext link to sending a message to a learning object. While simple, this linking ability significantly increases the ability of a non-programmer to create interesting active learning experiences.

In addition to the community building advantages discussed above, linked-active content also has advantages for the structure of a digital library for education. Consider the current manner in which active content is supported on the web. Currently, the limitations of HTML are overcome with plug-in technologies, such as Flash or Java, that essentially provide embedded browsers for content stored in a format that has little or no relation to the overall HTML in which the plug-in is embedded. Without good communication between the plug-in content and the overall HTML, even text and image portions of the software, which could be handled with HTML, must instead be embedded in the plug-in. This embedding leads to large chunks of content that are difficult to adapt and reuse. Although this active content often consists of sub-objects with relationships, similar to a hypertext web site, the limits of the plug-in technology prevent this substructure from being made explicit in a manner that can lead to better organized digital library collections.

To catalyze the creation of good active learning experiences, the infrastructure should not only support the creation of domain-specific software components, but actively promote coupling of new components to existing ones. For many of the activities that drive the web, such as product advertising and online sales, the functionality provided by HTML and plug-ins is sufficient. However, the creation of active learning experiences involves a mix of cross-domain components, such as those that handle text and images, along with domain-specific components, such as that in scientific simulations.

Consider a Virtual Lab for chemical education that provides a flexible simulation in which students may perform a large variety of experimental procedures in a manner that mimics that of a real laboratory [18]. While this flexibility supports a wide variety of approaches to chemical education, many approaches call for providing students with guidance so that they interact with the simulation in a meaningful way [19]. This guidance is typically *text* and *image* based: what to do at a certain stage, explanations of concepts, questions to be answered, etc. While much of the functionality needed to provide this guidance is domain independent, educational software presents a rather unique need for both general and domain-specific components. A flexible means to present text and images, such as that in HTML, can be made much more powerful by allowing the text and images to link to simulations. In the Virtual lab example, it should be possible to provide text links and buttons that cause chemical solutions to appear in the virtual lab, or that check the current state of the lab to see if the student has achieved a certain goal. This design promotes reuse of components, since each domain and each software project do not need to reinvent the general tools, but need only provide domain-specific software components.

4. PROGRAMMER AND AUTHOR ROLES

Producing a shift in the line between software creation and content authoring requires not just a change in technical infrastructure, but also a change in how programmers and educators view their roles. Although a programmer must create learning objects, a major goal of our work is to promote a change in the mind-set of the programmer, away from creating finished pieces of educational software and towards creation of learning objects that serve as viewers and manipulators of content.

The issues we are addressing may then be viewed as a specific instance of the more general objective of creating useful components for educational software. Our project is a collaboration with Andries van Dam and Anne Spalter of Brown University, who are considering the creation of components for use by programmers. Our emphasis here is on the creation of components for use by instructors. These instructor components are to be created by programmers, ideally using the software components developed by our Brown collaborators.

5. LIBRARY SERVICES

The goal of our research is to explore the requirements and benefits of allowing links between active content. Active content is that content contained in scientific simulations, tutorials, and other materials that, due to their interactive nature, go beyond the abilities of HTML and so are constructed using JAVA or a web plug-in technology. We will refer to the software components that present this content as learning objects. By "linked active content", we mean information passing between learning objects through conduits that are created during the authoring process rather than being hard-coded into the objects themselves. Such links can pass a message to a learning object, or query an object for data. (More complex means of transferring information between objects are discussed in Section 8.4.)

A principal motivation of this approach is to give an intuitive yet powerful ability to curriculum authors and instructors. It is not a big leap to go from creating a hypertext link to sending a message to a learning object, especially if the link creation is done through dialog boxes as described below. For example, in authoring curriculum around a combustion engine simulation, the author could insert the text "click here to start the engine", with the word "here" being a link that sends the "start" message to the engine. The extended link is similar to a method call in an object oriented programming language, and has the logical structure *target:action:data*, stored via XML.

The output of a completed authoring process is an XML file describing the initial arrangement and state of the learning objects, and the links between these objects that will drive the simulation and react to the student's interactions with it. A viewer, implemented in Java, is used to create the learning environment specified by this file. Note that the XML file is not a script file, but rather a set of linked objects with an initial configuration, analogous to a web site containing a set of linked documents and an initial entry page. This outcome reflects the shift from a simplified-programming to extended-web-authoring paradigm. Unlike current component assembly tools, where errors are typically reported by a compiler in a language that is often obtuse even to experienced programmers, the only types of errors that can arise here are broken links, and even these can be avoided via automated link-generation facilities.

The viewer is a browser that supports dynamic loading of Java objects, and that supports our extending linking structure between these objects. The responsibilities of the viewer are intentionally minimized to make the environment maximally extensible. The functionality resides in the learning objects (implemented as Java classes), and in the links between these objects, not in the viewer itself. We provide implementations of core objects, discussed below, that bring extended linking to web functionalities such as hypertext and imagemaps. However, these learning objects can be swapped out for other implementations. The viewer handles only the extended linking of these objects. Even the placement of objects on the screen is done using a replaceable learning object, analogous to the layout managers of Java. Our initial implementation of the layout manager mimics HTML frames in order to capitalize on potential instructor familiarity.

The environment consists of:

- **Core learning objects** such as a hypertext viewer and image-map viewer that build on their web counterparts by supporting extended linking. These learning objects consist of cross-domain objects for viewing text, images, setting up navigation structures, assessment and grading tools, etc.
- An **authoring environment**, that provides a simple graphical means for arranging the objects and creating links, and that helps the curriculum author organize the potentially large number of text snippets and image maps that make up a complete tutorial or exercise. The organizational aspects are handled by a file-explorer interface that presents the author with a hierarchical list of the corresponding text and image maps. This environment gives the author access to the library of learning objects that can be used to construct simulations. As objects are added to the library, the range of simulations that can be created grows.
- **A viewer**, implemented in Java that creates the learning environment described in the XML file output by the authoring environment.

We are also developing domain-specific learning objects for chemical education, such as a virtual chemistry lab and other simulations. All of the software components use the Java Beans API to expose the methods that will accept messages via the linking mechanism.

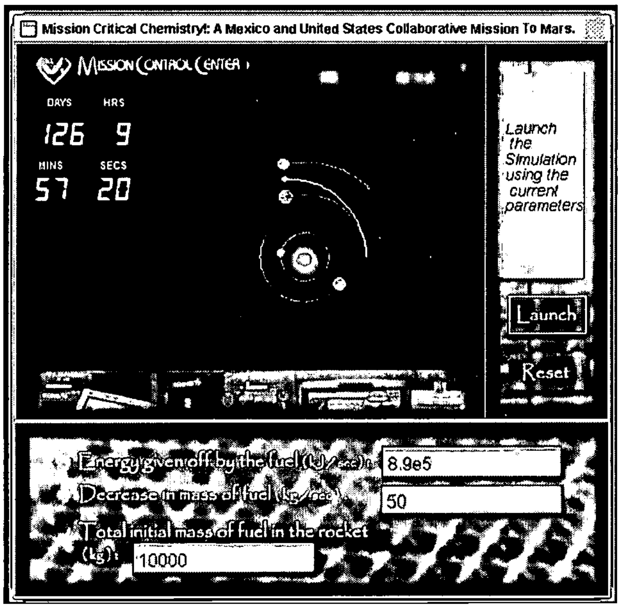
6. EXAMPLES

A major goal of our research is to explore the authoring flexibility that may result from allowing linking between active content. Just as hypertext brings more power to text authoring than may have been expected given its relative simplicity, linking of active content leads to considerably more power and flexibility in the construction of student activities than we anticipated at the start of this project. We will attempt to illustrate this power through the following two examples.

6.1 Mission to Mars

In creating a student activity that utilizes a scientific simulation, the role of the programmer is to develop the simulation while that of the curriculum author is to guide student interaction with the simulation. Under current web technologies, a programmer may provide the simulation as a Java applet, which can then be placed on a web page along with text and image maps that guide the student interaction. The text, especially if placed in a frame below the applet, can take full advantage of the hypertext abilities of HTML to present the material in an appropriate order and even provide nonlinear paths through the material. However, this text remains disconnected from the simulation.

One use of the extended linking mechanism we are developing is to allow the text to pass messages to the simulation.



Mission Critical Chemistry!: A Mexico and United States Collaborative Mission To Mars.

Energy given off by the fuel (kJ/sec): 8.9e5

Decrease in mass of fuel (kg/sec): 50

Total initial mass of fuel in the rocket (kg): 10000

Figure 1. The Mars Simulation exercise with simulation frame (top left), control frame (top right) and parameter frame (bottom).

Figure 1 shows a simulation of a rocket trajectory from Earth to Mars, built using our authoring system. A programmer created the trajectory simulation as a Java component with methods that set the fuel characteristics and other simulation parameters. The author may then use the core learning objects describe in Section 5 to guide student interaction with the simulation. Here, the author has created three frames: the *simulation_frame*, which contains the trajectory simulator object, and a *control_frame* and *parameter_frame*, each of which contains an image-map viewer.

The image-map viewer is a core learning object that builds on the paradigm of the standard web image map. A region (ellipse or polygon) of an image can be made into a "hot spot" that responds when the pointer passes over it, or is clicked/double-clicked. The authoring environment allows the author to create some simple but useful responses to activating a hot spot. For instance, in the *control_frame* of Figure 1, the pointer has activated a hot spot over "Launch" that brings up "bubble help" describing the action of this menu item. (Various filters are provided to allow the look of the bubble help to be customized.) Hot spots may also pop up additional images or a menu of links. In this case, clicking on the Launch hot spot sends a launch message to the trajectory simulator via the extended link: *trajectorySimulator:launch.*

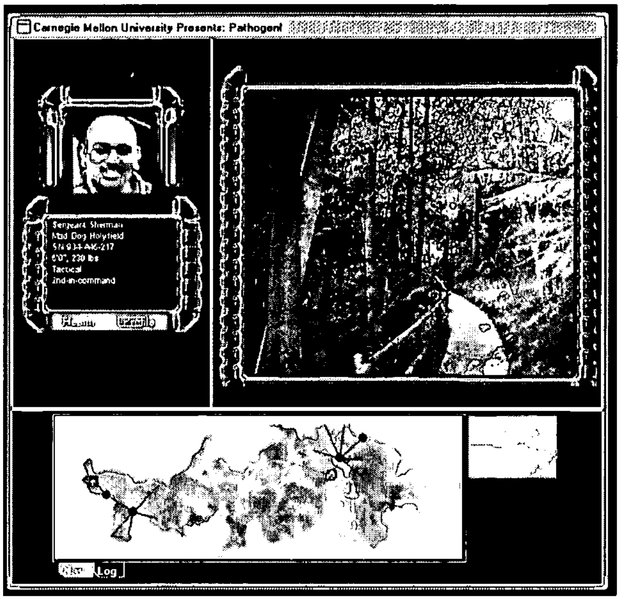
The image map also supports editable hot-spots, which serve as entry boxes for user input. For instance, in the *parameter_frame* of Figure 1, the student may enter and edit text. This text is then passed as data to the method associated with the hot spot. For instance, the upper most text box sends the message *trajectorySimulator:setHeat:8.9e5*. In this manner, the author is allowed to create simple control panels to the simulation. The construction of the control panel is quite different from the approach of JAVA Bean assembly environments. Here, the author first creates an image of the control panel using a tool such as Photoshop. The author then uses our imagemap editor tool to make the relevant portions of the image hotspots that link to the simulation.

In our current implementation, the author must type in the link using appropriate syntax. However, we are currently developing dialog-driven link generation. In this manner, when the author chooses a hot spot for a link, the authoring environment presents a dialog box showing possible targets for the link, such as a list of frames into which a new learning object might appear in response to the student clicking on the hot spot. For example, the author could select *simulation_frame* as the location where the object would appear. This would then prompt for the desired learning object from the library. Once a learning object is selected, the author is prompted with a list of instructions that can be sent to the chosen learning object. For example, a petri dish simulation for growing bacteria could have methods *setGrowthRate, startGrowth,* and *stopGrowth.* Finally, if the chosen method requires data, such in *setGrowthRate*, the curriculum author would be prompted for the relevant data. Thus, the author stipulates how a simulation will behave in response to student actions by doing little more than making hyperlink connections. Note also that more than one link can be attached to a single hotspot or link.

6.2 Pathogen

While conceptually simple, the functionality described above is quite powerful. At the simplest level, providing one frame containing text next to a frame containing a simulation allows authoring of a tutorial that guides a student through the simulation. It is also possible to create the illusion of moving through a virtual world by placing links on imagemaps that load other imagemaps, as is common in computer games such as King's Quest or Myst. Clicking on a door loads an image in which the door is open, and clicking again loads an image of the next room. Such images can be easily obtained with a digital camera. In addition, components may be designed that fit into the environment and provide the author with considerably more power, as illustrated by our Pathogen exercise.

The Pathogen exercise addresses topics typically covered in the first few weeks of an introductory college or high school chemistry course. These topics are put in the context of drug discovery, and in Figure 2, students lead a team of researchers on an island in search of plants with medicinal activity. (They will later bring these plants back to a pharmaceutical laboratory and help determine the active ingredients.)



Carnegie Mellon University Presents: Pathogen!

Sergeant Sherman
Mad Dog Holyfield
SN 934-A46-217
6'0", 230 lbs
Tactical
2nd-in-command

Figure 2: The Pathogen student exercise.

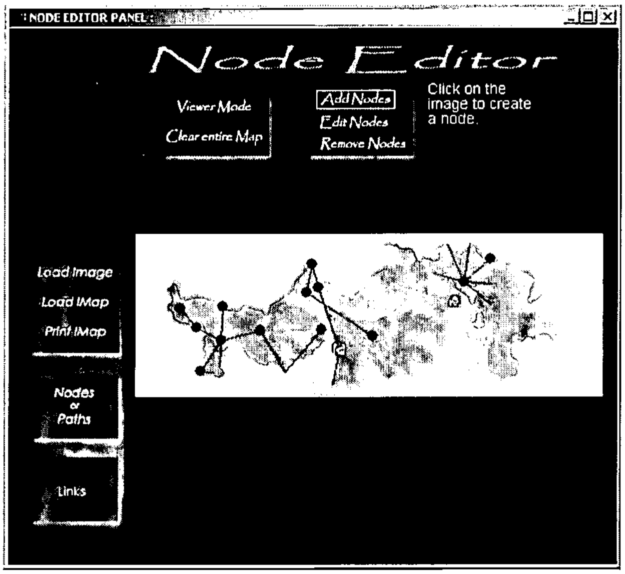
Occasionally, a team member becomes infected and must be given a appropriate drug to be cured. In determining the type and amount of drug, the student must solve a problem involving chemical concepts. If the drug is not appropriate, the team member is airlifted off the island and to a local hospital. The number of initial team members then sets the number of problems the student may get wrong before needing to get a new team and start again.

To support the navigation required by this application, a programmer created a map navigation component, which is loaded into the lower frame of Figure 2. The component shows a map of the island and the current location of team members, and allows the student to move these members around the island. When a team member enters a new location on the map, an image of that location is loaded into the upper-right *workspace* frame of Figure 2. This is an example of a component passing a message to another component. In this case, the map component passes a *loadImage* message to the image-map component in the *workspace* frame. The programmer also created an authoring tool, shown in Figure 3, to allow authors to start with any image and place nodes at various locations to construct a map.

This simple yet powerful navigation component is reused to allow the student to navigate through the pharmaceutical lab in the second part of this exercise. We are currently extending this component to support Quicktime VR images [17]. One can envision other types of navigation components that utilize, for instance, 3-D graphics.

Since the image in the *workspace* frame is displayed through the image-map viewer discussed above, it supports all of the extensions discussed there. For instance, clicking on a hot spot on the image can load a protein viewer showing the molecular structure of a protein relevant to the current exercise.

The upper left panel in Figure 2 contains a profile viewer component that displays information about the currently active team member. This component also has an associated authoring tool that allows an instructor to add his or her own team members.

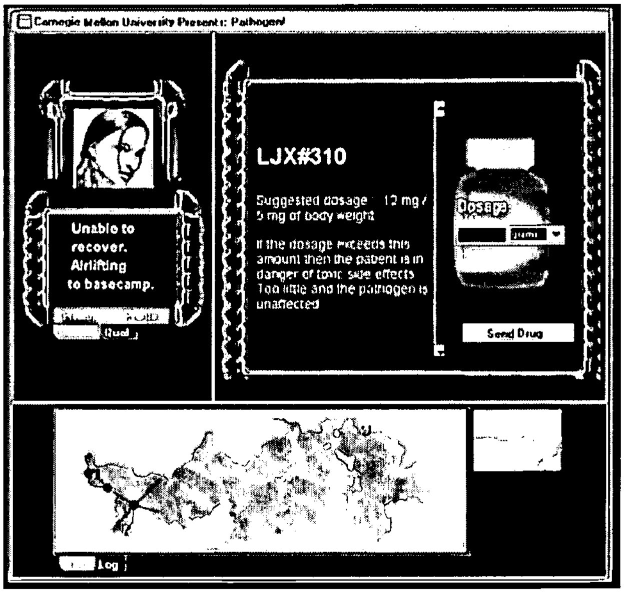


NODE EDITOR PANEL

Node Editor

Click on the image to create a node.

Figure 3. Authoring tool for creating and editing new maps.



Carnegie Mellon University Presents: Pathogen!

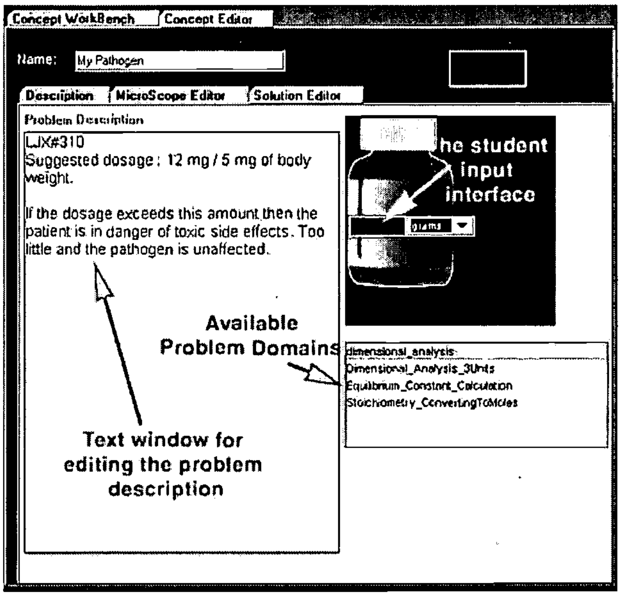
Unable to recover. Airlifting to basecamp.

LJX#310

Suggested dosage : 12 mg / 5 mg of body weight

If the dosage exceeds this amount then the patient is in danger of toxic side effects Too little and the pathogen is unaffected

Figure 4. The Pathogen student exercise displaying the written problem (upper right).


Concept WorkBench
Concept Editor
Description
MicroScope Editor
Solution Editor
grams
the student input interface
Available Problem Domains
Text window for editing the problem description

Name: My Pathogen

Problem Description

LJX#310
Suggested dosage: 12 mg / 5 mg of body weight.

If the dosage exceeds this amount then the patient is in danger of toxic side effects. Too little and the pathogen is unaffected.

dimensional_analysis
Dimensional_Analysis_3Units
Equilibrium_Constant_Calculation
Stoichiometry_ConvertingToMoles

Figure 5. The Pathogen instructor authoring tool showing the creation (or editing) of a specific drug problem.

When a team member becomes infected with a pathogen, a drug-bottle component appears in the *workspace* frame, as shown in Figure 4. The instructions on the drug bottle pose a chemical problem to be solved by the students. Again, an authoring tool, shown in Figure 5, is provided to allow instructors to add their own problems to the application. Note that this problem queries the profile viewer for information such as the body weight of the team member, which may be of relevance to the appropriate cure. This communication between viewers illustrates the use of extended linking to do simple queries, in addition to the message passing discussed above.

Note that in adding a powerful component such as those created for Pathogen, the programmer creates three items: a viewer, a format for the content displayed and/or manipulated by this viewer, and a authoring tool to allow non-programmers to create or modify this content.

Note also that the approach provides multiple entry points for instructors. For instance, an instructor may first simply use the drug authoring tool of Figure 5 to add their own chemistry exercises to Pathogen. Or they may use the profile authoring tool to replace the team members with students in their class. As they become more familiar with the approach, they may attempt more complex modifications and eventually assemble their own exercises.

7. ADVANTAGES

7.1 Instructor Versus Student Interfaces

The above examples illustrate a fundamental shift in the development of interactive software for education, whereby the programmer's primary focus is on designing an "instructor interface" so that the instructor can design a "student interface." Consider the current situation, where a simulation of a combustion engine would typically be constructed as a complete, stand-alone application. The programmer would need to attach a student interface to the simulation, and in so doing, would make curriculum decisions that severely limit potential reuse. If the application were meant to illustrate a college-level concept such as thermodynamic cycles, the resulting user interface would exclude use of the application in high schools to illustrate the ideal gas law. A potential alternative would be to create a simulation with a very flexible interface, but this could easily lead to a complex application that confuses students. With the above assembly environment, the programmer is able to design an "instructor interface," with the goal of providing sufficient flexibility that the curriculum author can design useful "student interfaces." This separation of responsibilities between the programmer and curriculum developer should lead to much more effective collaboration and reuse than is currently possible.

7.2 Browsing through Active Content

In creating a student exercise or tutorial, the curriculum author is essentially creating a means for a student to browse through active content. This browsing ability is supported by the ability to create links that load learning objects into frames. For instance, in creating an exercise about hemoglobin, an author can create two frames, one of which initially contains text describing the role of hemoglobin in the body and the other containing a protein viewer object displaying the 3D molecular structure of hemoglobin. In response to student clicks on text links or images, the instructor can load the virtual lab containing solutions in which hemoglobin has bound up various numbers of oxygen molecules, or an animation showing how hemoglobin sequentially binds up to 4 oxygen atoms.

The curriculum author is thereby allowed to focus on content, independent of the particular software viewer (text viewer, molecular viewer, virtual lab, QuickTime player) needed to display this content. The content is also well extracted from the viewer, in a web with relationships and interconnections that reflect the curriculum content, rather than integrated with the technology needed to display this content.

8. FUTURE DIRECTIONS

8.1 More Flexible Control Panels

The Mars example of Figure 1 showed the use of an editable hot spot to serve as a control to a simulation. This control may be extended to allow sliders, dials etc. to be attached to certain variables of a simulation. This is an area where graphical assembly of components is known to work well. For instance, users with little programming expertise are able to construct simple yet powerful front panels in LabView [16] that control real instruments. Also, attaching text boxes, sliders, etc. to JavaBeans is one area where even simple Bean assembly tools work well. We may adopt this paradigm by allowing users to drag controls on top of an image map to construct a control panel for a simulation. Attaching these controls to a simulation uses the extended linking mechanism; for instance, the author could arrange for a slider to act as a throttle for the engine by first setting the minimum and maximum value, and then linking it to the engine, choosing the "*engine:setspeed:value*" link to the engine object. Such controls are easily built into the hyperlink-like scheme for message passing, since they need to send the message only when altered. This model mimics the construction capabilities of a commercial JavaBeans beanbox assembly environment but delivers it in a non-programming, education-specific service that opens construction to authors. Simultaneously, it provides a service to which educational programmers can contribute their *industry standard* Java classes as learning objects. Of course, those programmers will need to take into consideration design specifications that emerge from the research being done on object design both by other groups and as a result of our own research.

8.2 Branching

Another issue for continued research is the means by which the environment responds to student input. In the Pathogen application discussed above, the student's response to the chemical problem posed on the drug bottle is checked through a special-purpose mechanism provided by the programmer. We are currently working on branching objects that can provide authors with a simple means to add conditional behavior, or if-then statements, to the environment. These form the basis for individualizing responses and feedback to the choices that a student makes in navigating the learning environment. We view complex branching behaviors as the domain of the programmer, and therefore they are meant to be programmed into learning objects. For instance, a simulation such as our Virtual Lab [20] provides the student with varied choices and feedback on the choices they make and actions they take.

One way to give authors branching capabilities is by allowing branching controls to be inserted into image maps, or control panels. The author could, for example, use a multiple choice panel component to allow a student to follow different links depending on his or her answer. Alternately, an integer response control and real-number response control would accept a value from the student, and follow various links depending on the value. Allowing the input to such controls to come from simulations (via extended linking) leads to a flexible simulation environment. For instance, the author can create text in a text-viewer that asks the student to adjust the air/fuel ratio of the combustion engine to achieve a certain power level. They can then add the text "click here when done", and link the word here to a branching control. The branching control is of the real-number type, linked to *engine:get_power*, and configured to link to *text_frame:load:power_correct.htm* if the value is in the correct range and to *text_frame:load:power_incorrect.htm* if the value is out of the correct range. This model becomes more powerful since multiple links can be attached to the same hypertext or image-map region.

8.3 Assessment Components

Student performance assessment is an essential functionality. Our environment allows assessment/grading components to be easily integrated into exercises with the simulations. The functionality will include the ability for students to login to the assignment, and then the collection of milestone data. The author can create links to the assessment objects as the simulation is created. For instance, in the above pathogen exercise, the link to *text_frame:load:power_incorrect.htm* could be coupled with the link *grading_object:milestone:("power question",incorrect)*, and similarly for the correct response. (This illustrates the importance of multiple links from a single point.) The assessment component would then keep track of the number of attempts before the correct response was obtained. Grading may be viewed as a special case of this assessment model, where the grade is based on achieving certain milestones. We are currently developing methods to store milestone data in a network database, for which we will provide a simple implementation using an open source database such as MySQL. Extensions could include learning objects that interface to standard course management systems such as those offered by Blackboard and WebCT or with assessment tools developed for distance learning [1].

8.4 Interobject Communication

Another possible extension is to add an additional mode of inter-object communication to the viewer. For instance, objects may be allowed to publish or monitor messages on a bus, as in the InfoBus[1] standard. While this would allow programmers to provide authors with more powerful control panel objects than that described above, it is important to assess the effects of this added complexity on the author. In particular, will issues of synchronization lead to new types of errors, beyond the broken links of the current design, and potentially frustrate the curriculum author?

9. DIGITAL LIBRARY SERVICES

We stress that our approach is to start from a simple design and add in only those functionalities that retain a high level of ease-of-use. This assembly environment is not meant to provide a single all-purpose solution to component assembly. Rather it is meant to complement assembly environments currently available that are based on the simplified-programming model such as ESCOT. Design of assembly environments involves decisions of flexibility vs. ease-of-use. The emphasis in this tool is on ease-of-use via an extended-web-authoring approach that should be intuitive to instructors, and on the inclusion of features that help organize the potentially large number of text and images that make up a learning environment. To the extent that this environment does not support a more complex mode of component assembly, one of the tools based on the simplified-programming model may be used to assemble components into an object for inclusion in this environment. Thus the two classes of tools actually complement one another. Our tool fills an important need, since sole reliance on more complex tools would exclude potentially valuable contributors.

Our research then consists of the development of two services in support of digital libraries for education. First is providing an authoring environment that illustrates the "learning object" approach to creating on-line learning activities. This set of core learning objects such as a hypertext viewer and image-map viewer extend their web counterparts by supporting links to active content. Second is exploring the extent to which the piecing together of learning objects can be moved from the domain of the programmer into the domain of the curriculum developer. We anticipate that this shift will lead to both a larger community creating, adapting and using materials for the digital library and a better organized collection.

10. ACKNOWLEDGMENTS

This work was funded by the National Science Foundation as part of the NSDL (National Science Mathematics Engineering and Technology Education Digital Library) program.

[1] InfoBus is a standard for data communication between JavaBeans components.

11. REFERENCES

[1] Project ADEPT - http://www.users.csbsju.edu/~tcreed/adept/index.html

[2] AGENTSHEETS - http://www.agentsheets.com

[3] ARIADNE - Collaborative browsing project on digital libraries. http://www.comp.lancs.ac.uk/computing/research/cseg/projects/ariadne/

[4] American Physics Society - Physics Academic Software. http://webassign.net/pasnew/

[5] Advanced Distributed Learning Intiative network., DOD http://www.adlnet.org/

[6] BELVEDERE http://www.ics.hawaii.edu

[7] EDUCAUSE - http://www.educause.edu/

[8] EOE – http://www.eoe.org

[9] ESCOT - http://www.sri.com/policy/ctl/html/escot.html

[10] ESLATE - http://e-slate.cti.gr/

[11] IEEE Learning Technology Standards Committee (LTSC) http://www.manta.ieee.org/groups/ltsc/

[12] JAVASKETCH - http://www.keypress.com/sketchpad/java_gsp

[13] Meyers, B.J. & Jones, T.B. Promoting Active Learning: Strategies for the College Classroom. Jossey-Bass, San Francisco, 1993.

[14] NEEDS - http://www.needs.org

[15] NSFCCR - http://www.education.siggraph.org/nsfcscr/nsfcscr.home.html

[16] National Instruments Corp., LabVIEW – http://www.ni.com, 11500 N Mopac Expwy, Austin TX, 78759.

[17] Quicktime VR – Apple Computer Inc., http://www.apple.com/quicktime

[18] Smith, J. M. What Steve Jobs Did Right, The Educational Potential of the Ideas Behind NeXTSTEP. Educom Review, 1994.

[19] Squires, D. Educational Software for Constructivist Learning Environments: Subversive Use and Volatile Design. Educational Technology, May-June 1999,

48-54.

[20] Virtual Laboratory, The IrYdium Project, Carnegie Mellon University, Chemistry Dept.

A Component Repository for Learning Objects

A Progress Report

Jean R. Laleuf
Brown University
Department of Computer Science Box 1910
(401) 863-7658
jrl@cs.brown.edu

Anne Morgan Spalter
Brown University
Department of Computer Science Box 1910
(401) 863-7615
ams@cs.brown.edu

ABSTRACT

We believe that an important category of SMET digital library content will be highly interactive, explorable microworlds for teaching science, mathematics, and engineering concepts. Such environments have proved extraordinarily time-consuming and difficult to produce, however, threatening the goals of widespread creation and use.

One proposed solution for accelerating production has been the creation of repositories of reusable software components or learning objects. Programmers would use such components to rapidly assemble larger-scale environments. Although many agree on the value of this approach, few repositories of such components have been successfully created. We suggest some reasons for the lack of expected results and propose two strategies for developing such repositories. We report on a case study that provides a proof of concept of these strategies.

Keywords
Components, design, digital library, education, learning objects, NSDL, reuse, software engineering, standards.

1. INTRODUCTION

Our vision for digital library content goes beyond scanned literature or searchable curriculum materials to include richly interactive explorable microworlds that take full advantage of the ever-increasing power of computers, software, and networks. These learning environments combine the best qualities of live demonstration (see Fig. 1) with interaction only possible on the computer. Unfortunately, our research has led us to the conclusion that it is an unexpectedly huge effort to create a complete collection of interactive learning experiences for even a single introductory course in a given discipline.

We have been trying to accelerate the development process by creating reusable components or learning objects that can be recombined in different ways to produce sets of learning environments. By components or learning objects we mean standardized pieces of code, usually class files or Java beans, which programmers can easily reuse in different programs.

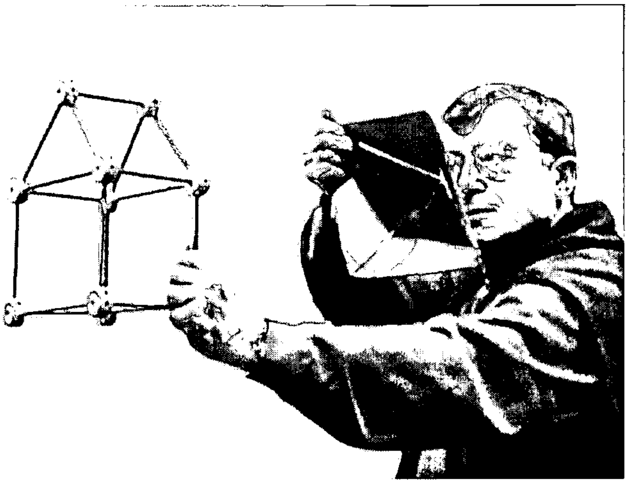


Figure 1: Professor van Dam uses a Tinkertoy house and a cardboard perspective viewing volume to demonstrate camera viewing transformations.

It may seem at first that creating a few dozen components would suffice for many courses. For example, an introductory calculus course would have a function-graphing component, some tools for finding derivatives and integrals, a series expander, expression editors and a few more objects, but the reality is far more complicated. The situation is similar to that faced by GUI designers. If one looks at an interface, it seems to be made up of a few basic elements, such as field boxes, sliders, buttons, and panes, but GUI libraries are huge and take enormous effort to develop.

Based on the GUI library analogy, we believe that a half a dozen programmers and researchers could never construct a major reusable educational components library in a reasonable amount of time. Only a concerted collaborative effort of tens of person-years can build the necessary content for any given field. In particular, although some underlying components, such as math libraries for learning objects, may be applicable across many science education domains, more domain-specific components must be created separately for each field.

The GUI library analogy assumes that we know exactly how to make these components, but in the field of educational software components there are many open research issues. For example, how does one analyze current simulations for decomposition into reusable components? How can one design components to be

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.

useful for educators (as well as programmers)? And how does one choose a proper level of granularity for a component?

We have been exploring two main strategies for repository creation. The first is the use of a categorization scheme to help programmers analyze and characterize types of components as they relate to educational purposes. The second is a method for addressing issues of object granularity, and determining what levels of object complexity are appropriate. We have applied these strategies in a proof-of-concept case study.

The work described here is part of an NSF NSDL grant, the CREATE project (A Component Repository and Environment for Assembly of Teaching Environments).

2. PREVIOUS WORK

It has long been hoped that instructional software developers would contribute to libraries of instructional software that would give others access to the benefits of software simulations and interactive exercises. There are a number of significant efforts to create develop and share educational software components, most of them based on the Java language. The ESCOT project is a testbed that seeks to encourage development and reuse of learning objects, with a current emphasis on middle-school mathematics, but their components are not available for general public use [6]. The E-Slate company sells educational components (as opposed to finished applications) and describes about two dozen of them on their Web site [7]. The Educational Object Economy project compiles interactive educational tools in the form of complete applets [5]. As far as we can tell, none of these undertakings provides complete sets of components for specific courses and even if they become quite successful, are aimed chiefly at educators with little or no programming experience. There is still a need for the digital library to house lower-level components, to be used by programmers to create fully customized educational environments. Further discussion of educational component use can be found in the IEEE Computer September 1999 Special Issue on Web based learning and collaboration [13].

In addition to work specific to educational software, the technical, social, economic, and administrative problems with general computer code reuse are now better understood [4, 11]. Problems include failure to organize and index reusable objects, failure to mandate that code be designed with reuse in mind, lack of an organizational structure dedicated to supporting reuse, and failure to recognize the domain dependency of reuse strategies.

The Exploratories project at Brown University, on which this paper's work is based, has worked for over five years to create learning objects with high levels of interactivity [8]. Our chief content area has been introductory computer graphics, including introductory linear algebra [2, 17]. We think of exploratories as combinations of "exploratoriums" [9] and laboratories, realized as two- and three-dimensional explorable worlds which are currently implemented as Java applets. Our applets are embedded in a hypertext environment and are used by a number of high school- and college-level courses around the world. They are freely available at our Web site [8].

When we began trying to reuse frequently occurring program elements, we quickly found that simply copying and pasting code was not a good strategy. Most classes worked well only in the program for which they were initially designed. The problems for reuse ranged from a lack of software interface standards (such as those imposed by the Java beans spec [16]) to difficulty in arriving at the right level of feature complexity. In the end, we found that everyone wound up rewriting the "reusable" elements. The strategies and project discussed in this paper were inspired by this situation.

The Exploratories project has tried to promote other aspects of reusability by creating reusable hypertext structures for Web-based curricula [19], describing methods for integrating learning objects into traditional curricula [20], categorizing pedagogical approaches and teaching techniques that can be used for interactive learning environments [18], and creating a Web-based repository structure for JavaBeans [3].

2.1 A Component Categorization Strategy

When we began making components, we found that they fell naturally into three different categories. We characterize all reusable educational components as either 1. Core Technologies, which have a high degree of domain independence and are typically quite fine-grained (e.g., Java GUI classes), 2. Support Technologies, which usually have some domain dependence and are typically medium-grained objects (e.g., BEA Systems JavaBeans designed for e-commerce (includes shopping cart beans, order tracking beans, etc.), or 3. Application Technologies, which are almost always highly domain dependent and coarse-grained (e.g., a Java applet that teaches about a particular chemical reaction). These categories are described in more detail below:

2.1.1 Core Technologies

Characteristics

- Domain independence
- High levels of reusability
- Self-contained functionality
- Adherence to high standards of design and reliability

Audience

Chiefly programmers but also some content developers (with little or no programming ability) using assembly tools

Granularity

Typically fine-grained (e.g., sliders, timers and buttons) but can also include coarse-grained objects (e.g., spreadsheets or even an entire pedagogical framework into which one can plug one's materials [19])

Examples

- Java GUI classes
- Various math function libraries
- IBM AlphaBeans [1]
- 3D Interaction and visualization widgets

Notes

Many commercial offerings fall into this category, but such components may also be built in-house.

2.1.2 Support Technologies

Characteristics

- Domain dependence
- Moderate levels of reusability
- Self-contained functionality

Audience

Chiefly programmers but also some content developers (with little or no programming ability) using assembly tools.

Granularity

Typically medium-grained objects, often aggregating finer-grained components (e.g., an image filtering widget that combines slider and windowing components)

Examples

BEA Systems JavaBeans designed for e-commerce (e.g., shopping cart beans, order tracking beans, etc.)

2.1.3 Application Technologies

Characteristics:

- Domain dependence
- Minimal levels of reusability
- Each object is a self-contained, fully-functional application or applet that attempts to achieve an educational goal.

Audience:

End users without programming experience. Application Technology components can be combined by both programmers or non-programmers (using a visual environment) and can be part of a larger pedagogical structure (e.g., a game or lab) or a full-blown curriculum.

Granularity:

Typically coarse-grained, although useful distinctions can be made about the level of concept granularity of these end products (e.g., fine-grained applets teaching a single concept vs. coarse-grained applets teaching a number of concepts in one module)

Examples:

- An applet that teaches characteristics of a particular chemical reaction.
- An applet that teaches how to make a scenegraph in 3D graphics.

Notes:

These objects may also be part of a larger group of similar objects (e.g., a series of applets to teach increasingly complex topics in a subject). Because these objects attempt to achieve an educational goal and are targeted primarily at end-users, pedagogical, user interface, and information structure concerns are highly important in the creation of objects for this category.

Note that in comparison to other fields, especially e-commerce, educational technology is particularly deficient in the support technologies category. We know of no substantial repositories for such domain-specific components and therefore observe educational software developers repeatedly building these types of components. We believe that one of the chief problems facing current efforts in educational component technology is a lack of distinction between support and core technologies. This is compounded further by a deficiency in understanding the design features necessary to promote reuse of support technologies.

2.2 A Granularity Strategy

Although the categories described above provide guidance as to what levels of granularity are appropriate for specific types of components, individual components may in fact be of any granularity regardless of the category in which they fall. In our experience of creating components, we found that, in fact, creating them solely at any single level of granularity either meant having too many features and not enough flexibility, or having so fine-grained a set of components that a great deal of work still needed to be done to create the final product.

Our strategy for dealing with granularity and organizing the results is, for a given component, to produce a **complete** set of sub-components, thus providing objects at **all** levels of granularity, from application technology components (e.g., self-contained interactive applets) down to core technology code (e.g., a coordinate system package which includes a complete set of interfaces and behaviors for coordinate systems). This decomposition is important to complete even if there is no current need for some of the components it generates.

When this strategy is employed, programmers given a component from any category will be able to customize that component and also to retrieve, customize, reuse and reconfigure any sub-components it might aggregate. Teachers working with programmers would have virtually no constraints on how they could reconfigure what they see on the screen to meet their own needs. By including all levels of component granularity in the development process, we hope to create component repositories that will be broadly applicable to diverse areas of science education. In addition, completely decomposed components describe a hierarchy that is helpful for documentation purposes. The downside of this strategy is that it requires a large upfront investment, as discussed in the Conclusion.

3. Case Study: The Camera Viewing Transformation

Our case study is based on a set of applets that teach students in an introductory graphics course about 3D camera transformations.

One of the basic tasks in computer graphics is generating a representation of a three-dimensional scene and displaying it on the two-dimensional computer screen. This is performed in a manner very similar to positioning a camera in front of a real-world scene, taking a photograph, and looking at the resulting image.

For the sake of mathematical simplicity and efficiency, a series of manipulations must be performed on the scene and its geometry before its two-dimensional representation can be drawn on the screen. Pedagogically, these are best described as changes in position and shape of the objects concerned. The continuum of these changes is of particular interest to students.

3.1 History

3.1.1 Text Illustrations

Despite the inherently continuous nature of the camera transformations material, the canonical reference text in computer graphics, Computer Graphics: Principles and Practice [10], provides only five pairs of discrete snapshots of these continuous manipulations (see Fig. 2). Unfortunately, these images are at best hard to decipher and are typically found confusing and unintuitive. The essential difficulty of these images is that they are preset, providing no opportunity for exploration or discovery through manipulation of the scene and the operations it undergoes.

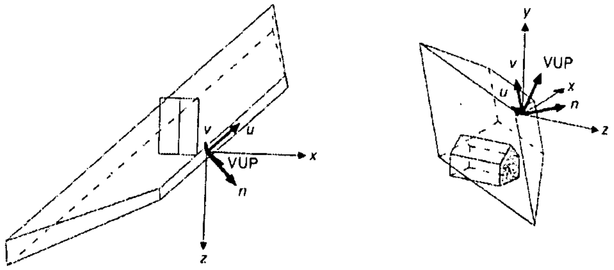


Figure 2: One of five pairs of diagrams used to illustrate the camera viewing transformations.

3.1.2 Models used in class

In his class on introductory computer graphics, Professor Andries van Dam does a lot to impart the continuous nature of the scene transformations through the use of Tinkertoy props, as shown in Fig. 1. This resolves many of the issues presented by the book's illustrations, as he is able to move the props around and explain how they move and change orientation in time. Students see the continuity of motion and more fully grasp the concepts presented to them. It is difficult, however, to show more than one model changing at a time and Tinkertoys lack the flexibility necessary to accurately display the concluding mathematical transformations. These include scaling and other non-rigid body, distorting transformations, requiring much hand-waving by the professor and implicit visualization by the students.

3.1.3 Customized software

One of the first attempts at using computer technology to resolve the problems above was a program written at Brown for an Evans and Sutherland vector display in the late '70s [12]. The resulting program allowed one to manipulate the parameters of the computer camera and then have the program animate the succession of transformations to produce a 2D image. Students observed the continuity of the changes and saw how one shape smoothly turned into another over time. In the early '90s the program was reimplemented for a raster display, again using an experimental software system.

This second version was more powerful and compelling than the first. It was heavily used for many years but had several limitations, such as the fact that it only ran in our research lab. In addition, because of the nature of the program design, it was impossible to easily modify it and when new viewing models were introduced in the class, the demo became somewhat confusing. As with all software, it also suffered from code rot. Again, the custom nature of the software made this difficult to address. When the underlying software system was abandoned by the research group, the program eventually ceased to function. A video was made, but the lack of interaction greatly diminished its usefulness.

3.2 Current incarnation

The idea of an interactive, computer-based educational tool describing the camera transformations was revisited in early fall of 1999 when one of the course's teaching assistants offered to throw together a rough prototype replicating a subset of the FLESH program's capabilities.

Working off this successful prototype, we engaged in a more formal design process, aiming to better think out the various pedagogical considerations and also use the new applet as a proving ground for our ideas on component design and organization. We thought through the potential components in terms of our three categories and began to decompose each one according to our theory of granularity completeness.

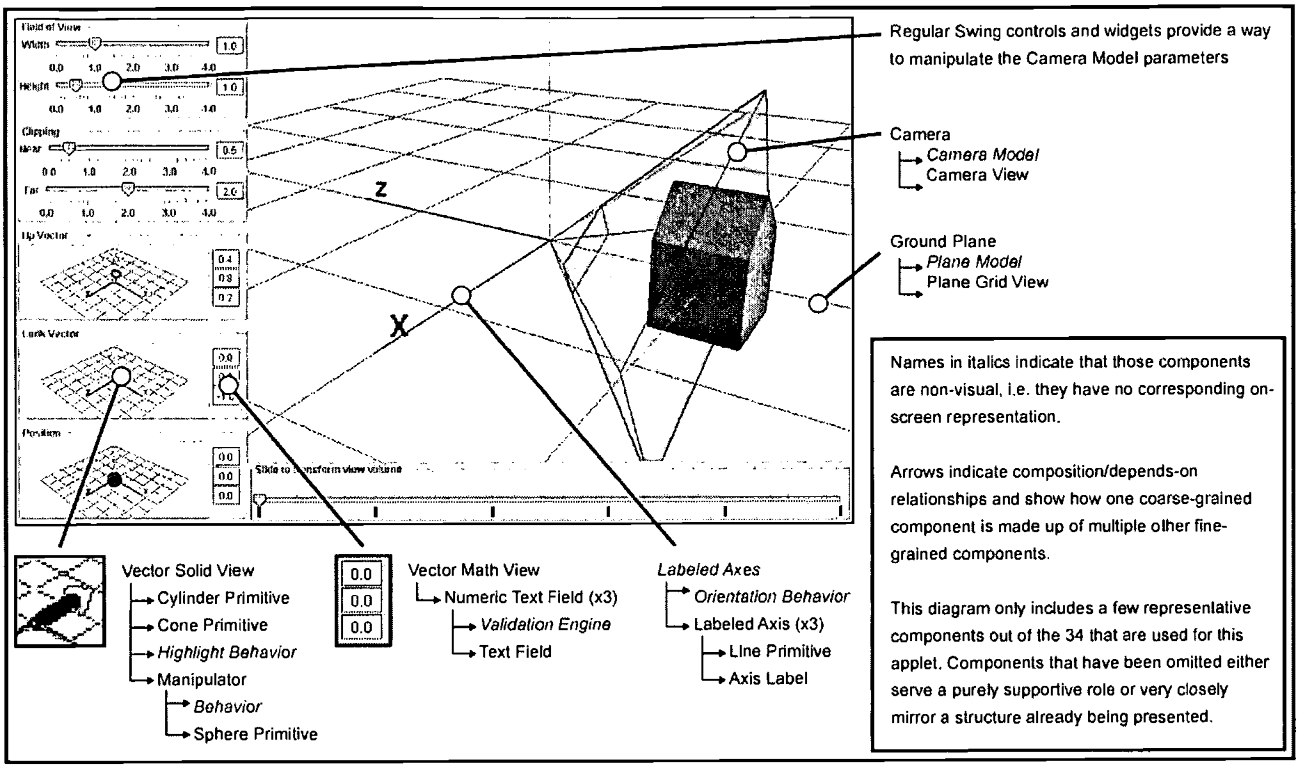
3.2.1 Pedagogical considerations

We sought primarily to provide a means by which a user could visualize the various manipulations undergone by a synthetic scene when its two-dimensional representation is generated. It needed to be useful for the professor demonstrating the concepts to his class during a lecture as well as accessible to students who wanted to revisit the material and explore and experiment on their own time.

To these ends we sought to have our new versions encompass all the capabilities of other methods of teaching the material (such as the diagrams and models) while enabling a students to freely explore and answer any question that might occur to them. We chose to provide an interactive, three-dimensional computer rendering to allow one to view the scene from multiple angles. We also gave the users complete control over the passage of time in the simulation, allowing them to advance or backtrack as needed to better understand a particular concept. Furthermore, we provided full, interactive and graphical control over the various camera parameters (position, orientation, field of view, etc.) at all stages of the simulation in order to offer students as many possibilities for exploration as possible.

3.2.2 Component structure

In Fig. 3, we see a sample of the components used in the current incarnation of the camera viewing applet. The transparent gray truncated pyramid represents the computer's perspective view volume, and the scene being viewed consists entirely of the simple house. A representative component structure, the *Vector Solid View*, has been selected to show how a coarse-grained, high-level component can be broken down into fine-grained, customizable components.



Names in italics indicate that those components are non-visual, i.e. they have no corresponding on-screen representation.

Arrows indicate composition/depends-on relationships and show how one coarse-grained component is made up of multiple other fine-grained components.

This diagram only includes a few representative components out of the 34 that are used for this applet. Components that have been omitted either serve a purely supportive role or very closely mirror a structure already being presented.

Figure 3: Decomposition of a sample set of the components used in the camera viewing transformation applet.

Vector Solid View:

Cylinder primitive
The cylinder primitive provides a visual representation of the stem of the vector's arrow representation. It would be relatively simple to replace it with another primitive such as a rectangular solid.

Cone primitive
Serving a function similar to the cylinder primitive component, the cone primitive provides a visual representation of the vector arrow's head. As with the cylinder, changing the type of primitive used would be trivial.

Highlight Behavior
This non-visual component provides user feedback functionality by highlighting the head and stem of the vector arrow when the user's mouse passes over them. This allows users to intuitively understand that something will happen if they click on the different parts of the vector.

Manipulator
This component allows users to visually interact with the vector, dragging it around to change its orientation. It is composed of three different components, one of which has been omitted from the diagram for simplicity's sake but which will be described below.

> *Behavior*
> This component plugs into the Java3D interaction framework to allow users to interact with the vector by dragging it to the desired position.
>
> *Sphere Primitive*
> This object provides visual feedback to users while they drag the vector arrow to a new orientation.
>
> *Spherical Geometry Intersection Utility Class*
> Although omitted from the diagram, this non-visual component serves a critical role by performing the calculations necessary for the Behavior component to function correctly. Encapsulating this functionality in a single component facilitates long-term maintenance, allowing it to be replaced at a later date with a more efficient implementation should one be developed.

3.3 Reusability

Having decomposed the Camera Viewing Transformation Applet into a number of fine-grained components, one can now turn to the issue of these components' reuse. More specifically, one must look at whether or not these components are at all reusable outside of the one or two applets that were in the designer's mind when they were created.

Fig.s 4-6 show the component usage for three different classes of components. In Fig. 4, the camera viewing applets all reference the same set of components. Indeed, our four different viewing applets (Parallel Camera Parameters, Parallel Camera Transformation, Perspective Camera Parameters, and Perspective Camera Transformation [8]) each teach different aspects of the transformation process but use the same set of components, passing in different parameters to get the desired effects. Notice that the camera applets use a mixture of both core technology and support technology components. Also, some support technology components also reuse core technology components. For example, the Camera Interpolator component reuses all the components in the 3D Interpolators core component package.

In Fig. 5, we see that an applet that reuses many of the components used in the camera viewing applets. This applet helps students understand the rendering process called radiosity, in

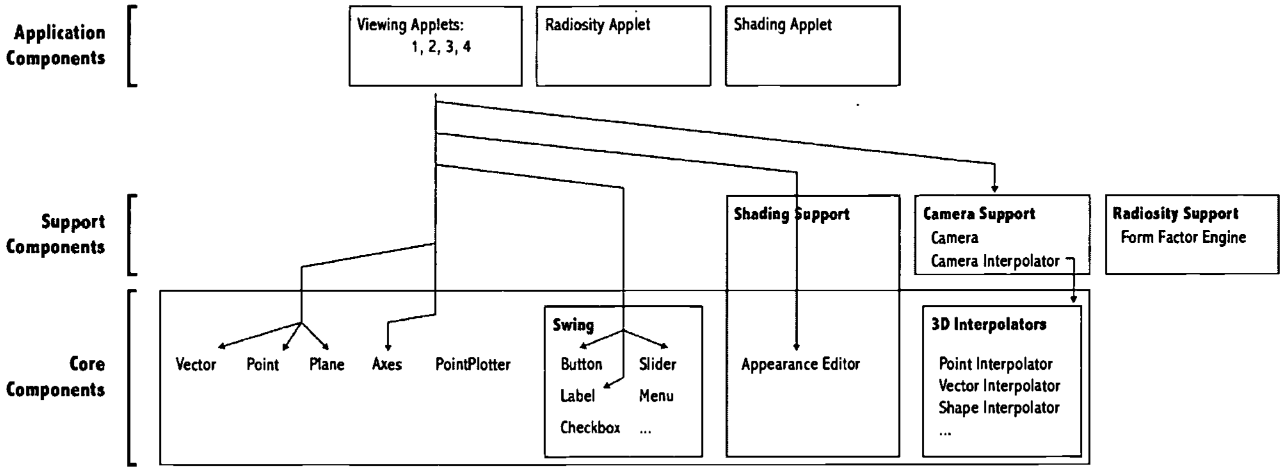


Figure 4: The component reuse graph for the four camera viewing transformation applets.

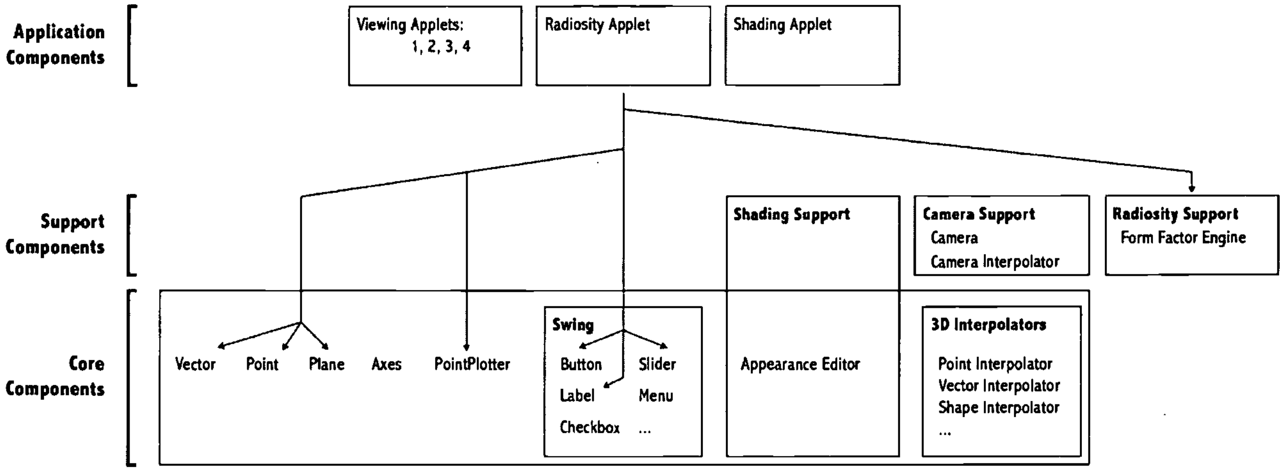


Figure 5: The component reuse graph for the radiosity applet.

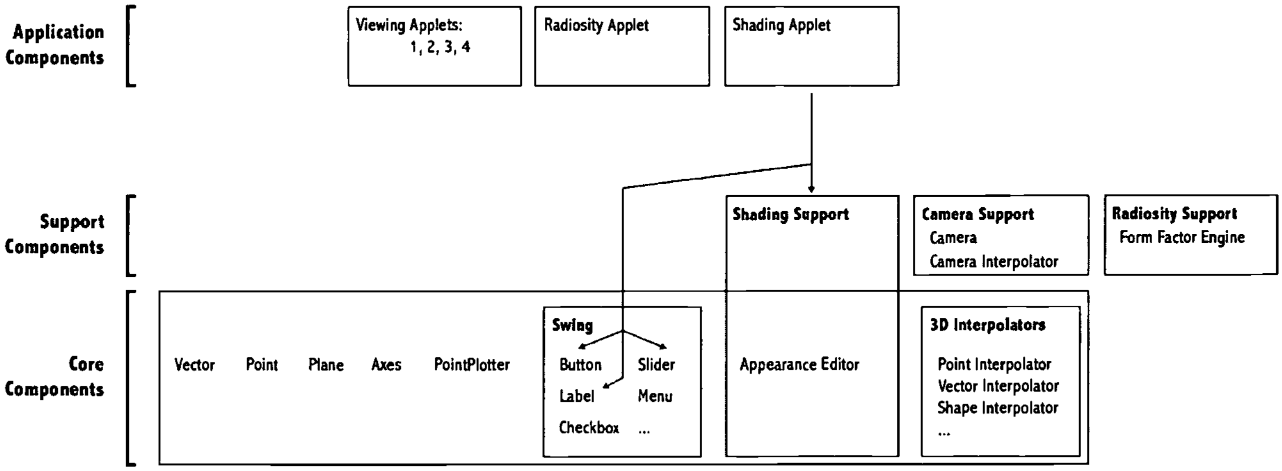


Figure 6: The component reuse graph for the shading applet.

which energy transport is simulated to calculate diffuse light reflections in a scene. We see reuse of the Vector, Point and Plane core technology components as well as the standard Swing components. The Axes component has not been reused because the PointPlotter component is being used instead. Notice also that although the components in the Camera Support package are not being used, those in the Radiosity Support package are. This is in keeping with the fact that support technology components are typically domain-specific and, although reusable for similar applets, are generally not reusable by applets from different domains.

Fig. 6 demonstrates that components can, in some cases, be classified into two different categories depending on how they are used by the application component. In this case, the Appearance Editor component is being used as a support technology component because it plays a central supportive role for the shading applet. Contrast this to Fig. 4, in which this very same component acts as a core technology component with respect to the camera viewing applets. This ability of components to migrate from role to role depending on how they are used by other components adds more flexibility to the classification schema laid out above and thereby enhances its power.

3.3.1 Standards

All of our components adhere to the JavaBeans specification because doing so facilitates integration of components into commercial software design packages. A properly-designed JavaBean, for instance, can be used without any problems in Sun's BeanBox, Inprise's JBuilder, Forte's Netbeans IDE, or IBM's VisualAge. It can be used with equal success in the educational authoring environments being produced by our colleagues at ESCOT and e-Slate. The universal acceptance of the JavaBeans standard therefore makes it a valuable requirement for all our component designs and will make our components that much more powerful down the road.

3.4 Future Work

Our digital library grant is a collaborative one with chemistry professor Dave Yaron at Carnegie Mellon. Professor Yaron and other members of the grant team at Carnegie Mellon have been creating environments in which non-programming educators can customize interactive experiences and embed them in their own pedagogical materials. We have begun to join our different approaches to the problem of reuse and customization in the digital library efforts in a collaborative project to aid chemistry education.

3.4.1 Molecular Visualization Applet

We are working on a molecular visualization applet to increase the reuse of existing components as well as introduce new components developed in-house and by third parties.

A current prototype aims to reuse the plane model and view components as well as some of the core 3D object manipulation components that allow users to intuitively move objects in a 3D space.

This prototype expands the core technology space by using a VRML file loader imported from a third party source and using it to import VRML models retrieved from the Protein Data Bank, thereby leveraging content from an existing digital library. We are also planning on introducing a precise collision detection package to the core packages as well as more basic 3D interaction widgets.

Finally, we will provide a protein docking engine component that might find significant reuse in other software dealing with protein interactions.

The resulting set of components will be glued together using our Carnegie Mellon collaborators' non-programming environment, thereby providing the power of various in-house and third-party software components in an easy to use yet powerful building tool [15].

3.4.2 Increased Compatibility Effort

Recognizing that it would be a grave mistake to isolate our work from complementary research and development performed by other research groups, we will also be maximizing compatibility between the components we develop and development environments and test beds such as those produced by the E-Slate and ESCOT projects. Recent developments have made these environments more compatible with the JavaBeans standard and we anticipate increased productivity by leveraging the tools they provide.

3.4.3 Metadata Standards for Harvesting

We have started tagging the applets we produce with standard IMS [14] metadata tags in order to make our efforts easily harvestable by collectors of digital library material. In the near future, we intend to extend this effort by tagging individual components, thereby allowing harvesters into our component libraries as well as our applet catalogs.

4. DISCUSSION AND CONCLUSION

When we proposed our framework and granularity strategy as part of our grant application, we had limited experience using either one for real-life applications. By creating the camera viewing transformation applet set, we were able to test out these theories in practice. We feel that the results offer a proof of concept of these strategies.

We have found, however, that there are also disadvantages to using a structured, component-centric approach with an emphasis on reusability. Although the up-front design results in components that are well designed, this design methodology greatly increases the length of the development cycle. For example, the viewing techniques applets used in our case study took roughly four months to design, write and test. When compared with the four days it took to write the applet's initial buggy prototype, it may not seem to have been worth the time or effort. In addition, it takes more skilled programmers to develop truly reusable components, and taking this approach meant that we could no longer rely solely on undergraduates, our previous source of programmers.

If a digital library of reusable learning objects is ever to become a reality, however, we must continue to invest the upfront time. In our case, although we took four months to design, implement, and test a single applet, we also produced several dozen components that we consider to be stable and generally reusable. Indeed, they were easily reused by the undergraduate who programmed the radiosity applet. For us, the long-term benefits of this far outweigh the short-term gains of the hacked-together prototype, whose code was not reusable.

We believe the complete decomposition approach we have adopted for component creation allows us to offer full customization of components by content authors, thereby increasing the degree to which individual components are reused. We are also now convinced that the effort and time we must devote to designing and developing not only good components but also a large amount of underlying infrastructure will enable us to reap huge rewards down the line and significantly enhance the field of educational software component technology.

Although starting a good repository of reusable learning objects is a time-consuming and expensive task, ultimately doing so should dramatically reduce the time needed to create interactive learning environments. Not only will mature programmers' time be reduced but inexperienced developers will be able to make sophisticated learning environment by re-using learning objects that encapsulate advanced functionality that they would not be able to easily program on their own.

Developers' new creations and the new components that are a part of them, can, in turn, be contributed to the library, creating a snowball effect of additional content, both in complete instructional applications and in new building blocks from which future applications can be built.

5. ACKNOWLEDGMENTS

We would like to acknowledge helpful comments of Andy van Dam and contributions to the Exploratories project from Rosemary Michelle Simpson. We would also like to thank our sponsors, the National Science Foundation, through our NSDL and STC grants, and Sun Microsystems.

6. REFERENCES

[1] IBM. alphaBeans: JavaBeans by IBM, http://alphaworks.ibm.com/alphabeans/

[2] Jeff E. Beall, Adam M. Doppelt, and John F. Hughes. "Developing an Interactive Illustration: Using Java and the Web to Make It Worthwhile," in Proceedings of 3D and Multimedia on the Internet, WWW and Networks, 16-18 April 1996, Pictureville, National Museum of Photography, Film & Television, Bradford, UK, 1996.

[3] BeanHaus. Java Bean Repository, http://www.beanhaus.org

[4] Stephanie Doublait. "Standard Reuse Practices: Many Myths vs. a Reality," in Standard View, Vol.5, No. 2, June, 1997.

[5] EOE Foundation. Educational Objects Economy: Building Communities that Build Knowledge, http://www.eoe.org.

[6] ESCOT project. Educational Software Components of Tomorrow, http://www.sri.com/policy/ctl/html/escot.html.

[7] E-Slate project. An exploratory learning environment, http://e-slate.cti.gr/

[8] Exploratories project. Web-based educational software, http://www.cs.brown.edu/exploratory/

[9] Exploratorium. The San Francisco Exploratorium: museum of science, art, and human perception, http://www.exploratorium.edu/

[10] James D. Foley, Andries van Dam, Steven K. Feiner, John F. Hughes. "Computer Graphics: Principles and Practice," Addison-Wesley, 1996, ISBN 0-201-84840-6.

[11] Richard P. Gabriel. "Patterns of Software: Tales from the Software Community," Oxford University Press, August 1996.

[12] R. F. Gurwitz and R. W. Thorne and A. van Dam and I. B. Carlbo. "BUMPS: A Program for Animating Projections," in Proceedings of ACM SIGGRAPH '80, pp. 231-237, 1980.

[13] IEEE. "Web-Based Learning and Collaboration" special issue, Computer, Vol. 32, No. 9, September 1999.

[14] IMS Global Learning Consortium, Inc., http://www.imsproject.org/

[15] IrYdium Project. Java Enhanced Chemical Education, http://ir.chem.cmu.edu/irProject/

[16] JavaBeans. "Specification for the Java 2 Platform," http://java.sun.com/products/javabeans/glasgow/

[17] Rosemary Michelle Simpson, Anne Morgan Spalter, and Andries van Dam. "Exploratories: An Educational Strategy for the 21st Century," in ACM SIGGRAPH '99 Conference Abstractions and Applications, 1999.

[18] Anne Morgan Spalter, Michael LeGrand, Saori Taichi, and Rosemary Michelle Simpson. "Considering a Full Range of Teaching Techniques for Use in Interactive Educational Software: A" Practical Guide and Brainstorming Session, in Proceedings of IEEE FIE 2000 (Frontiers in Education), October 2000.

[19] Anne Morgan Spalter and Rosemary Michelle Simpson. "Reusable Hypertext, Structures for Distance and JIT Learning," in Proceedings of ACM Hypertext 2000, June 2000.

[20] Anne Morgan Spalter and Rosemary Michelle Simpson. "Integrating Interactive Computer-Based Learning Experiences Into Established Curricula," in Proceedings of ACM ITICSE 2000 (Innovation and Technology in Computer Science Education), July 2000.

Designing e-Books for Legal Research

Catherine C. Marshall,[1] Morgan N. Price, Gene Golovchinsky, Bill N. Schilit

FX Palo Alto Laboratory, Inc.
3400 Hillview Ave., Bldg. 4
Palo Alto, CA 94304 USA

cathycmarshall@yahoo.com, MorganNPrice@yahoo.com, {gene, schilit}@pal.xerox.com

ABSTRACT

In this paper we report the findings from a field study of legal research in a first-tier law school and on the resulting redesign of XLibris, a next-generation e-book. We first characterize a work setting in which we expected an e-book to be a useful interface for reading and otherwise using a mix of physical and digital library materials, and explore what kinds of reading-related functionality would bring value to this setting. We do this by describing important aspects of legal research in a heterogeneous information environment, including mobility, reading, annotation, link following and writing practices, and their general implications for design. We then discuss how our work with a user community and an evolving e-book prototype allowed us to examine tandem issues of usability and utility, and to redesign an existing e-book user interface to suit the needs of law students. The study caused us to move away from the notion of a stand-alone reading device and toward the concept of a document laptop, a platform that would provide wireless access to information resources, as well as support a fuller spectrum of reading-related activities.

Keywords

e-books, information appliances, field study, physical and digital information resources, legal education, legal research, digital libraries.

1 INTRODUCTION

Dynabook, Alan Kay's imagined dynamic electronic medium [9], is often cited as an inspiration for current work on electronic books. Now we have real products (e.g., RCA's REB1100), tablet computers for reading. We wanted to explore the future of such devices as interfaces to today's heterogeneous libraries, and to understand how they could support research activities typical among knowledge workers who use such information resources. We used XLibris [15], analytic reading software running on a pen tablet computer, as an example of an e-book, and legal research as a discipline in which analytic reading software could bring value.

When we began our study, we conceived of ideal e-books as stand-alone reading devices that would be based on a paper document metaphor, would use pen interaction and freeform digital ink, and would support research activities by using readers' annotations as indications of their interests [16]. Indeed, XLibris (shown in Figure 1) was just such a working prototype, a good foil for our investigation of the kinds of "beyond paper" functionality an e-book could bring to bear on legal research.

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

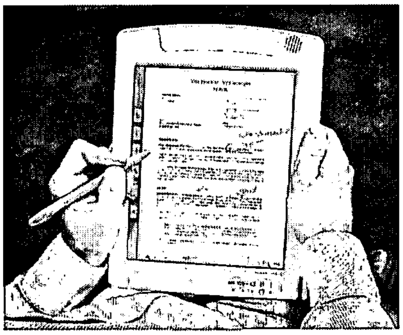


Figure 1. Our prototype e-book at the outset of the study: the XLibris analytic reading software running on a Fujitsu pen tablet computer.

We chose legal education for our study for several reasons. Early discussions with our research colleagues suggested that attorneys read and mark up documents from diverse sources, from physical as well as digital collections [4]. Furthermore, attorneys' reading is purposeful: they use such documents in subsequent work activities, including writing and collaboration [1]. These practices have their roots in legal education.

By studying legal education, we wanted to do more than characterize current practice; we also wanted to evaluate existing designs and to generate new design insights about both the usability and utility of XLibris. What would an e-book for legal research actually do for its users? We started by observing how paper and online resources are used in legal research; how law students read, annotate, organize, and use their materials; and the role of legal research and reading in a larger scope of activities like writing and collaboration. We then worked with this potential user community to evaluate the effectiveness of XLibris. Our hope is that insights that we gathered from our study of legal research, reading, writing, and collaboration would be more generally applicable in other educational and research settings, as well as in legal work.

In the following sections, we describe the study, our observations and the broad implications for design. We conclude with the details of the redesign that emerged from our observations.

2 STUDY DESCRIPTION

Our field study focused on an annual Moot Court competition at a first-tier law school. Moot Court is the venue in which students practice advocacy as they argue hypothetical cases. Controversial issues—cases heard in appellate-level courts that suggest

[1] Author's current address:
Microsoft Corporation
One Microsoft Way - 32/1079
Redmond WA, 98052

cathymar@microsoft.com

unresolved points of law—typically form the basis for Moot Court problems. We chose Moot Court for our field study because it is characterized by one of the major legal digital library providers as the closest experience a law student has to preparing for and engaging in real courtroom advocacy.

The students start their research from a Transcript of Record that lays out the facts of the case and cites relevant prior cases. From their research, they retrieve, print, read, and annotate cases, and consult secondary materials such as law journals. The materials they collect are organized and used to produce a brief, a document of constrained length that presents a position on the legal issues; these materials form the basis for the oral arguments as well.

Our study took place over three months, from the distribution of the transcript of record to the final competition. We interviewed and observed the two faculty members who organized the competition, and nine second- and third-year students who participated in the Moot Court competition. The interviews were open-ended and semi-structured; we observed the students and faculty interacting with online resources, meeting to coordinate writing and research tasks, and attending classes. Interviews and observations took place where the participants normally worked, in settings like the law library, shared on-campus offices, and dorm rooms. We audio-recorded and transcribed interviews, photographed salient aspects of the work settings, and took field notes of our observations.

After the competition ended, we collected the students' and faculty members' documents—source materials they had drawn on for their research and the briefs they wrote for the competition. We analyzed these documents to understand patterns of annotation.

To assist us in the design process, interviews concluded with a demonstration of the evolving XLibris prototype. We used documents drawn from the Moot Court research in the demonstration; these familiar documents made the system more transparent and more compelling to the law students. The students used the device briefly, commented on specific design elements and functionality, and reflected more generally on how it might be useful in their work.

3 RESULTS

We divide our findings into two parts: a characterization of current practice and its implications for the design of an information appliance for obtaining, reading, annotating, and organizing digital materials in the legal domain; and design insights about the existing technology, XLibris. This way, the characterization and insights are not constrained by limitations of existing technology, yet the technology can grow and be shaped by the field study.

To set the stage, we first discuss work settings and mobility. We then discuss legal research, the traditional province of online legal information services and law publishers. Reading, organizing, and annotating documents form a core set of topics when we talk about working with legal documents. Finally, we examine reading-related practices, writing and collaboration.

Our XLibris-related findings are related to the usability and utility of the device itself, in addition to more general reactions that the demonstration provoked. We then use these findings about legal work and the prototype as a basis for the redesign.

3.1 Work settings and mobility

The law students worked in a variety of settings, each of which offered them access to unique local resources. Settings included the law library, especially work areas such as the computer/printer room and carrels; shared offices, mostly associated with the law reviews published at the law school; dorm rooms and homes; classrooms; other law libraries and other on-campus libraries; and *ad hoc*, unpredictable work sites, especially those used while traveling.

Important localized physical resources at these settings included computers (shared and personal) that provided access to services and applications, network connections, printers (in particular, the free printers in the law library; these are supplied by Lexis and Westlaw legal information services), paper books and journals (in personal collections or at the library), knowledgeable people (e.g. faculty, Lexis and Westlaw representatives, librarians, peers), comfortable, quiet places to read and write, places to store materials across uses, and places to spread out materials during a work session.

Distributed local resources give rise to an increased need for mobile work habits. Thus, like other students and an increasing number of professionals, law students may work in many places on any given day, carrying materials with them in heavy backpacks. For example, even if the students do legal research from their home computers, they frequently re-retrieve materials at the law library so they may be printed without cost.

This resource-centered mobility also causes the students' documents to be decentralized. When the documents are electronic, they are not necessarily stored on a server, but rather stored on computers' local disks or transported on floppy disks (which affects the way the students share files as well). Paper documents may be kept at various locations—for example, on-campus lockers, dorm rooms, library carrels—or taken along in the students' backpacks.

From settings and mobility to design

What are the design consequences of the students' current patterns of resource-centered mobility? Mobility may mean many things: a student may move from a desk to a nearby comfortable chair to read a case; a student consulting a legal treatise in the library may go to a different floor to re-retrieve a case and print it; or a student may be traveling, and use another university's law library. For these mobile work situations, a portable reading device can form the bridge between paper resources and electronic ones; the hardware may be brought to where associated resources are available. This physical/digital bridge function, coupled with the variability of network access in the various work settings, underscores the importance of *wireless* access and the need to consider how materials get on and off the device, a finding that confirms Jones et al.'s study [8].

Our observations also echo those of Elliott [5], who interviewed judges about their use of online legal resources. She reports that having Lexis terminals in chambers was not convenient, as it forced the judge "to excuse himself in the middle of a trial, go to his chambers, dial into Lexis/Nexis, print out a citation, and take it back to the bench to read."

The variability of settings also suggests that the ability to keep materials in place across sessions cannot be taken for granted; for example, shared tables or workspaces in the law library (besides personal carrels) must be cleared off. Thus our design should take advantage of the computer's ability to readily support casual, persistent layouts of many documents in a workspace.

3.2 Legal research

Legal research is the process of gathering the materials together *to meet the needs of the task at hand.* As Sutton [17] points out, criteria of whether a given item is relevant are based on use, not on abstract notions of topicality or on a set of rules governing relevance judgment. Legal research might involve consulting books or law review articles in the law library, grabbing last semester's textbook off a bookshelf, searching an online service, pursuing specific case citations, or other strategies for amassing relevant background material for brief writing and oral advocacy.

Students began their Moot Court research by identifying a key case or cases from the Transcript of Record. Is this artificial? Not really; subsequent conversations with attorneys and the reflections of the students themselves showed that much legal research starts with some knowledge of an important case or cases in an area. If this information is not available, research is often initiated by consulting a treatise (an encyclopedic reference that summarizes the issues and case law in a specific area) or a law review article. Students used expressions like finding a "launching pad," "raid[ing] the cases," or "looking for a thread to pull." One student described her experience doing research in the books as a summer associate:

> "The first firm I worked at was very pro-books... I was pretty much taught to look in Witkin first... It's a California law treatise. ...it'll give you case citations, and then you can narrow your search that way. Once you have a case on point, or a case kind of on point, you can Key Cite it or something."

Once the students got started on their research, they continued to use citations as points of departure. They used citations in two different ways. First, citations are obvious links to precedent. If a student sees the same citation over and over again, referenced from multiple cases, it may well be valuable for current work. Students kept lists of cases to look for next or annotated case printouts with proposed follow-up citations. As we saw in an earlier study [12], these potentially interesting references may not be pursued, given limitations of time and attention.

Second, they evaluated citations, not just by looking at them, but by investigating if they are still "good law"—whether, for example, they have been overturned—and whether they are sufficiently authoritative. In US law the authority of a precedent is intimately tied to its currency, to the court in which it was decided, and to whether or not it is a good fact match to the current case. Reverse citation facilities such as Lexis's Shepards or Westlaw's Key Cite are typically used to determine whether a case is still good law. Both services provide annotated "back links" to, and metadata for, the cases that have cited the case in question. Thus this activity also amounted to link following.

Does this tendency to follow and evaluate explicit citations mean that the students do not perform full text searches? They describe searching in a few situations, although it seems to be of secondary importance if a starting point—a key case, a treatise, or even a comprehensive law review article—is at hand. Full-text search is used to identify a key case at the outset if none is available, to check breadth and coverage if there is time, and to look for very recent cases, as Shepards links may lag court decisions by as much as six months [2].

Important resources for legal research continue to be a mix of physical and digital materials; paper books still play an important role in legal research, but they are used in concert with online legal services. These materials may be maintained institutionally (such as treatises, law books, and journals available through the law library) or they may be part of a student's personal collection of books and files.

The scope of these resources creates a rhythm of paper and electronic research that is seamless to the students. Though the alternation between print and electronic forms may seem inefficient, it is a very fast and effective way for the students to pull together the collection of materials that they actually want to read. Thus a treatise, a paper book, may lead to a specific case citation; this case may then be retrieved directly from an online service, printed, read, and marked-up. Students may type in case citations from this printout to pursue them further. Later, they may retype portions of the case to use as quotes in their written briefs or as notes that will contribute to their writing. Alternatively, once an electronic version of a case is located, research may remain electronic – the student may choose to Shepardize the case or follow some embedded links to precedent cases.

The students print not only to read, but also to perform triage [13] to sort through the cases themselves, or through long lists of potentially interesting cases that they have generated by performing a search or by Shepardizing a case. In fact, several of the students saved Shepards or Key Cite lists with their case printouts, and some of these lists were annotated as part of their triage.

From legal research to design

We observed four trends in the students' legal research: The continued importance and authority of books; research strategies that are link-based rather than search based; the advantage of electronic resources for case evaluation; and alternating use of print and electronic resources.

These trends suggest a set of design consequences for a legal e-book. First, there is a need to support hypertext links. Much legal research involves pursuing explicit citations. Furthermore, citations accrete influence; citations that are seen by the researcher many times are likely to be pursued. Second, we must consider the role of paper in the use of such a device; paper resources and paper practices will persist even given the availability of e-books and electronic services. Finally, there is significant potential value to "waving a wand over a case citation," quickly consulting reverse citations to check the validity of the case being read. This last design consequence highlights the importance of good metadata, and the associated benefits of making the metadata readily available to guide on-the-spot research decisions (for example, "Should I follow this link?" or "Is this case still good law?").

These trends also suggest that digital libraries co-exist with more traditional resources, and must accommodate work in this hybrid environment. Ignoring this reality in the design of interfaces to digital libraries may reduce their usefulness in the real world.

3.3 Reading and annotation

To frame a discussion of reading and annotation, it is important to examine first the form of the materials; the form necessarily shapes and constrains any subsequent activities like the ability to mark on documents or carry them around.

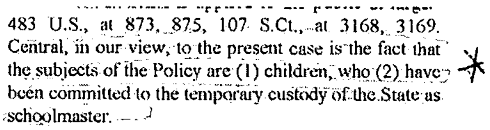
For the most part, the participants in our study read printed documents, with the notable exception of cases they retrieved and skimmed while they were writing. Working with documents on paper allowed the students to read opportunistically, carrying materials around and finding places to read that were relatively free from distractions. We noted that the students and faculty members often printed more than they read.

Do readers then read once, and move on? Not necessarily. Legal practice demands re-reading. A first read may be a scan or a quick skim to see if the material is even relevant, or to get a general idea of what is covered. Subsequent readings may be careful: students reported reading documents they were using in Moot Court front-to-back. Or they may involve skipping to the relevant sections of the document: students reported skipping a case's dissent, or using the headnotes (human-authored indices to the specific points of law covered in a case) to navigate into the body of a case. Re-reading during writing may be very quick, just to remind the student of what is in the materials, or to find a particular passage of interest.

When readers read for a specific activity like writing a legal brief, they are likely to mark on the documents they are reading. Annotation is a prevailing practice, although some readers annotated far more than others and one did not annotate at all.

Re-annotation is also common, concomitant with the kind of re-reading we describe above. If a student is apt to make long, extensive annotations on the first round, he or she may cull them during subsequent readings, either by marking them again, or by using emphasis marks like asterisks to set them off from the original markings (e.g. Figure 2). For example, one student said:

> "You're supposed to use the highlighting to tell you to go back and read it. But sometimes I highlight as I read, and so I have to go back and mark things so that I remember to definitely go back to that. So that's two iterations I guess."



483 U.S., at 873, 875, 107 S.Ct., at 3168, 3169. Central, in our view, to the present case is the fact that the subjects of the Policy are (1) children, who (2) have been committed to the temporary custody of the State as schoolmaster.

Figure 2. A reader's asterisk. The reader plans to revisit material associated with this mark.

It is notable that the students can articulate their own marking strategies. Many annotators are unaccustomed (and sometimes unable) to explaining their annotation practices [12]. However the law students had reflected on their own annotation practices and those of their peers. For example, one student said of her own marking strategy:

> "Usually with the cases, I try to write 'facts,' 'issue...' I'll write 'issue' next to the issue. It replaces briefing. Book briefing [an outlining technique] is just kind of just writing the issue, then you've got your facts and your holding...Some people do the holding in blue and they'll do the issue in pink. I don't do that."

This reflection helps demonstrate the importance of the practice to many of the students. As we have seen in other settings, the students' annotation strategies may vary in ways that are related to the form of the materials. For example, books sometimes receive different treatment than printouts because they are regarded as long term references.

Annotations may also reflect disciplinary practices [11]. As we demonstrated in the quote above, for the law students annotating carries over from the case analysis techniques they learn in class. These techniques give students a way of looking at legal decisions in terms of, for example, issues (what points of law are addressed) and holdings (what the decision's import is). Students sometimes use annotations to identify such aspects of a case, and will even revise them as they continue to read the materials. Such structured interpretation leads to a greater use of annotation tactics like color-coding than may be found in other disciplines.

In spite of the well-developed disciplinary marking strategies that the students exhibit and are able to discuss, these strategies are neither fixed over time nor consistent. They change throughout and beyond schooling as the reader becomes more efficient and comfortable with legal work and unnecessary or unworkable complexities (multi-color coding schemes) are discarded. Sometimes the exigencies of the situation dictate a change in strategy (for example, a favorite pen is left at home). Finally, annotations are crucially tied to situational factors. For example, a lawyer reported that she annotates the same case differently for different uses. The students confirm this. Annotations they make in class that capture what the professor is saying are considered more important than (and are readily distinguishable from) marks the students have made in their own readings of the material. As Wolfe pointed out in her study of how students value the annotations of experts, the source of the interpretation is very important [18].

From reading and annotation to design

Reading is a difficult activity to support; it is hard to improve on paper and pen. We began our study with the assumption that freeform ink annotations are important to analytic readers. This assumption held as we worked with law students. What are further design implications of these reading and annotation practices?

First, re-reading seems like a good target for computational support; readers are already inventing strategies to help themselves read. They skip, scan, or skim through the documents using their own marks or the properties of the documents themselves.

Furthermore, annotations vary in importance and usefulness; the marks readers make on documents have different functions and different degrees of value. Yet annotations are a fundamental technique for signaling what is important in a document. They help readers re-read the material (focusing on the most pertinent and useful portions of a longer document), and they guide readers' future use of source materials in associated activities like writing. Techniques may be applied to find particular kinds of annotations (e.g. see Figure 2) or to use collective marks across different annotations [12].

3.4 Organizing

The documents that students gathered to use in the Moot Court competition are organized in different ways, particular to how they will be used. When research begins, documents may be organized by the court that heard the case, by the date of the case, or simply in a stack.

Once the students began writing their briefs, they tended to move more toward a writing-based organization, creating categories like "pro" cases (cases that support the student's side of the argument), "con" cases (cases that present counterarguments), and cases with matching facts. Cases with a close fact match are particularly interesting, in that they must in some way be addressed. Upon encountering a "con" case very similar to the Moot Court problem, one student wrote, "Deal with this!" on top of the document. These writing-based organizations were frequently implemented as piles on the floor and desktop: fluid organizations that could be rapidly accessed and changed. Naturally if the student worked in multiple venues, this organization could not be preserved across sessions; the student needed recreate it each time, in each place.

One student described the shift from research to writing this way:

> "I find for me I like to print them out. It's sort of like the old way of using index cards. If I print them out, I can staple it and then I can throw them in different piles, and then the piles can change. And then when I'm writing, I look at the pile. I bring it up to here, and then I start writing based on those cases. And that I know in my mind: this is argument 1; this is argument 2; this is argument 3."

This tendency toward activity-based organizational strategies suggests that there is no canonical way of organizing materials. Documents are organized and re-organized to meet the needs of the task at hand and to reflect the student's understanding. It was difficult to ascertain whether the collection's structure would become more uniform after the task ended, as the students did not keep Moot Court documents that they felt they could re-retrieve. Most acknowledged this would change when they became practitioners, and indeed subsequent conversations with practicing attorneys revealed that files within a firm or office may have a standard structure to facilitate sharing.

From organization to design

The design implications of these organizational strategies are threefold: First, it would be advantageous to provide readers with a way of organizing materials across sessions. Many work settings demand that loosely organized documents (e.g., piles on the floor) be picked up and put away, disrupting the organization. Second, as Mander et al. also observed, the difference between organizing materials for research and writing, as well as the difference between transient and archival structures, points to need for multiple ways to organize documents [10]. Finally, the notebooks offer evidence that there is a perceived advantage to keeping documents "in one place." Moving from physical systems of organizing documents to electronic tools may be perceived as a disadvantage however, since computer screens offer far less space and flexibility for spreading out and manipulating working papers. On the other hand, the capability to switch among multiple ways of organizing information on the computer should accommodate not only different working situations, but different cognitive styles as well.

3.5 Writing

The written brief is a key element of the Moot Court competition and indeed of certain kinds of legal work. One prevailing strategy for brief-writing was to outline the important parts of the argument and find the right quotation—a passage or key statement of a rule of law that has come out of a precedent—to illustrate or support the argument. This strategy entails finding the quotation, either from the student's annotations, or simply from memory; transcribing the relevant quote; correctly citing the case, conformant with the prescribed citation form. One student said:

> "I looked at the cases, and looked at the different modes of analysis that the opinions used and there seemed to be two types of tests. One, the Lee v. Weisman analysis, and the other one is the Lemon analysis, which is a three step test, a three prong test. So I kind of used that to structure my outline, and then tried to plug in cases and quotes that I could use for each part of the test."

The students often reported that they would like to perform a word search to re-locate the quotes that they have already read (and potentially annotated) when they are writing. One student gave the following account:

> "after having read through all these various cases, I remember a citation to Brown vs. Board of Education, in which they used a quotation about education being the most important function of government. I couldn't off the top of my head remember which case it was in, or where in that particular case it was located. So essentially what that involved is me going through every single case looking through all my annotations to find this one quotation that I remember having read."

The mechanics of legal writing also involves creating the citation form that conforms to legal practice; each authority that is cited must be in the "blue book" form; "that bringer of much grief, the Uniform System of Citation." [2]

Additional research is also provoked by writing: the students find they need to fill in holes, or to check the authority of a particular case. This interleaved research is different from the student's initial reading. When research is spurred by writing, the student may check the new materials quickly, without printing them. One student said:

> "In the course of writing my brief or paper or whatnot, if there's something I need to look up, a citation or a particular case that's referenced, and I don't want to go to the library to do it or print it out, then I'll do it here, and just kind of flip back and forth between my word processor and Westlaw. Just to transcribe what I need."

As we noted earlier, organizing materials to support the writing process is different than organizing them during the earlier phases of research. The students cite the importance of having several key cases to hand, and arraying them on the floor or tabletop in piles for ready access as they are writing their briefs.

From writing to design

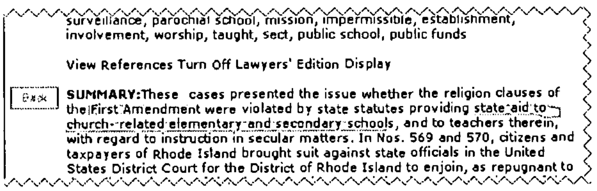
What is the key design insight we can take away from the law students and their writing? What seems clear is that they switch back and forth between activities; these switches may be frequent and fluid. They look for new sources to fill in holes in their arguments. They search for quotes in familiar materials as well. Our initial concept of a dedicated reading appliance may not be the best way of addressing this fluid shift in activities. In other words, a *document laptop* may be more appropriate than a dedicated reader. We discuss this notion further in our accounts of the students' reactions to the XLibris prototype.

4 XLIBRIS REDESIGN

Our interviews and observations of the law students suggested several directions of redesign for XLibris. We redesigned the XLibris interface in terms of functionality and also in terms of appearance; the latter flowed from the need to convey the former. Broadly-speaking, our redesign focused on navigation, link-following, and re-visiting previously-read documents; on retrieval, on annotation, and on managing, organizing and categorizing documents.

4.1 Navigation

Our observations of the students' link following and navigation among gathered documents caused us to re-examine navigation controls. The initial design was based loosely on the Web model of navigation: a "previous" button and a "next" button moved the reader back and forward through the document views. Unlike the Web model, backtracking did not prune branches, but kept all visited documents in the same queue. The intent was to compensate for known deficiencies of the Web browsing model (see [3] for a discussion of alternatives), but the result was equally confusing. The crux of the problem was that the same controls were used for different purposes—short-term exploratory backtracking ("Where does this link go? Oh, no, that wasn't it.") and managing or reading several documents simultaneously. We saw both kinds of activity in our observations: quick skimming of documents (perhaps the destination of a citation link), and working with multiple documents.



surveillance, parochial school, mission, impermissible, establishment, involvement, worship, taught, sect, public school, public funds

View References Turn Off Lawyers' Edition Display

SUMMARY:These cases presented the issue whether the religion clauses of the First Amendment were violated by state statutes providing state aid to church-related elementary and secondary schools, and to teachers therein, with regard to instruction in secular matters. In Nos. 569 and 570, citizens and taxpayers of Rhode Island brought suit against state officials in the United States District Court for the District of Rhode Island to enjoin, as repugnant to

Figure 3. A fragment of a page: a link target is identified by the placement of the "Back" button

In the redesign, we split the two: backtracking from a link traversal was accomplished by a dynamically-added "Back" button, positioned near the target of the link (Figure 3); multi-document use was accommodated by providing a overlaid semi-transparent menu of recent views, from which the reader could select at random (Figure 4). Each item in the menu corresponds to a different document or a specific view (e.g., workspace, clippings, etc.). Changes to the views and documents are reflected in these thumbnail representations. Thus readers can switch easily among the recently-used views and documents without resorting to more elaborate navigation. Additional evaluation is required to know whether this menu should contain only document views, or other organizational views (see below) as well.

4.2 Retrieval

The original XLibris design included a way of launching queries based on freeform digital ink annotations of the underlying text [6]. Because the students tended to follow citations and use Shepards rather than running queries to find useful documents; search was more typically used to find documents that had already been read. Thus we replaced the experimental feature with a keyword search dialog that worked over the documents already loaded into XLibris. A simple ranking algorithm that preferred passages with many different keywords over those with just a few was used to order matching passages.

To handle the kind of reference-following the students preferred, we added the ability to traverse standard http references to Web-based materials. This capacity to load documents into XLibris incrementally changed the flavor interaction from a purely reading-oriented to a hybrid of reading and browsing. Again the effect was to reduce the cost, cognitive and temporal, of transitions between the various activities that make up document-centered research.

4.3 Annotation

The original XLibris "Clippings" view showed the reader only the annotated portions of documents, thereby allowing her to revisit those passages easily. Furthermore, the list could be filtered by color. While this design supported some forms of re-retrieval, it could not accommodate some of the students' practice.

We saw many examples of reviewing and re-annotating previously-read and marked passages (see Figure 2 and the accompanying discussion). A student would read a document and mark it up; subsequently she would review the annotations and identify the more important ones with additional marks. In XLibris, this required many steps: to place a second mark on a passage shown in the Clippings view, a reader had to navigate to the page containing that passage, mark on it, and then move back to the Clippings view, an awkward and distracting operation. We redesigned this interface by allowing readers to mark on the clippings directly, without moving to the containing page; a separate button was provided for moving to the document. The Clippings view and the document view were coordinated: marks made on one view were available in the other. Thus clippings became miniature windows onto parts of documents of particular interest to the reader.

Another shortcoming we identified in the Clippings view was its automatic nature: the view would update when a new mark was added or an old one erased. This made it difficult to collect ideas and references in a persistent manner. We therefore added a new view that was modeled after a yellow legal note pad. Readers could clip passages from the Clippings view to the notebook; once there, they could position and resize the views as desired. Passages pasted into the notebook behaved similarly to the Clippings view: marks made on them affected the document, and vice versa.

This interface (Figure 5) was designed to accommodate the transitions from retrieval to organization to writing: passages found while reading could be clipped and organized thematically in the notebook view, and then could be copied through the system clipboard to the word processing program of choice. Both the text and the image could be pasted in; the canonical "blue book" citation for the source would be included automatically.

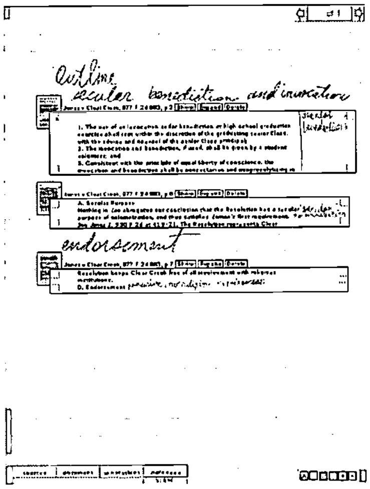


Figure 5. A notebook page with three clipped passages

4.4 Organization

We observed various approaches to organizing documents, reflecting differences in cognitive style and activity. The original design included an overview of document thumbnails which the user could drag to form rudimentary piles. If more documents were added to the workspace, a new page of thumbnails was added to the overview. The reader could not add new pages manually, and it was not possible to group documents thematically by moving them to different pages, similar to student notebooks.

Our redesign accommodated these observations: we provided a way to add blank pages to the workspace and to drag documents between those pages. This allowed readers to organize documents thematically: one worksheet for pro cases, another for con, for example. This ability to move objects between worksheets also applied to the notebook view: the reader could now move relevant clippings to appropriate pages, further increasing XLibris's capability to organize information.

We also allowed readers to mark directly on the workspace pages to label them. Finally, we added a sortable metadata list view that allows readers to group documents by such aspects as court and date, a feature useful in the initial triage stage. Of course other metadata could be used for other kinds of documents.

5 REACTIONS TO THE XLIBRIS PROTOTYPE

At the close of each semi-structured interview we demonstrated XLibris to gauge general reactions to reading on an e-book, and to better match the design to legal research. Early demonstrations were of the original system; later demonstrations showed the redesigned XLibris. The demonstrations were hands-on; the students and faculty members held the device, marking on documents, turning pages, and trying out other features. Because XLibris had Moot Court documents, the students were able to see familiar content in the system. They readily engaged with the prototype, reading, turning pages and marking. We explained why we were showing them the prototype, and elicited as much discussion from them about it as possible.

Reactions were generally positive: students would start reading and annotating right away, using the paper document metaphor in the way we expected. They confirmed the desirability of a mobile device. They did, however, have some important questions about the relationship between XLibris (and e-books) and the technologies they currently use for reading and research.

First, they wanted to understand the relationship between XLibris, paper, and books. Many of them asked if they could print the documents displayed on XLibris. They were not always sure why they would want to print them, but they were certain that this capability was necessary. On the flip side, many of them told us that they still used "the books" for some of their research, and expressed skepticism about the ultimate utility of XLibris in book-oriented research. After demonstrating to us how difficult he found online statutes to use, one student explained that:

> "I have a difficult time one just coming up with [keywords] ... Then when you do get something, it's not what you're looking for. I think it's because ... there's a geography to statutes that you don't have with cases ... that lend themselves to having hardcopy, simply because of the way it's broken down into various titles, and you can easily go to where you need to go."

Second, they asked about how XLibris would interact with their PCs. The questions centered on how other activities would interleave with reading on the device. How would the results of their online research get onto the device? Would they be able to cut and paste quotes from the documents in XLibris to the one they were writing in Microsoft Word? In short, they were acutely aware of the overhead an e-book might add to their current work.

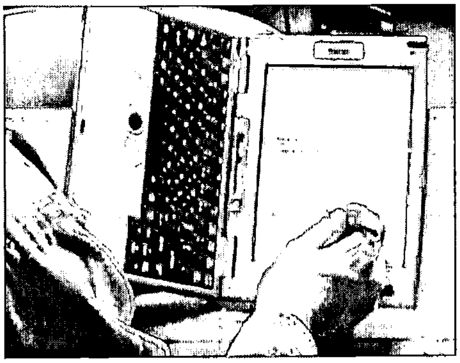
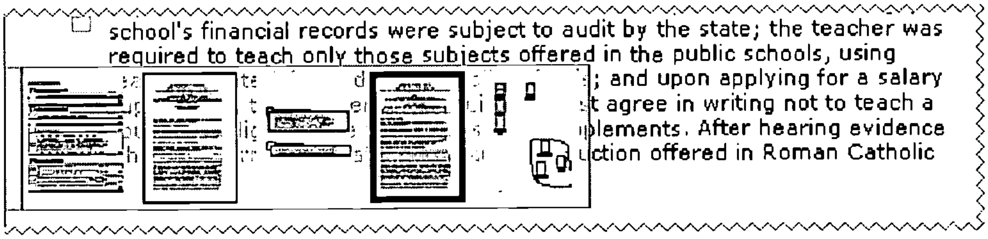


Figure 6. Example configuration of a document laptop, a Fujitsu Lifebook B-Series pen mini-notebook, rotated to display a document in portrait mode.

Finally, they wondered about the relationship between XLibris and their laptops. Many of them had already expressed frustration with their laptops, complaining about their weight, bulk, and durability. Now we were introducing a second computer-like device. One student confided that she would not want to carry both. Others asked if they could perform normal computer work on the pen tablet computer, for example, getting their email, or using a word processor with a keyboard.

The implications of these questions are far reaching. Again, the concept of a document laptop (Figure 6) seems to win out over that of a large dedicated e-book. Reading is interleaved with other activities. This observation leads us to emphasize the ability to read more effectively on existing hardware, and take advantage of its form factor. Transitions into and out of XLibris—e.g., pasting quotes from sources into a word processor—are easier if the reading and writing applications are on the same computer. Recent trends toward lighter hardware and wireless peripherals



school's financial records were subject to audit by the state; the teacher was required to teach only those subjects offered in the public schools, using

; and upon applying for a salary
t agree in writing not to teach a
plements. After hearing evidence
ction offered in Roman Catholic

Figure 4. Semi-transparent overlay showing several recent views (1,3,5) and documents (2,4)

may make it easier to combine the advantages of the tablet and laptop form factors in the same device.

6 CONCLUSION

Surely the shift from e-book to document laptop represents the greatest sea-change in our thinking about legal work. When we began this study, we assumed we would be introducing a dedicated reading device. Now we believe that the advantages afforded by such a device are offset by the need to interleave other activities with reading.

We are less likely to think of reading devices as peripherals tethered to a stationary PC. Reading is so opportunistic, and paper is such a flexible medium, that it seems inappropriate to tie legal work to a place and time. Wireless access to materials may be just the "in" that makes a document laptop a useful and desirable piece of technology.

Our observations of legal research left us with two important insights. First, we cannot underestimate the importance of the notion of a starting place, one that might easily be a paper treatise. Second, we saw that link following is at least as important as the ability to perform broad queries. A document representation that includes links and functionality that implements link traversal now seems essential.

Through our observations, we came away with three compelling scenarios for using a document laptop to perform legal research: (1) immediate access to current legal materials; (2) the ability to re-retrieve familiar materials; and (3) the ability to suspend and resume interrupted work that involves many documents.

Reading and annotation were the original terms of engagement for XLibris. It would seem like there is little more to be said about these two areas. Yet we have seen new styles of working with annotations (e.g. outlining styles that use short points amplified by extracted quotations), and are investigating different reading phenomena (e.g. re-reading).

We also found opportunities to revisit issues of navigation within and among documents, and to explore additional ways of managing and organizing documents and passages.

In short, the field study produced exciting insights and possibilities for e-books. It also introduced new questions and issues about how a document appliance-turned-laptop will function in legal work, and provided additional evidence for the hybrid nature of document collections. Future designs of information appliances, e-books, and other interfaces to digital libraries must consider the simultaneous use of paper and digital documents and the fluid transitions between them.

7 ACKNOWLEDGMENTS

We thank Elizabeth Churchill for her help with some of the field work, Peter Hodgson for his graphic design and Photoshop wizardry, and Kei Tanaka for debugging. We thank Jim Baker for supporting this research.

8 REFERENCES

1. Adler, A., Gujar, A., Harrison, B., O'Hara K. and Sellen, A. A diary study of work-related reading: design implications for digital reading devices, in *Proceedings of CHI98* (Los Angeles, CA, April 1998), ACM Press, 241-248.
2. Berring, R.C. *Legal Information and the Search for Cognitive Authority*. UC Berkeley School of Law, Public Law and Legal Theory Working paper No. 99-1, September 1999. Available at http://papers.ssrn.com/paper.taf?abstract_id=184050
3. Bieber, M. and Wan, J. Backtracking in a multiple-window hypertext environment, in *Proceedings of ECHT '94* (Edinburgh, UK, September 1994), ACM Press, 158-166.
4. Blomberg, J., Suchman, L., & Trigg, R. Reflections on a Work-Oriented Design Project. *Human-Computer Interaction, 11*,3 (1996), 237-265.
5. Elliott, M. *Digital Library Design for Organizational Usability in the Courts*. Available on the web at http://edfu.lis.uiuc.edu/allerton/95/s3/elliott.html. Retrieved on Jan 9, 2001.
6. Golovchinsky, G., Price, M.N., and Schilit, B.N. From reading to retrieval: freeform ink annotations as queries, in *Proceedings of SIGIR '99*.(Berkeley, CA, August 1999), ACM Press, 19-25.
7. Hill, W.C. and Hollan, J.D. Edit Wear and Read Wear, in *Proceedings of CHI 92* (Monterey, CA, April 1992), ACM Press, 3-9.
8. Jones, M., Rieger, R., Treadwell, P. and Gay, G. Live from the stacks: user feedback on mobile computers and wireless tools for library patrons. *Proceedings of ACM Digital Libraries 2000* (San Antonio, TX, June 2000), ACM Press, 95-102.
9. Kay, A. and Goldberg, A. Personal Dynamic Media, *Computer 10*,3 (March 1977), 31-41.
10. Mander, R., Salomon, G., and Wong, Y.Y. A pile metaphor for supporting casual organization of information, in *Proceedings of CHI*92 (Monterey, CA, April 1992), ACM Press, 627-634.
11. Marshall, C. Toward an ecology of hypertext annotation, in *Proceedings of ACM Hypertext '98*, (Pittsburgh, PA, June 1998) ACM Press, 40-49.
12. Marshall, C.C., Price, M.N., Golovchinsky, G., and Schilit, B.N. Introducing a digital library reading appliance into a reading group, in *Proceedings of ACM Digital Libraries 99* (Berkeley, CA, August 1999) ACM Press, 77-84.
13. Marshall, C.C. and Shipman, F.M. III Spatial hypertext and the practice of information triage, in *Proceedings of Hypertext 97* (Southampton, UK, April 1997), ACM Press, 124-133.
14. Poon, A., Weber, K., Cass, T. Scribbler: A Tool for Searching Digital Ink. In *CHI95 Conference Companion* (Denver CO, May 1995), ACM Press. pp. 252-253.
15. Schilit, B.N., Golovchinsky, G., and Price, M.N. Beyond paper: supporting active reading with free form digital ink annotations, in *Proceedings of CHI98* (Los Angeles, CA, April 1998), ACM Press, 249-256.
16. Schilit, B.N., Price, M.N., Golovchinsky, G., Tanaka, K., and Marshall, C.C. As we may read: the reading appliance revolution. *IEEE Computer 32*,1 (January 1999), 65-73.
17. Sutton, S.A. The role of attorney mental models of law in case relevance determinations: an exploratory analysis. *JASIS* 45,3 (1994) 186-200.
18. Wolfe, J. Effects of annotations on student readers and writers. *Proceedings of ACM Digital Libraries 2000* (San Antonio, TX, June 2000) ACM Press, 19-26.

Mapping the Interoperability Landscape for Networked Information Retrieval

William E. Moen
School of Library and Information Sciences, Texas Center for Digital Knowledge,
University of North Texas, P.O. Box 311068, Denton, Texas 76203
940-565-3563
wemoen@unt.edu

ABSTRACT

Interoperability is a fundamental challenge for networked information discovery and retrieval. Often treated monolithically in the literature, interoperability is multifaceted and can be analyzed into different types and levels. This paper discusses an approach to map the interoperability landscape for networked information retrieval as part of an interoperability assessment research project.

Categories and Subject Descriptors

H.3.7 **[Information Storage and Retrieval]**: Digital Libraries – *standards, systems issues, user issues.*

General Terms

Standardization

Keywords

Interoperability; networked information discovery and retrieval; Z39.50, testbeds.

1. INTRODUCTION

The ability of two information systems to communicate, execute instructions, share data, or otherwise interact is a fundamental requirement in the networked environment. Typically these interactions are subsumed under the term *interoperability*. In the traditional library automation environment and more recently in the digital library context, we have recognized the complexity of linking information retrieval systems. Yet, there are increased expectations for seamless, transparent, and reliable access and sharing of information in and between digital libraries. Increasingly the literature identifies interoperability as one of the fundamental problem facing networked information discovery and retrieval (NIDR) [3].

This paper outlines a preliminary and evolving framework for addressing interoperability. Mapping the NIDR interoperability landscape is part of a broader research and demonstration project for assessing interoperability. The U.S. federal Institute of Museum and Library Services awarded a National Leadership Grant to the School of Library and Information Sciences and Texas Center for Digital Knowledge at the University of North Texas to establish a research and demonstration Z39.50 interoperability testbed [5].

A map or conceptual framework of interoperability allows us to situate our focal activities since not all types of interoperability will be assessed through the initial testbed. By identifying the multiple factors that threaten interoperability, we can control some of those factors in the testbed while acknowledging that subsequent phases of the research can address other factors. For example, the initial Z39.50 interoperability testbed focuses on three types or levels of interoperability: protocol syntax level, protocol service level, and semantic level. Each of these levels, and particularly the semantic level, is multifaceted. Lynch and Garcia-Molina describe semantic interoperability as a "grand challenge" research problem [3]. The testbed focuses on some aspects of semantic interoperability.

Similar to Paepcke, et al. [1], we suggest that multiple levels and categories of interoperability can be identified, and these may differ depending on the application. Too often the literature treats interoperability monolithically or simply from a system perspective (i.e., the level of two systems interacting). Mapping the interoperability landscape can help focus attention on specific interoperability problems and assessment methodologies. We suggest further that the degree of interoperability between information systems may be dependent on the distance between communities whose information systems attempt to interact. Further, it is ultimately the user who benefits when systems interoperate, and we propose that user assessments of interoperability should be factored into the methodology.

2. DEFINITIONS

In the context of the networked environment, a number of definitions of interoperability surface [2, 6]. Miller goes beyond a system orientation and describes a more expansive perspective on interoperability. He suggests multiple aspects of the concept including Technical, Semantic, Inter-community, and Legal [4].

Within the context of digital libraries and networked information retrieval, we assume the following operating principle: *systems will interoperate*. Miller's approach, however, underscores the principle that not only systems but also organizations will need to interoperate. This brings attention to various environmental factors that can affect interoperability.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

3. INTEROPERABILITY FACTORS

A number of diverse factors challenge interoperability:

- Multiple and disparate operating and IR systems
- Multiple protocols
- Multiple metadata schemes
- Multiple data formats
- Multiple languages and character sets
- Multiple vocabularies, ontologies, and disciplines.

Our map of interoperability will account for a variety of factors.

4. COMMUNITY AND DOMAIN CONTEXTS FOR INTEROPERABILITY

The context of "information communities" provides a way to frame the challenges of achieving interoperability. The diversity of factors above may be reduced within a particular community.

For cross-catalog information retrieval in traditional libraries, the diversity listed above is radically reduced. Data in the catalogs are relatively homogenous, and there is commonality in the metadata scheme for structuring the records, etc. The challenges to achieve interoperability in a virtual catalog application may be less than in cases when one crosses community boundaries. For example, if a user queries repositories of library bibliographic records and a museum's object records concurrently through Z39.50, diversity of metadata schemes and vocabulary increases.

Our preliminary map proposes the following NIDR communities:

- Focal–community NIDR: factors affecting interoperability are minimized (e.g., libraries)
- Extended–community NIDR: increased diversity (e.g., cultural heritage information held by libraries and museum)
- Extra–community NIDR: factors affecting interoperability are maximized (e.g., libraries interacting with geospatial repositories).

Challenges to interoperability NIDR increase (and likely the costs of achieving interoperability increase) as one moves further outside of a focal community. Figure 1 illustrates potential interaction among communities.

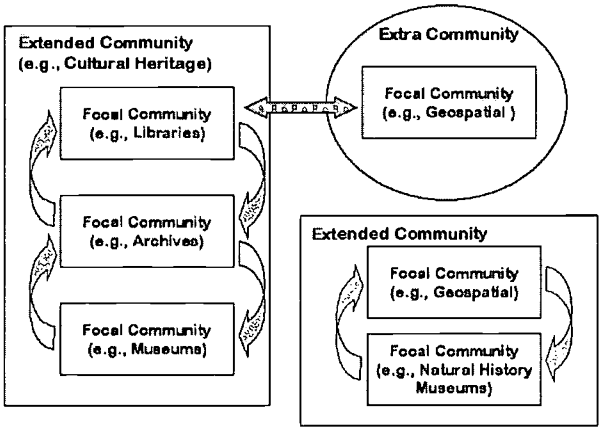


Figure 1. Information Retrieval Among/Across Communities

Another useful perspective for the interoperability conceptual map is in terms of domains. There may be some overlap here with the perspective of communities, but addressing differences in domains can highlight factors such as vocabularies and ontologies. Our mapping identifies the following two categories:

- Intra-domain/discipline
- Extra-domain/discipline.

As in the case of communities, the challenges to interoperability within a domain may be less than between domains. Semantic differences may present major barriers to interoperability.

Within a community or domain, relative homogeneity reduces interoperability challenges. Heterogeneity increases as one moves outside of a focal community/domain, and interoperability is likely more costly and more difficult to achieve.

5. SUMMARY

The work on mapping the interoperability landscape is in its initial phase. The resulting map situates our interoperability assessment research in that landscape and provides points of reference to other researchers for assessing and overcoming interoperability problems. Ultimately, this map may assist in realistically assessing the challenges and costs in making real the promises of providing seamless and transparent access (i.e., interoperability) within the networked environment.

6. ACKNOWLEDGMENTS

We thank the U.S. federal Institute of Museum and Library Services for support of the interoperability testbed and papers resulting from our research.

7. REFERENCES

[1] Paepcke, A., et al. Interoperability for digital libraries worldwide. Communications of the ACM 41(April 1998), 33-43.

[2] Lynch, C. Interoperability: The standards challenge for the 1990s. Wilson Library Bulletin, 67 (March 1993), 38-42.

[3] Lynch, C. & Garcia-Molina, H. Interoperability, scaling, and the digital libraries research agenda: A report on the May 18-19, 1995 IITA digital libraries workshop. (1995). Available URL: http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/

[4] Miller, P. Interoperability: What it is and why should I want it. Ariadne, 24. Available URL: http://www.ariadne.ac.uk/issue24/interoperability/intro.html

[5] Moen, W. E. Realizing the vision of networked access to library resources: An applied research and demonstration project to establish and operate a Z39.50 interoperability testbed. (February 2000). Available URL: http://www.unt.edu/zinterop

[6] Preston, C.M. & Lynch, C.A. Interoperability and conformance issues in the development and implementation of the government information locator service (GILS). In W.E. Moen and C.R. McClure, The Government Information Locator Service (GILS). Syracuse, NY: School of Information Studies, Syracuse University (1994).

Distributed Resource Discovery: Using Z39.50 to Build Cross-Domain Information Servers

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley
(510)642-6046
ray@sherlock.berkeley.edu

ABSTRACT

This short paper describes the construction and application of Cross-Domain Information Servers using features of the standard Z39.50 information retrieval protocol[11]. We use the Z39.50 Explain Database to determine the databases and indexes of a given server, then use the SCAN facility to extract the contents of the indexes. This information is used to build "collection documents" that can be retrieved using probabilistic retrieval algorithms.

Keywords

Distributed Information Retrieval, Cross-Domain Resource Discovery, Distributed Search

1. INTRODUCTION

Information seekers must be able to identify information resources that are pertinent to their needs. Today they are also required to have the knowledge and skills to navigate those resources, once identified, and extract relevant information. The widespread distribution of recorded knowledge across the network landscape of the World Wide Web is only the beginning of the problem. The reality is that the repositories of recorded knowledge on the Web are only a small part of an environment with a bewildering variety of search engines, metadata, and protocols of very different kinds and of varying degrees of completeness and incompatibility. The challenge is to not only to decide how to mix, match, and combine one or more search engines with one or more knowledge repositories for any given inquiry, but also to have detailed understanding of the endless complexities of largely incompatible metadata, transfer protocols, and so on.

As Buckland and Plaunt[1] have pointed out, searching for recorded knowledge in a digital library environment involves three types of selection:

1. Selecting which library (repository) to look in;

2. Selecting which document(s) within a library to look at; and

3. Selecting fragments of data (text, numeric data, images) from within a document.

In the following discussion we focus on the first type of selection, that is, discovering which digital libraries are the best places for the user to begin a search. Our approach to the second selection problem has been discussed elsewhere[6,7].

Distributed information retrieval has been an area of active research interest for many years. Distributed IR presents three central research problems that echo the selection problems noted by Buckland and Plaunt. These are:

1. How to select appropriate databases or collections for search from a large number of distributed databases;

2. How to perform parallel or sequential distributed search over the selected databases, possibly using different query structures or search formulations, in a networked environment where not all resources are always available; and

3. How to merge results from the different search engines and collections, with differing record contents and structures (sometimes referred to as the collection fusion problem).

Each of these research problems presents a number of challenges that must be addressed to provide effective and efficient solutions to the overall problem of distributed information retrieval.

These problems been approached in a variety of ways by different researchers focusing on different aspects of retrieval effectiveness, and on construction of the index resources required to select and search distributed collections [2,3,4,5,8,9].

In this paper we present a method for building resource discovery indexes that is based on the Z39.50 protocol standard instead of requiring adoption of a new protocol. It does not require that remote servers perform any functions other than those already established in the Z39.50 standard.

2. Z39.50 FACILITIES

The Z39.50 Information retrieval protocol is made up of a number of "facilities", such as Initialization, Search, and Retrieval. Each facility is a logical group of services (or a single service) to perform various functions in the interaction between a client (origin) and a server (target). The facilities that we will be concerned with in this paper are the Explain Facility and the Browse Facility.

2.1 The Explain Facility

The Z39.50 Explain facility permits the client to obtain information about the server implementation, including information on databases supported, attribute sets used (an attribute set specifies the allowable search fields and semantics for a database), diagnostic or error information, record syntaxes and information on defined subsets of record elements that may be

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

requested from the server (called elementsets). The server (optionally) maintains a database of Explain information about itself and may maintain Explain databases for other servers. The explain database appears to the client as any other database, and uses the Z39.50 Search and Retrieval facilities to query and retrieve information from it. There are specific attributes, search terms and record syntaxes defined in the standard for the Explain database to facilitate interoperability among different server implementations.

2.2 Z39.50 Browse Facility

As the name of this facility implies, it was originally intended to support browsing of the server contents, specifically the items extracted for indexing the databases. The single service in the Browse facility is the Scan service. It is used to scan an ordered list of terms (subject headings, titles, keyword, text terms, etc.) drawn from the database. Most implementations of the Scan service directly access the contents of the indexes on the server and return requested portions of those indexes as an ordered list of terms along with the document frequency for each term.

3. IMPLEMENTATION

Our implementation relies on these two Z39.50 Facilities to derive information from Z39.50 servers (including library catalogs, full-text search systems, and digital library systems) in order to build a GlOSS-like[5] index for distributed resources. The procedure followed is:

1. Search the Explain Database to derive the server information about each database maintained by the server and the attributes available for searching that server.
2. For each database, we determine whether Dublin Core attributes are available, and if not, we select from the available attributes those most commonly associated with Dublin Core information.
3. For each of the attributes discovered, we send a sequence of Scan requests to the server and collect the resulting lists of index terms. As the lists are collected they are verified for uniqueness (since a server may allow multiple search attributes to be processed by the same index) so that duplication is avoided.
4. For each database an XML *collection document* is constructed to act as a surrogate for the database using the information obtained from the server Explain database and the Scans of the various indexes.
5. A database of collection documents is created and indexed using all of the terms and frequency information derived above.

Probabilistic ranking methods[6] are used to retrieve and rank these collection documents for presentation to the user for selection, or for automatic distributed search of the most highly ranked databases using method similar to [2] and [9]. Although the Z39.50 protocol has been used previously for resource discovery databases [8], in that work random samples of the records in the collection were used to build the indexes. The method described here gives the ability to use the server's own processing of its records in extracting the terms to be matched in the resource discovery index.

4. CONCLUSION

This brief paper has presented a standards-based method for building a resource discovery database from distributed Z39.50 servers. We plan to combine this system with others that support SDLIP, and possible other protocols for further investigation of distributed search for Digital Libraries.

5. ACKNOWLEDGMENTS

This work was sponsored by NSF and JISC under the International Digital Libraries Program award #IIS-9975164.

6. REFERENCES

[1] Buckland, M. K. and Plaunt, C. Selecting Libraries, Selecting Documents, Selecting Data. In ISDL'97: (Tsukuba City, Japan, 1997). http://www.dl.ulis.ac.jp/ISDL97/proceedings/.

[2] Callan, J. P., Lu, Z. and Croft, W. B. Searching Distributed Collections with Inference Networks. In SIGIR '95: (Seattle, WA, 1995), ACM Press, 21-28.

[3] Danzig, P.B., Ahn, J., Noll, J. and Obraczka, K. Distributed Indexing: A Scalable mechanism for Distributed Information Retrieval. In SIGIR '91 (Chicago, IL, 1991), ACM Press, 220-229.

[4] French, J. C., et al. Evaluating Database Selection Techniques: A Testbed and Experiment. In SIGIR '98 (Melbourne, Australia, 1998), ACM Press, 121-129

[5] Tomasic, A., et al. Data Structures for Efficient Broker Implementation. ACM Transactions on Information Systems, 15 (July 1997), 254-290.

[6] Larson, R. R. and Carson, C. Information Access for A Digital Library: Cheshire II and the Berkeley Environmental Digital Library. In Proceedings ASIS '99 (Washington, DC, 1999), Information Today, 515-535.

[7] Larson, R. R. TREC Interactive with Cheshire II. Information Processing and Management, 37 (2001), 485-505

[8] Lin, Y., et al. Zbroker: A Query Routing Broker for Z39.50 Databases. In: CIKM '99: (Kansas City, MO, 1999), ACM Press, 202-209.

[9] Xu, J. and Callan, J. (1998) Effective Retrieval with Distributed Collections. In SIGIR '98, (Melbourne, Australia, 1998), ACM Press, 112-120.

[10] Z39.50 Maintenance Agency. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995), Washington: Library of Congress, 1995

The Open Archives Initiative: Building a Low-Barrier Interoperability Framework

Carl Lagoze
Digital Library Research Group
Cornell University
Ithaca, NY
+1-607-255-6046
lagoze@cs.cornell.edu

Herbert Van de Sompel
Digital Library Research Group
Cornell University
Ithaca, NY
+1-607-255-3085
herbertv@cs.cornell.edu

ABSTRACT

The Open Archives Initiative (OAI) develops and promotes interoperability solutions that aim to facilitate the efficient dissemination of content. The roots of the OAI lie in the E-Print community. Over the last year its focus has been extended to include all content providers. This paper describes the recent history of the OAI – its origins in promoting E-Prints, the broadening of its focus, the details of its technical standard for metadata harvesting, the applications of this standard, and future plans.

Categories and Subject Descriptors

D.2.12 [**Software Engineering**]: Interoperability – *Interface definition languages.*

General Terms

Experimentation, Standardization.

Keywords

Metadata, Interoperability, Digital Libraries, Protocols.

1. INTRODUCTION

In October 1999, a meeting was held in Santa Fe to discuss mechanisms to encourage the development of E-Print solutions. The group at this meeting was united in the belief that the ubiquitous interconnectivity of the Web provides new opportunities for the timely dissemination of scholarly information. The well-known physics archive run by Paul Ginsparg at Los Alamos National Laboratory has already radically changed the publishing paradigm in its respective field. Similar efforts planned, or already underway, promise to extend these striking changes to other domains.

The result of this meeting was the formation of the Open Archives Initiative (OAI) and beginning of work on a framework facilitating the federation of content providers on the Web. Since that first meeting, the OAI has undergone a period of intensive development both organizationally and technically. The original focus on E-Prints has broadened to encompass content providers from many domains (with an emphasis on what could be classified "scholarly" publishing), a refined and extensively tested technical framework has been developed, and an organizational structure to support the Initiative has been established.

The name *Open Archives Initiative* reflects the origins of the OAI in the E-Prints community where the term *archive* is generally accepted as a synonym for a repository of scholarly papers. Members of the archiving profession have justifiably noted the strict definition of an "archive" within their domain; with implications for preservation of long-term value, statutory authorization and institutional policy. The OAI uses the term "archive" in a broader sense: as a repository for stored information. Language and terms are never unambiguous and uncontroversial and the OAI respectfully requests the indulgence of the archiving community with this less constrained use of "archive".

Some explanation of the use of the term "Open" in OAI is also due. Our intention is "open" from the architectural perspective – defining and promoting machine interfaces that facilitate the availability of content from a variety of providers. Openness does not mean "free" or "unlimited" access to the information repositories that conform to the OAI technical framework. Such terms are often used too casually and ignore the fact that monetary cost is not the only type of restriction on use of information – any advocate of "free" information recognize that it is eminently reasonable to restrict denial of service attacks or defamatory misuse of information.

This paper documents the development of the Open Archives Initiative and describes the plans for the OAI for the near future. At the time of completion of this paper (May 2001), the OAI has released the technical specifications of its metadata harvesting protocol. The substantial interest in the OAI heretofore indicates that the approach advocated by the OAI – establishing a low-entry and well-defined interoperability framework applicable across domains – may be the appropriate catalyst for the federation of a broad cross-section of content providers. The coming year will indicate whether this is true and whether the technical framework defined by the metadata harvesting protocol is a sufficient

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

underpinning for the development of usable digital library services.

2. E-PRINT ORIGINS

The initial meeting and developments of the Open Archives Initiative are described in detail in an earlier paper [1]. This section summarizes that material from the perspective of current developments and events.

The origins of the OAI lie in increasing interest in alternatives to the traditional scholarly publishing paradigm. While there may be disagreements about the nature of what changes need to take place, there is widespread consensus that change, perhaps radical change, is inevitable. There are numerous motivating factors for this change. An increasing number of scholarly disciplines, especially those in the so-called "hard sciences" (e.g., physics, computer science, life sciences), are producing results at an increasingly rapid pace. This velocity of change demands mechanisms for reporting results with lower latency times than the ones experienced in the established journal system. The ubiquity of high-speed networks and personal computing has created further consumer demand for use of the Web for delivery of research results. Finally, the economic model of scholarly publishing has been severely strained by rapidly rising subscription prices and relatively stagnant research library budgets.

In some scholarly fields, the development of alternative models for the communication of scholarly results – many in the form of on-line repositories of EPrints – has demonstrated a viable alternative to traditional journal publication. Perhaps the best known of these is the Physics archive[1] run by Paul Ginsparg [2] at Los Alamos National Laboratory. There are, however, a number of other established efforts (CogPrints[2], NCSTRL[3], RePEC[4]), which collectively demonstrate the growing interest of scholars in using the Internet and the Web as vehicles for immediate dissemination of research findings. Stevan Harnad, among the most outspoken advocate of change, views such solutions as the first step in radical transformation of scholarly publishing whereby authors reclaim control over their intellectual property and the publishing process [3].

The October 1999 meeting in Santa Fe[5] of what was then called the UPS (Universal Preprint Service) was organized on the belief that the interoperability among these E-Print archives was key to increasing their impact. Interoperability would make it possible to bridge across, or federate, a number of archives. Issues related to interoperability are well described elsewhere [4]. It is sufficient to say here that establishing such a framework requires both technical and organizational agreements.

There are many benefits of federation of E-Print repositories. Scholarly endeavors are increasingly multi-disciplinary and scholars should be able to move fluidly among the research results from various disciplines. Federation and interoperability also encourage the construction of innovative services. Such services might use information from various repositories and process that information to link citations, create cross-repository query interfaces, or maintain current-awareness services. The benefits of federation were demonstrated by earlier work joining the Los Alamos archive with the NCSTRL system [5], as well as in the UPS prototype [6] that was prepared for the Santa Fe meeting.

Interoperability has numerous facets including uniform naming, metadata formats, document models, and access protocols. The participants at the Santa Fe meeting decided that a low-barrier solution was critical towards widespread adoption among E-Print providers. The meeting therefore adopted an interoperability solution known as *metadata harvesting*. This solution allows E-Print (content) providers to expose their metadata via an open interface, with the intent that this metadata be used as the basis for value-added service development. More details on metadata harvesting and the OAI technical agreements are provided in Section. 4.

The result of the meeting was a set of technical and organizational agreements known as the *Santa Fe Convention*. The technical aspects included the agreement on a protocol for metadata harvesting based on the broader Dienst protocol [7], a common metadata standard for E-Prints (the Open Archives Metadata Set), and a uniform identifier scheme. The organizational agreements coming out of the meeting were informal and involved the establishment of email lists for communication amongst participants, a rudimentary registration procedure, and the definition of an acceptable use policy for consumers of harvested metadata.

The Santa Fe meeting closed with enthusiasm among the participants to refine the agreements and pursue implementation and experimentation. Within a relatively short period the technical specifications were completed and were posted on a publicly accessible web site[6] along with other results of the Santa Fe meeting. A number of the participants quickly implemented the technical agreements and others experimented with a number of prototype services.

3. BEYOND E-PRINTS

Soon after the dissemination of the Santa Fe Convention in February 2000 it became clear that there was interest beyond the E-Print community. A number of other communities were intrigued by a low-barrier interoperability solution and viewed metadata harvesting as a means to this end.

In particular, strong interest came from the research library community in the US. Key members of this community met at the so-called *Cambridge Meetings*, sponsored by the Digital Library Federation and the Andrew W. Mellon Foundation, at Harvard University in the first half of 2000. The goal of the meetings was

[1] http://www.arxiv.org.

[2] http://cogprints.soton.ac.uk.

[3] http://www.ncstrl.org.

[4] http://netec.mcc.ac.uk/RePEc/.

[5] The Santa Fe meeting was sponsored by the Council on Library and Information Resources (CLIR), the Digital Library Federation (DLF), the Scholarly Publishing & Academic Resources Coalition (SPARC), the Association of Research Libraries (ARL) and the Los Alamos National Laboratory (LANL).

[6] http://www.openarchives.org.

to explore the ways that research libraries could expose aspects of their collections to Web search engines. The participants, who included not only representatives from research libraries but also from the museum community, agreed that exposing metadata in a uniform fashion was a key step towards achieving their goal [8].

Additional evidence of the broad-based interest in Santa Fe Convention came in the form of well-attended Open Archives Initiative workshops held at the ACM Digital Library 2000 Conference in San Antonio [9] and the European Digital Library Conference in Lisbon [10]. Participants at both of these workshops included publishers, librarians, metadata and digital library experts, and scholars interested in E-Prints.

Responding to this wider interest required a reconsideration of a number of decisions made by members of the Open Archives Initiative at the Santa Fe meeting and in the months following.

- The original mission of the OAI was focused on E-Print solutions and interoperability as a means of achieving their global acceptance. While this goal was still shared by the majority of participants, it was deemed too restrictive and possibly alienating to communities not directly engaged or interested in E-Prints.
- A number of aspects of the technical specifications were specific to the original E-Print focused mission and needed to be generalized for applicability to a broad range of communities.
- The credibility of the effort was uncertain due to the lack of organizational infrastructure. Communities such as the research library community are hesitant to adopt so-called "standards" when the stability of the organization responsible for promotion and maintenance of the standard are questionable.

The issue of organizational stability was addressed first. In August 2000, the DLF (Digital Library Federation) and CNI (Coalition of Networked Information) announced organizational support and resources for the ongoing OAI effort. This support announcement was made in a press release[7] that also contained the formation of an OAI steering committee with membership from a cross-section of communities and a level of international participation (membership of the Steering Committee is listed in the Appendix in Section 7).

The OAI steering committee immediately addressed the task of compiling a new mission statement that reflected the broader scope. This mission statement is as follows:

The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. Continued support of this work remains a cornerstone of the Open Archives program. The fundamental technological framework and standards that are developing to support this work are, however, independent of the both the type of content offered and the economic mechanisms surrounding that content, and promise to have much broader relevance in opening up access to a range of digital materials. As a result, the Open Archives Initiative is currently an organization and an effort explicitly in transition, and is committed to exploring and enabling this new and broader range of applications. As we gain greater knowledge of the scope of applicability of the underlying technology and standards being developed, and begin to understand the structure and culture of the various adopter communities, we expect that we will have to make continued evolutionary changes to both the mission and organization of the Open Archives Initiative.

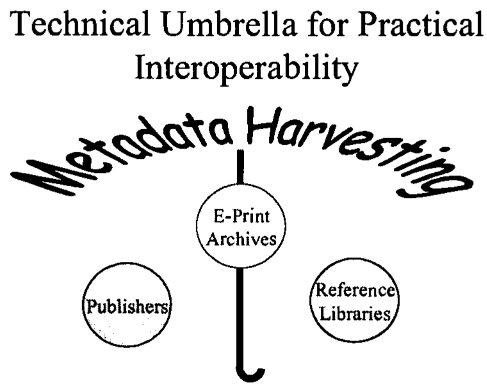


Figure 1 - A framework for multiple communities

This mission is illustrated in Figure 1, where the technical framework is designed as an umbrella that can be exploited by a variety of communities.

A key element of this mission statement is the formulation of the OAI as an experiment: "*an organization and an effort explicitly in transition*". The organization is well aware that the technical infrastructure it proposes – metadata harvesting – has to be proven as an effective means of facilitating interoperability or even what that interoperability will achieve. The organizational structure and strategy reflects a belief among the steering committee that "proving the concept" will require a delicate balance between stability and flexibility. Furthermore, there is strong consensus that goals and scope of the OAI should be controlled – while interoperability is a wide-open area with many potential areas of investigation, the OAI should resist expanding its scope until its current technical goals are met and justified.

The Steering Committee also took steps to address the E-Print focus of the Santa Fe technical agreements and to fix other problems that were revealed in testing of those agreements. A technical committee was formed and a meeting organized at Cornell University in September 2000. The results of that meeting are reported in the next section.

4. TECHNICAL FRAMEWORK

The technical framework of the Open Archives Initiative is intended to provide a low-barrier approach to interoperability. The membership of the OAI recognizes that there are functional limitations to such a low-barrier framework and that other interoperability standards, for example Z39.50, address a number of issues in a more complete manner. However, as noted by Bill Arms [11], interoperability strategies generally increase in cost (difficulty of implementation) with an increase in functionality. The OAI technical framework is not intended to replace other approaches but to provide an easy-to-implement and easy-to-

[7] http://www.openarchives.org/OAISC/oaiscpress000825.htm.

deploy alternative for different constituencies or different purposes than those addressed by existing interoperability solutions. As noted earlier, experimentation will prove whether such low-barrier interoperability is realistic and functional.

At the root of the technical agreement lies a distinction between two classes of participants:

- *Data Providers* adopt the OAI technical framework as a means of exposing metadata about their content.
- *Service Providers* harvest metadata from data providers using the OAI protocol and use the metadata as the basis for value-added services.

The remainder of this section describes the components of this technical framework. A theme carried through the framework and the section is the attempt to define a common denominator for interoperability among multiple communities while providing enough hooks for individual communities to address individual needs (without interfering with interoperability). More details on the technical framework are available in the Open Archives Metadata Protocol specification available at the OAI web site.[8]

4.1 Metadata

The OAI technical framework addresses two well-known metadata requirements: interoperability and extensibility (or community specificity). These issues have been a subject of considerable discussion in the metadata community [12, 13] – the OAI attempts to answer this in a simple and deployable manner.

The requirement for metadata interoperability is addressed by requiring that all OAI data providers supply metadata in a common format – the Dublin Core Metadata Element Set [14]. The decision to mandate a common element set and to use the Dublin Core was the subject of considerable discussion in the OAI. One approach, which has been investigated in the research literature [15], is to place the burden on the consumer of metadata rather than the provider, tolerating export of heterogeneous metadata and relying on services to map amongst the representations. OAI, however, is purposely outside the domain of strict research, and in the interest of easy deployment and usability it was decided that a common metadata format was the prudent decision.

The decision to use the Dublin Core was also the result of some deliberation. The original Santa Fe convention took a different course – defining a metadata set, the Open Archives Metadata Set, with some functionality tailored for the E-Print community. The broadening of the focus of the OAI, however, forced reconsideration of this decision and the alternative of leveraging the well-known and active work of the DC community in formulating a cross-domain set for resource discovery was chosen.

Those familiar with the Dublin Core will note that all fields in DC are *optional*. The OAI discussed requiring a number of DC elements in OAI records. While such requirement might be preferable from the perspective of interoperability, the spirit of experimentation in the OAI persuaded the committee to keep all elements optional. The committee agreed that it would be desirable at this early stage to encourage metadata suppliers to expose DC metadata according to their own needs and thereby reveal a market of community-developed metadata practices.

It should be noted that the specific decision was to use *unqualified* Dublin Core as the common metadata set. This decision was made based on the belief that the common metadata set in OAI is explicitly purposed for coarse granularity resource discovery. As discussed elsewhere [16], qualification of DC, for the purpose of more detailed *description* (rather than simple *discovery*) is still an area of some contention and threatens to interfere with the goal of simple resource discovery. The OAI takes the approach of strictly separating simple discovery from community-specific description.

Community-specific description, or metadata specificity, is addressed in the technical framework by support for parallel metadata sets. The technical framework places no limitations on the nature of such parallel sets, other than that the metadata records be structured as XML documents, which have a corresponding XML schema for validation (as described in section 4.2). At the time of completion of this paper (January 2001), initial steps have been taken to encourage the development of community-specific harvestable metadata sets. Representatives of the E-Print community have been working on a metadata set targeted at the E-Print community under the name EPMS. Representatives of the research library community have proposed a similar effort and there are calls for proposals from other communities (e.g., the museum community, Open Language Archives).

4.2 Records, Repositories, and Identifiers

The OAI technical framework defines a *record*, which is an XML-encoded byte stream that serves as a packaging mechanism for harvested metadata. A record has three parts:

- *header* – containing information that is common to all records (it is independent of the metadata format disseminated in the record) and that is necessary for the harvesting process. The information defined in the header is the unique identifier for the record (described below), and a datestamp indicating the date of creation, deletion, or latest date of change in the metadata in the record.
- *metadata* – containing metadata in a single format. As noted in section 4.1, all OAI data providers must be capable of emitting records containing unqualified DC metadata. Other metadata formats are optional.
- *about* – an optional container to hold data about the metadata part of the record. Typically, the "about" container could be used to hold rights information about the metadata, terms and conditions for usage of the metadata, etc. The internal structure of the "about" container is not defined by the protocol. It is left to individual communities to decide on its syntax and semantics through the definition of a schema.

A sample OAI record is shown in Figure 2.

Metadata records are disseminated from *Repositories*, which are network accessible servers of data providers. An OAI-conformant repository supports the set of OAI protocol requests defined in Section 4.4. Abstractly, repositories contain *items*, and each metadata record harvested from a repository corresponds to an item. (There is a many-to-one relationship of records to items, since metadata can be expressed in multiple formats). The nature of an item - for example, what type of metadata is actually stored

[8] http://www.openarchives.org/OAI/openarchivesprotocol.htm.

in the item, what type is derived on the fly, and whether the item includes the "full content" described by the metadata - is outside the scope of the OAI protocol. This admittedly indistinct nature of an item is intentional. The OAI harvesting protocol is meant to be agnostic as to the nature of a data provider – it supports those that have content with fixed metadata records, those that computationally derive metadata in various formats from some intermediate form or from the content itself, or those that are metadata stores or metadata intermediaries for external content providers.

```
<header>
  <identifier>oai:arXiv:9901001</identifier>
  <datestamp>1999-01-01</datestamp>
 </header>
 <metadata>
  <dc xmlns="http://www.openarchives.org/OAI/dc.xsd">
   <title>Quantum slow motion</title>
   <creator>Hug, M.</creator>
   <creator>Milburn, G. J.</creator>
   <date>1999-01-01</date>
   <type>e-print</type>
  <identifier>http://arXiv.org/abs/9901001</identifier>
  </dc>
 </metadata>
  <about>
  <dc xmlns="  httpd://www.openarchives.org/OAI/dc.xsd>
    <rights>Metadata may be used without restrictions</rights>
  </dc>
  </about>
```

Figure 2 – Sample OAI Record

As illustrated in Figure 2 each record has an identifier. The nature of this identifier deserves some discussion. Concretely, the record identifier serves as a key for extracting metadata from an item in a repository. This key, parameterized by a metadata format identifier, produces an OAI record. Since the identifier acts in this manner as a key it must be unique within the repository; each key corresponds to metadata derived from one item. The protocol itself does not address the issues of inter-repository naming or globally unique identifiers. Such issues are addressed at the level of registration, which is described in Section 4.5

The record identifier is expressly *not* the identifier of the item – the issue of identifiers for contents of repositories is intentionally outside the scope of the OAI protocol. Undoubtedly, many clients of the OAI protocol will want access to the full content described by a metadata record. The protocol recommends that repositories use an element in metadata records to establish a linkage between the record and the identifier (URL, URN, DOI, etc.) of the associated item. The mandatory Dublin Core format provides the *identifier* element that can be used for this purpose.

4.3 Selective Harvesting

A protocol that only enabled consumers of metadata to gather all metadata from a data provider would be cumbersome. Imagine the transactions with large research libraries that expose the metadata in their entire catalog through such a protocol!

Thus, some provision for selective harvesting, which makes it possible in the protocol to specify a subset of records to be harvested, is desirable. Selection, however, has a broad range of functionality. More expressive protocols include provisions for the specification of reasonably complete predicates (in the manner of database requests) on the information requested. The OAI decided that such high functionality was not appropriate for a low-barrier protocol and instead opted for two relatively simple criteria for selective harvesting.

- *Date-based* – As noted in Section 4.2, every record contains a date stamp, defined as "the date of creation, deletion, or latest date of modification of an item, the effect of which is a change in the metadata of a record disseminated from that item". Harvesting requests may correspondingly contain a date range for harvesting, which may be total (between two dates) or partial (either only a lower bound or an upper bound). This date-based harvesting provides the means for incremental harvesting. For example, a client may have a weekly schedule for harvesting records from a repository, and use the date-based selectivity to only harvest records added or modified since the last harvesting time.
- *Set-based* – The protocol defines a *set* as "an optional construct for grouping items in a repository for the purpose of selective harvesting of records". Sets may be used in harvesting requests to specify that only records within a specific grouping should be returned from the harvesting request (note that each item in a repository may be organized in one set, several sets, or no sets at all). Each repository may define a hierarchical organization of sets that can have several top-level nodes, each of which is a *set*. Figure 3 illustrates a sample set hierarchy for a fictional E-Print repository. The actual meaning of the sets is not defined within the protocol. Instead, it is expected that communities that use the OAI protocol may formulate well-defined set configurations with perhaps a controlled vocabulary for set names, and may even develop mechanisms for exposing these to service providers. As experiments with the OAI protocol proceed in the future, it will be interesting to see how communities exploit the set mechanism and if it provides sufficient functionality.

Institutions
- Cornell University
- Virginia Tech

Subjects
- Computer Science
- High Energy Physics

Figure 3 - Sample Set Hierarchy

Even with the provisions for selective harvesting, it is possible that clients will make harvesting requests of repositories that are large and burdensome to fulfill in a single response. Some other protocols make provision for such cases with the notion of *state* and *result sets* – a client explicitly opens a session, conducts transactions within that session, and then closes a session. Yet, session maintenance is notably complex and ill suited for protocols such as HTTP, which is intended as the carrier protocol for OAI requests and responses. Instead the OAI uses a relatively simple flow control mechanism that makes it possible to partition large transactions among several requests and responses. This flow control mechanism employs a *resumption token*, which is returned by a repository when the response to a harvesting request is larger than the repository may wish to respond to at one time. The client can then use the resumption token to make subsequent requests until the transaction is complete.

4.4 Open Archives Metadata Harvesting Protocol

The initial protocol that came out of the Santa Fe meeting was a subset of the Dienst protocol. While that subset protocol was functionally useful for metadata harvesting, aspects of its legacy context presented barriers to simple implementation. The current technical framework is built around a more focused and easier to implement protocol – the Open Archives Metadata Harvesting Protocol.

The Open Archives Metadata Harvesting Protocol consists of six requests or verbs. The protocol is carried within HTTP POST or GET methods. The intention is to make it simple for data providers to configure OAI conformant repositories by using readily available Web tools such as libwww-perl[9]. OAI requests all have the following structure:

- *base-url* – the Internet host and port of the HTTP server acting as a repository, with an optional path specified by the respective HTTP server as the handler for OAI protocol requests.
- *keyword arguments* – consisting of a list of key-value pairs. At a minimum, each OAI protocol request has one key=value pair that specifies the name of the OAI protocol request.

Figure 4 shows the encoding of a sample OAI protocol request using both HTTP GET and POST methods. The request is the *GetRecords* verb, and the specific example requests the return of the record with identifier *oai:arXiv:hepth01* in *dc* (Dublin Core) format.

The response to all OAI protocol requests is encoded in XML. Each response includes the protocol request that generated the response, facilitating machine batch processing of the responses. Furthermore, the XML for each response is defined via an XML schema. [17-19]. The goal is to make conformance to the technical specifications as machine verifiable as possible – a test program should be able to visit an OAI repository, issue each protocol request with various arguments, and test that each response conforms to the schema defined in the protocol for the response.

The remainder of this section summarizes each of the protocol requests.

GET Request

```
http://ana.oa.org/OAI-script?
        verb=GetRecord&
        identifier=oai:arXiv:hep-th01&
        metadataPrefix=dc
```

POST Request

```
POST http://an.oa.org/OAI-script
Content-Length: 62
Content-Type: application/x-www-form-urlencoded
verb=GetRecord&
identifier=oai:arXiv:hep-th01&
metadataPrefix=dc
```

Figure 4 - Sample OAI Request Encoding

4.4.1 GetRecord

This verb is used to retrieve an individual record (metadata) from an item in a repository. Required arguments specify the identifier, or key, of the requested record and the format of the metadata that should be included in the record.

4.4.2 Identify

This verb is used to retrieve information about a repository. The response schema specifies that the following information should be returned by the *Identify* verb:

- A human readable name for the repository.
- The base URL of the repository.
- The version of the OAI protocol supported by the repository.
- The e-mail address of the administrator of the repository.

In addition to this fixed information, the protocol provides a mechanism for individual communities to extend the functionality of this verb. The response may contain a list of *description* containers, for which a community may define an XML schema that specifies semantics for additional description of the repository.

4.4.3 ListIdentifier

This verb is used to retrieve the identifiers of records that can be harvested from a repository. Optional arguments permit selectivity of the identifiers - based on their membership in a specific set in the repository or based on their modification, creation, or deletion within a specific date range.

4.4.4 ListMetadataFormats

This verb is used to retrieve the metadata formats available from a repository. An optional argument restricts the request to the formats available for a specific record.

4.4.5 ListRecords

This verb is used to harvest records from a repository. Optional arguments permit selectivity of the harvesting - based on the membership of records in a specific Set in the repository or based

[9] http://www.ics.uci.edu/pub/websoft/libwww-perl/.

on their modification, creation, or deletion within a specific date range.

4.4.6 ListSets

This verb is used to retrieve the set structure in a repository.

4.5 Data Provider Conformance and Registration

The OAI expects that data providers will fall into three layers of participation, each higher layer implying the preceding layer(s);

1) *OAI-conformant* – These are data providers who support the protocol definition. As stated earlier, conformance is testable since there are XML-schemas to validate all responses. No doubt, the OAI will not be able to track every provider using the protocol since use of it does not require any licensing or registration procedure.

2) *OAI-registered* – These are data providers who register in an OAI-maintained database, which will be available through the OAI web site. Registration will entail that the data provider gives a BASE-URL, which the registration software will then use to test compliance.

3) *OAI-namespace-registered* – These are data providers who choose to name their records in conformance with an OAI naming convention for identifiers. Names that follow this convention have the following three components:

   a) *oai* – A fixed string indicating that the name is in the OAI namespace.

   b) *<repoID>* - An identifier for the repository that is unique within the OAI namespace.

   c) *<localID>* - An identifier unique within the respective repository.

   An example of a name that uses this naming scheme is: *oai:arXiv:hep-th01*
   The advantage for repositories of adopting this naming convention is that record identifiers will be resolvable via a central OAI resolution service, that will be made available at the OAI web site. The intention is to make this resolver OpenURL-aware[10], as a means to support open linking [20-22] based on OAI identifiers. The attractiveness of such an approach has been demonstrated in an experiment conducted in the DOI namespace[11]. A process for fast-track standardization of OpenURL has recently started with NISO.

5. TESTING AND REFINEMENT

Participants of the September 2000 technical meeting at Cornell developed a rough outline of the technical framework. However, the task of normalizing and putting the framework into the form of a specification was undertaken at Cornell University, where Open Archives Initiative activities are coordinated. Continuous feedback from the alpha-test group that implemented consecutive versions of the protocol played an important role in this activity. Participants in the alpha-test group were solicited from both the original Santa Fe Convention E-Print community and from the attendees at the DLF sponsored Cambridge meetings. These solicitations led to a quite comprehensive and diverse testing community, organized around an alpha testers email list. The complete list of alpha testers is shown in the Appendix in section 8. It includes representatives from the E-Print community, museums, research libraries, repositories of publisher metadata, and collectors of web site metadata. In addition, the alpha test included two rudimentary service providers who constructed search interfaces based on metadata harvested from the OAI-conformant alpha testers.

Three alpha-testers deserve special mention. Hussein Suleman at Virginia Tech created and continuously updated a repository explorer that allowed alpha-testers to examine the compliance of their repositories to the most recent version of the protocol document. Simeon Warner (Los Alamos National Laboratory and arXiv.org) and Michael Nelson (University of Northern Carolina and NASA) did extensive proofreading of new versions of the protocol document before release to the alpha-group.

The results of these tests are quite encouraging. First, the protocol specification has passed through a number of revisions and has been vetted extensively for errors and ambiguities. Second, virtually all the testers remarked at the ease of implementation and conceptual simplicity of the protocol.

6. THE ROAD AHEAD

At the beginning of 2001 the Open Archives Initiative began the next phase of its work, public deployment and experimentation with the technical framework. To initiate this process two public meetings were scheduled. A US meeting was be held in Washington DC on January 23. Registration for this meeting was closed when the maximum of 140 participants was reached. The participants represented a wide variety of communities. A European meeting was scheduled for February 26 in Berlin. Participants at these meetings heard a complete overview of the goals of the OAI and the particulars of the technical framework. The meetings also provided an opportunity for the development of communities within the Open Archives framework. Communities may take the form of groups of data providers that exploit the extensibility of the Open Archives Harvesting Protocol to expose purpose-specific metadata or the development of targeted services.

These meetings were meant to be a "kick off" for an extended period of experimentation (at least one year) with the harvesting protocol. The OAI intends during this period to keep the protocol as stable as possible. This experimentation phase is motivated by the belief that the community needs to fully understand the functionality and limits of the interoperability framework before considering major changes or expansion of functionality.

During this experimentation period, three large research and implementation projects, in the U.S. and in Europe, plan to experiment with the functionality of the OAI technical framework:

1. *National Science Digital Library*[12] – NSDL is a multi-participant project in the US funded by the National Science Foundation with the goal of creating an online network of learning environments and resources for science, mathematics, engineering, and technology education. Our group at Cornell is funded under the core infrastructure

[10] http://www.sfxit.com/OpenURL

[11] http://sfxserv.rug.ac.be:8888/public/xref/

[12] http://www.ehr.nsf.gov/ehr/due/programs/nsdl/.

portion of the NSDL. In the context of the OAI alpha testing we experimented with a harvesting service that will later form the basis for other services in our NSDL infrastructure. Future plans include working with our partners at the San Diego Super Computer Center to harvest metadata via OAI, post-process and normalize it for storage in a metadata repository, and then make that metadata searchable using the SDLIP [23] protocol.

2. *Cyclades*[13] - This is a project funded by the European Commission with partners in Italy, Germany, and the U.K. The main objective of Cyclades is to develop advanced Internet accessible mediator services to support scholars both individually and as members of communities when interacting with large interdisciplinary electronic archives. Cyclades plans to investigate the construction of these services on the Open Archive foundation.
3. *Digital Library Federation Testbed* – As a follow-up to the Cambridge meetings (described earlier in this paper) a meeting[14] of interested project participants was convened in October 2000 by The Andrew W. Mellon Foundation to explore technical, organizational, and resource issues for broad-based metadata harvesting and to identify possible next steps. The result of this meeting was the commitment by a number of institutions (research libraries, other information-based organizations) to expose metadata from a number of collections through the Open Archives technical infrastructure, and experiment with services that use this metadata.

These projects promise to expose what is possible using the OAI framework and how it might be changed or expanded.

We are intrigued by the period of experimentation that lies ahead and encouraged by the widespread interest in the Open Archives Initiative. In the context of this enthusiasm, we are aware of the need to be circumspect in our approach towards the OAI and its goals. As Don Waters, a member of the OAI Steering Committee, pointed out, the technical proposals of the OAI include a number of assumptions about issues not yet fully understood.

- What is the value of a common metadata set?
- What are the interactions of native metadata set with the minimal, conventional set?
- What are the incentives and rewards for institutions and organizations for participating in such a framework?
- What are the intellectual property issues vis-à-vis harvestable metadata?
- Will this technical framework encourage new models of scholarly communication?

These, and many other questions, are all in need of thorough examination. Too often members of the digital library community have made casual statements that "interoperability is good", "metadata is important", and that "scholarly publishing is changing". At the minimum, we hope that the OAI will create a framework for serious investigation of these issues and lay the foundation for more informed statements about the issues critical to the success of our field.

7. Appendix A – OAI STEERING COMMITTEE

Names are followed by affiliations:

- Caroline Arms (Library of Congress)
- Lorcan Dempsey (Joint Information Systems Committee, UK)
- Dale Flecker (Harvard University)
- Ed Fox (Virginia Tech)
- Paul Ginsparg (Los Alamos National Laboratory)
- Daniel Greenstein (DLF)
- Carl Lagoze (Cornell University)
- Clifford Lynch (CNI)
- John Ober (California Digital Library)
- Diann Rusch-Feja (Max Planck Institute for Human Development)
- Herbert Van de Sompel (Cornell University)
- Don Waters (The Andrew W. Mellon Foundation)

8. Appendix B – ALPHA TEST SITES

The following institutions participated in the alpha testing of the technical specifications:

- CIMI Museum Consortium
- Cornell University
- Ex Libris
- Los Alamos National Laboratory
- NASA
- OCLC
- Old Dominion University
- UKOLN Resource Discovery Network
- University of Illinois Urbana Champaign
- University of North Carolina
- University of Pennsylvania
- University of Southampton
- University of Tennessee
- Virginia Tech

9. ACKNOWLEDGMENTS

Our thanks to the many members of the Open Archives Initiative Community – steering committee, technical committee, alpha testers, and workshop attendees – who have participated in the development of the OAI over the last year. Thanks to Naomi Dushay for her careful reading and editing. Support for work on the Open Archives Initiative comes from the Digital Library Federation, the Coalition for Networked Information, the National Science Foundation (through grant no. IIS-9817416) and DARPA (through contract no. N66001-98-1-8908).

[13] http://cinzica.iei.pi.cnr.it/cyclades/,

[14] http://www.clir.org/diglib/architectures/testbed.htm.

10. REFERENCES

[1] Van de Sompel, H. and C. Lagoze, *The Santa Fe Convention of the Open Archives Initiative.* D-Lib Magazine, 2000. **6**(2).http://www.dlib.org/dlib/february00/vandesompel-oai/vandesompel-oai.html.

[2] Ginsparg, P., *Winners and Losers in the Global Research Village.* The Serials Librarian, 1997. **30**(3/4): p. 83-95.

[3] Harnad, S., *Free at Last: The Future of Peer-Reviewed Journals.* D-Lib Magazine, 1999. **5**(12).http://www.dlib.org/dlib/december99/12harnad.html.

[4] Paepcke, A., *et al.*, *Interoperability for Digital Libraries Worldwide, Communications of the ACM.* 1998, **41**(4): 33-42.

[5] Halpern, J.Y. and C. Lagoze. *The Computing Research Repository: Promoting the Rapid Dissemination and Archiving of Computer Science Research.* In *Digital Libraries '99, The Fourth ACM Conference on Digital Libraries.* 1999. Berkeley, CA.

[6] Van de Sompel, H., T. Krichel, and M.L. Nelson, *The UPS Prototype: an experimental end-user service across e-print archives*, in *D-Lib Magazine.* 2000.

[7] Lagoze, C. and J.R. Davis, *Dienst - An Architecture for Distributed Document Libraries.* Communications of the ACM, 1995. **38**(4): 47.

[8] *A New Approach to Finding Research Materials on the Web*, . 2000, Digital Library Federation. http://www.clir.org/diglib/architectures/vision.htm.

[9] Anderson, K.M., *et al.*, *ACM 2000 digital libraries : proceedings of the fifth ACM Conference on Digital Libraries, June 2-7, 2000, San Antonio, Texas.* 2000, New York: Association for Computing Machinery. xiii, 293.

[10] Borbinha, J. and T. Baker, *Research and advanced technology for digital libraries: 4th European Conference, ECDL 2000, Lisbon, Portugal, September 18-20, 2000 : proceedings.* Lecture notes in computer science 1923. 2000, Berlin ; New York: Springer. xvii, 513.

[11] Arms, W.Y., *Digital libraries.* Digital libraries and electronic publishing. 2000, Cambridge, Ma.: MIT Press.

[12] Lagoze, C., C.A. Lynch, and R.D. Jr., *The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata.* 1996, Cornell University Computer Science. http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593.

[13] Lagoze, C. *Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience.* in *Seminar on Metadata.* 2000. Archiefschool, Netherlands Institute for Archival Education and Research, The Hague.

[14] Weibel, S., *The Dublin Core: A simple content description format for electronic resources.* NFAIS Newsletter, 1998. **40**(7): p. 117-119.

[15] Chang, C.-C.K. and H. Garcia-Molina. *Mind your vocabulary: query mapping across heterogeneous information sources.* In *International Conference on Management of Data and Symposium on Principles of Database Systems.* 1999. Philadelphia: ACM: 335-346.

[16] Lagoze, C. and T. Baker, *Keeping Dublin Core Simple.* D-Lib Magazine, 2001. **7**(1).

[17] Fallside (ed.), D.C., *XML Schema Part 0: Primer.* 2000, World Wide Web Consortium. http://www.w3.org/TR/xmlschema-0/.

[18] Thompson, H.S., *et al.*, *XML Schema Part 1: Structures.* 2000, World Wide Web Consortium. http://www.w3.org/TR/xmlschema-1/.

[19] Biron, P.V. and A. Malhotra, *XML Schema Part 2: Datatypes.* 2000, World Wide Consortium. http://www.w3.org/TR/xmlschema-2/.

[20] Van de Sompel, H. and P. Hochstenbach, *Reference Linking in a Hybrid Library Environment , Part 3: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment.* D-Lib Magazine, 1999. **5**(10).http://www.dlib.org/dlib/october99/van_de_sompel/10van_de_sompel.html.

[21] Van de Sompel, H. and P. Hochstenbach, *Reference Linking in a Hybrid Library Envronment:, Part 1: Frameworks for Linking.* D-Lib Magazine, 1999. **5**(4).http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html.

[22] Van de Sompel, H. and P. Hochstenbach, *Reference Linking in a Hybrid Library Environment, Part 2: SFX, a Generic Linking Solution*, in *D-Lib Magazine.* 1999.

[23] Paepcke, A., *et al.*, *Search Middleware and the Simple Digital Library Interoperability Protocol.* D-Lib Magazine, 2000. **5**(3).http://www.dlib.org/dlib/march00/paepcke/03paepcke.html.

Enforcing Interoperability with the Open Archives Initiative Repository Explorer

Hussein Suleman
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
+1 540 231-3615
hussein@vt.edu

ABSTRACT

The Open Archives Initiative (OAI) is an organization dedicated to solving problems of digital library interoperability by defining simple protocols, most recently for the exchange of metadata. The success of such an activity requires vigilance in specification of the protocol as well as standardization of implementation. The lack of standardized implementation is a substantial barrier to interoperability in many existing client/server protocols. To avoid this pitfall we developed the Repository Explorer, a tool that supports manual and automated protocol testing. This tool has a significant impact on simplifying development of interoperability interfaces and increasing the level of confidence of early adopters of the technology, thus exemplifying the positive impact of exhaustive testing and quality assurance on interoperability ventures.

Categories and Subject Descriptors

D.2.12 [**Software Engineering**]: Interoperability – *Interface definition languages.*

C.2.2 [**Computer-Communication Networks**]: Network Protocols – *Protocol verification.*

General Terms

Reliability, Experimentation, Standardization, Verification.

Keywords

Interoperability, Protocol, Testing, Validation.

1. CONTEXT

The work of the Open Archives Initiative (OAI) was initiated in connection with a meeting of representatives of various electronic pre-print and related archives (e.g. NDLTD, arXiv, NCSTRL) in Santa Fe, USA in October 1999 [2]. From this meeting emanated an agreement among the archivists to support a common set of principles and a technical framework to achieve interoperability. This started the process of defining standards, broadened to include digital libraries other than pre-print archives.

A technical working group provided input into the process of creating and testing those standards under widely differing conditions. This process culminated in the announcement of a new protocol for interoperability, the Open Archives Initiative Protocol for Metadata Harvesting [3], in January 2001. These standards are now being disseminated to all interested parties who wish to adopt a low-cost approach to interoperability, with support from a growing number of members of the Open Archives community.

2. MOTIVATION

After the inaugural meeting of the OAI in 1999, a handful of archivists began to implement the agreed-upon interoperability protocol at their distributed sites. This effort was immediately hampered by a varying interpretation of the protocol specification. This was largely due to the difficulty of precisely specifying a protocol that would both be general and applicable to multiple domains. The client/server architecture chosen by the OAI led to a classic "chicken-and-egg" problem since client implementations would need to interface correctly with server implementations – there was subsequently a low degree of confidence in the correctness of early implementations in each category. Coupled with this, even when clients and servers subscribed to the same interpretation, there was not high confidence that other client/server pairs would interoperate successfully.

As one approach to address these concerns, we developed a protocol tester that would allow a user to perform low-level protocol tests on a server implementation without the need for a corresponding client implementation. This now-widely used software, the Repository Explorer, aids in standardizing the protocol understood by various different archives subscribing to the OAI model of interoperability.

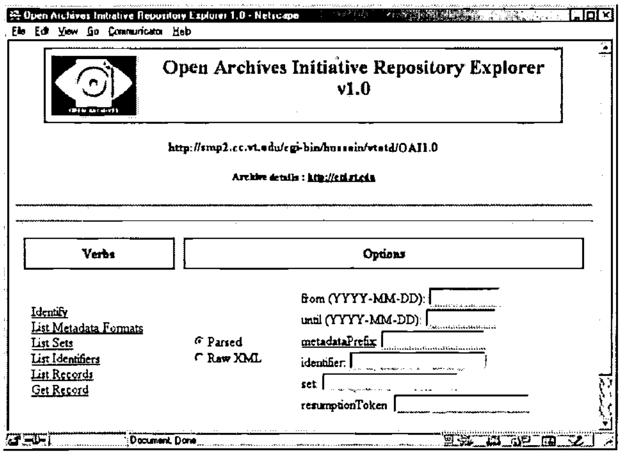
3. DESIGN OF REPOSITORY EXPLORER

The Repository Explorer is implemented as a web-based application (see Figure 1) to take advantage of the ubiquitous nature of WWW clients, and to alleviate the need to install multiple components on client machines to support all the software components used during testing.

The software supports both manual and automatic testing, but with an emphasis on the former. In automatic testing mode, a series of protocol requests, with legal and illegal combinations of parameters, are issued to the archive being tested, and the responses are checked for compliance with the expected range of responses. In manual mode, the software allows a user to browse through the contents of the archive using only the well-defined

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

interface provided by the protocol – in this instance the user has full control over all parameters and can test individual features of the protocol.


Open Archives Initiative Repository Explorer 1.0 - Netscape
File Edit View Go Communicator Help
Document: Done

Open Archives Initiative Repository Explorer v1.0

http://smp2.cc.vt.edu/cgi-bin/hussein/vtetd/OAI1.0

Archive details : http://etd.vt.edu

Verbs		Options
Identify List Metadata Formats List Sets List Identifiers List Records Get Record	● Parsed ○ Raw XML	from (YYYY-MM-DD): until (YYYY-MM-DD): metadataPrefix identifier: set resumptionToken

Figure 1. Repository Explorer basic interface

4. VALIDATION PROCEDURE

The OAI protocol is a request/response protocol that works as a layer over HTTP, with responses in XML. The Repository Explorer performs validation at multiple levels in an attempt to detect the widest range of possible errors. Figure 2 depicts the flow of response data received from a server during the validation process.

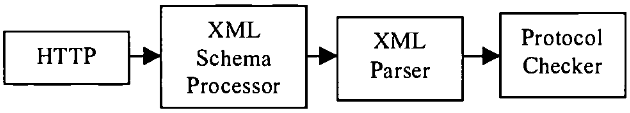


Figure 2. Outline of validation/testing process

Each step of this validation procedure performs incremental checking as described below:

1. When submitting an HTTP request, HTTP errors need to be detected and handled. The OAI protocol requires that explicitly illegal requests generate errors as HTTP status codes, and these are checked for.
2. Once the request is issued and the response is successful, the returned XML needs to be checked for validity. Since all XML responses are specified in the protocol specification using the XML Schema Language [1], this part of the validation is accomplished using an XML Schema processor, which attempts to match the schema with the generated XML data stream. Many structural and encoding errors in the XML are detected in this phase of testing.
3. Unfortunately, schema processing does not always work flawlessly because of its many external dependencies, e.g., schemata are downloaded from the WWW as required. As a redundant mechanism, XML errors are also detected during the parsing and tree generation phase.
4. Lastly, the tree representation is checked for semantic correctness. For example, where controlled vocabularies are used but not encoded into the XML Schema, these are checked at this stage (e.g., the standard list of metadata formats used by the OAI)

Once checks are performed, if the software is being used in manual mode, a new interface is generated to display the data received from the server and present the user with additional options to perform further operations.

5. CONCLUSIONS

Over the course of implementing the original and revised protocol, most if not all implementers used the Repository Explorer to test their implementations. Their positive feedback supported the original motivation that a compliance test would greatly ease the process of implementation.

Many lessons were learnt during the process of designing the test software, the most significant being that testing is a decidedly non-trivial problem if the target (namely, the protocol specification) is not stationary. However, designing a new version of the protocol in tandem with updating the test software ensured that the new specifications contain mechanisms (like self-identification) that can be exploited to perform more exhaustive automatic tests.

6. FUTURE WORK

Further standardization of the software libraries used in development can lead to a tool suite that will not only be adaptable to future versions of the OAI protocol, but also possibly to other client/server protocols. While XML tools are still not widely deployed and are notorious for high degrees of complexity, future combinations of XSD (Schema), XSLT (Transformation) and XPath (Path Specifications) could make protocol testing a fully automated process for emerging client/server protocols using XML as underlying technology.

The test suite is under constant development as best practices emerge among users of the community, thus ensuring that the Repository Explorer maintains its status as a compliance test, for which it was intended.

7. ACKNOWLEDGMENTS

Many thanks go to the group of OAI protocol authors and testers who used the Repository Explorer and provided useful feedback.

8. REFERENCES

[1] Fallside, David C. (editor). XML Schema. W3C, October 2000. http://www.w3.org/XML/Schema

[2] Van de Sompel, Herbert and Carl Lagoze. The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, Volume 6, Number 2, February 2000. http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

[3] Van de Sompel, Herbert and Carl Lagoze. The Open Archives Initiative Protocol for Metadata Harvesting. Open Archives Initiative, January 2001. http://www.openarchives.org/OAI/openarchivesprotocol.htm

Arc — An OAI Service Provider for Cross-Archive Searching

Xiaoming Liu, Kurt Maly, Mohammad Zubair, and Michael L. Nelson
Old Dominion University
Department of Computer Science
Norfolk, VA 23592 USA
+1 757 683 4017
{liu_x, maly, zubair, nelso_m}@cs.odu.edu

ABSTRACT

The usefulness of the many on-line journals and scientific digital libraries that exist today is limited by the lack of a service that can federate them through a unified interface. The Open Archive Initiative (OAI) is one major effort to address technical interoperability among distributed archives. The objective of OAI is to develop a framework to facilitate the discovery of content in distributed archives. In this paper, we describe our experience and lessons learned in building *Arc*, the first federated searching service based on the OAI protocol. *Arc* harvests metadata from several OAI compliant archives, normalizes them, and stores them in a search service based on a relational database (MySQL or Oracle). At present we have over 165K metadata records from 16 data providers from various domains.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: *collection, dissemination, standards.*

General Terms

Design, Experimentation, Standardization, Languages

Keywords

Digital Library, Open Archive Initiative

1. INTRODUCTION

A number of free on-line journals and scientific digital libraries exist today, however, there is a lack of a federated service that provides a unified interface to all these libraries. The Open Archive Initiative (OAI) [1] is one major effort to address technical interoperability among distributed archives. The objective of OAI is to develop a framework to facilitate the discovery of content in distributed archives. The OAI framework supports data providers (archives) and service providers. The service provider develops value-added services that are based on the information collected from data providers.

These value-added services could take the form of cross-archive search engines, linking systems, and peer-review systems. OAI is becoming widely accepted and there are many archives currently or soon-to-be OAI compliant. *Arc* (http://arc.cs.odu.edu) is the first federated search service based on the OAI protocol, and its concept originates from the Universal Preprint Service (UPS) prototype [2]. We encountered a number of problems in developing *Arc*. Different archives have different format/naming conventions for specific metadata fields that necessitate data normalization. Arbitrary harvesting can over-load the data provider making it unusable for normal purposes. Initial harvesting when a data provider joins a service provider requires a different technical approach than periodical harvesting that keeps the data current.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

2. Architecture

The *Arc* architecture is based on the Java servlets-based search service that was developed for the Joint Training, Analysis and Simulation Center (JTASC) [3]. This architecture is platform independent and it can work with any web server. Moreover, the changes required to work with different databases are minimal. Our current implementation supports two relational databases, one in the commercial domain (Oracle), and the other in public domain (MySQL). The architecture improves performance over the original UPS architecture by employing a three-level caching scheme [3]. Figure 1 outlines the major components; however, for brevity we discuss only the components relevant to this paper.

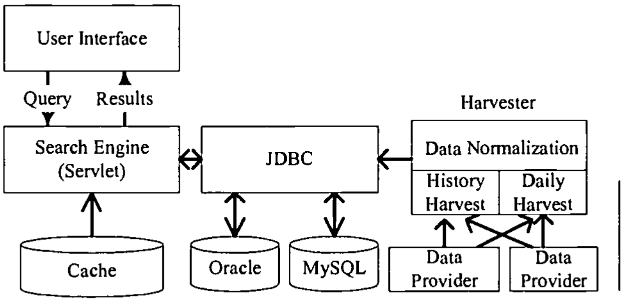


Figure 1. *Arc* **Architecture**

2.1 Harvester

Data providers are different in data volume, partition definition, service implementation quality and network connection quality. All these factors influence the harvesting procedure. Historical and newly-published data harvesting have different requirements. When a data provider joins a service provider for the first time, all past data (historical data) needs to be harvested, followed by periodic harvesting to keep the data current. Historical data are high-volume and more stable, the harvesting process generally runs once, and a chunk-based harvest is preferred to reduce accessing time and data provider overhead. To harvest newly published data, data size is not the major problem but the scheduler must be able to harvest new data as soon as possible and guarantee completeness – even if data providers provide incomplete data for the current date. The OAI protocol provides flexibility in choosing the harvesting strategy; theoretically, one data provider can be harvested in one simple transaction, or one is harvested as many times as the number of records in its collection. But in reality only a subset of this range is possible; choosing an appropriate harvesting method has not yet been made into a formal process. We defined four harvesting types in *Arc*: (a) bulk-harvest of historical data (b) bulk-harvest of new data (c) one-by-one-harvest of historical data (d) one-by-one-harvest of new data. From our tests, these four strategies in combination can fulfill various requirements of a particular collection.

2.2 Database Schema

OAI uses Dublin Core (DC) as the default metadata set. All DC attributes are saved in the database as separate fields. The archive name and partition are also treated as separate fields in the database for supporting search and browse functionality. In order to improve system efficiency, most fields are indexed using full-text properties of the database, which makes high performance queries over large dataset possible. The search engine communicates with the database using JDBC and Connection Pool.

2.3 Search Interface specification

The search interface supports both simple and advanced search as well as result sorting by date stamp, relevance ranking and archive. Simple search allows users to search free text across archives. Advanced search allows user to search in specific metadata fields. Users can also search/browse specific archives and/or archive partitions in case they are familiar with specific data provider. Author, title, abstract search are based on user input, the input can use boolean operators (AND, OR, NOT). Archive, set, type, language and subject use controlled vocabularies and are accumulated from the participating archives' source data.

Table 1. Collections Harvested by *Arc* (by Feb 7, 2001)

Archive Name	URL	Records
arXiv.org e-Print Archive	arxiv.org	151650
Cognitive Science Preprints	cogprints.soton.ac.uk	999
NACA	naca.larc.nasa.gov	6352
Networked Digital Library of Theses and Dissertations	www.ndltd.org	2401
Web Characterization Repository	repository.cs.vt.edu	131
NCSTRL in Comell	www.ncstrl.org	2080
NASA Langley Technical Report Server	techreports.larc.nasa.gov/ltrs	2323
Nine Other Collections For the OAI alpha test	arc.cs.odu.edu/help/archives.htm	9467

3. Results

We maintain two test sites, the public *Arc* cross archive service and another site for the OAI alpha test only. So far, in the public service, we have seven archives available for search (Table 1). In the OAI alpha test site, we have nine archives across different domains, including documents from Heinonline.org, Open Video Project, Language Resource Association, Library of Congress and other organizations. The alpha test data is not open to the public.

4. Lessons learned

Little is known about the long-term implications of a harvest-based digital library, but we have had the following initial experiences. The effort of maintaining a quality federation service is highly dependant on the quality of the data providers. Some are meticulous in maintaining exacting metadata records that need no corrective actions. Other data providers have problems maintaining even a minimum set of metadata and the records harvested are useless. We have not yet fully addressed the issue of metadata normalization. Some normalization was necessary to achieve a minimum presentation of query results. However, we did so on an ad hoc basis with no formal definition of the relationship mappings. A controlled vocabulary will be of great help for a cross archive search service to define such metadata fields as 'subject' or even 'organization'. XML syntax errors and character-encoding problems were surprisingly common and could invalidate entire large data sets. We also faced a trade-off in frequency of harvests: too many harvests could over burden both the service and data providers, and too few harvests allow the data in the service provider to potentially become stale.

5. REFERENCES

[1] Van de Sompel, H. and Lagoze, C. The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, 6(2), February 2000. http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

[2] Van de Sompel, H., Krichel, T., Nelson, M. L., Hochstenbach, P., Lyapunov, V. M., Maly, K., Zubair, M., Kholief, M., Liu, X. and O'Connell, H. The UPS Prototype: An Experimental End-User Service across E-Print Archives, *D-Lib Magazine* 6(2), February 2000. http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html

[3] Maly, K., Zubair, M., Anan, H., Tan, D. and Zhang, Y. Scalable Digital Libraries based on NCSTRL/Dienst. In *Proceedings of the 4th European Conference on Digital Libraries – ECDL 2000* (Lisbon, Portugal, September 2000), pp. 169-179.

Managing Change on the Web

Luis Francisco-Revilla, Frank Shipman, Richard Furuta, Unmil Karadkar, and Avital Arora

Center for the Study of Digital Libraries and Department of Computer Science
Texas A&M University
College Station, TX 77843-3112, USA

{l0f0954, shipman, furuta, unmil, avital}@csdl.tamu.edu

ABSTRACT

Increasingly, digital libraries are being defined that collect pointers to World-Wide Web based resources rather than hold the resources themselves. Maintaining these collections is challenging due to distributed document ownership and high fluidity. Typically a collection's maintainer has to assess the relevance of changes with little system aid. In this paper, we describe the Walden's Paths Path Manager, which assists a maintainer in discovering when *relevant* changes occur to linked resources. The approach and system design was informed by a study of how humans perceive changes of Web pages. The study indicated that structural changes are key in determining the overall change and that presentation changes are considered irrelevant.

Categories and Subject Descriptors

I.3.7 [**Digital Libraries**]: User issues;
H.5.4 [**Hypertext/Hypermedia**]: Other (maintenance)

General Terms

Algorithms, Management, Design, Reliability, Experimentation, Human Factors, Verification.

Keywords

Walden's Paths, Path Maintenance.

1. INTRODUCTION

The work in building a library is not only in the collection of materials, it is also in their organization and maintenance.

Books in a traditional library are actively managed—when acquired they must be organized, indexed and catalogued. When considered obsolete, they must be removed. Librarians go through great length in order to keep the collection up to date. New editions replace old ones, while old versions may be moved to archival sections or be discarded.

Digital libraries present an electronic counterpart of traditional libraries. Digital librarians are necessary to perform similar functions in order to maintain collections of electronic documents or electronic pointers to documents.

It is not rare to hear people refer to the Web as a new kind of library—or at least that it should be more like a library. However on the Web, not all electronic collections are well maintained—the degree of management of electronic collections varies considerably. Moreover, the characteristics of the Web raise many challenges for the maintainer. Collections on the Web can be extremely distributed—not only the location of the documents is distributed, but ownership is also distributed.

At the moment, the Web is less like a library and more like a huge bookshelf, containing millions of documents and links to documents. This mega shelf is subject to the activities of an army of people (and computer-based agents) that keep acting upon it in ways such as adding, removing, indexing, cataloguing, rearranging, modifying, and copying documents. Additionally, all these activities tend to occur with little or no coordination, worsening the situation.

Better coordination, social protocols, and institutionalized management of collection can alleviate the situation in some cases, particularly when the owner of the collection and the owner(s) of the documents agree to coordinate their activities. The emerging digital topical libraries, while homed on the Web, are better maintained than the Web as a whole. They present a more cohesive and organized structure in order to support their patrons. Different automatic mechanisms have been devised in order to help manage the documents. Indexing and cataloguing are considerably easier to perform than in the pre-electronic days. Nevertheless, considerable human effort is still required to monitor the changes in documents and collections.

In addition to the digital library that is provided by the "owner" of the documents (e.g., the ACM Digital Library) there are a growing number of specialized libraries with distributed ownership.[1] In these cases librarians often are limited to working with the pointers or links to the documents.

Digital documents are typically very fluid, i.e., changes are a common occurrence. Although detecting changes is easily

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

[1] One example is the National Science Foundation's NSDL effort, described at http://www.ehr.nsf.gov/due/programs/nsdl/

automated, assessing the relevance of changes is not. Typically humans are needed when facing the task of relevance assessment. In fact, "relevance" is a highly abstract, contextual, and subjective concept. Different individuals can disagree in their relevance judgment about any given change. Nevertheless, as difficult as it may be, assessing the relevance of a change is still necessary.

This paper presents an approach to aid humans in managing fluid collections of documents, in particular Web pages that have been authored and are owned by third parties. The approach provides a way of automatically assessing the relevance of changes in the Web pages. This method is explored in the Path Manager, a system designed to infer relevance of changes to Web pages and to communicate this information to the collection's maintainer.

The paper is divided into seven sections. Section two presents background about this work and its motivation. Sections three and four describe the approach used and its implementation as a prototype system. Section five gives an overview of a study to better understand how humans perceive changes, which we conducted to better understand how to refine the Path Manager. Sections six and seven conclude the paper.

2. MOTIVATION

Web readers access a rich and vast collection of information resources authored and published by a multitude of sources. In this collection it is possible to find information on virtually every topic, varying in aspects such as point of view, validity, semantics, rhetoric, and contextual situation. In order to better exploit this collection, different systems have been developed with the goal of aiding readers by providing an interpretation or contextualization of Web pages.

Walden's Paths is an application that allows teachers to construct trails [2] or paths using Web pages authored by others [7], [19]. The paths represent a higher order construct, or meta-document, that organizes and adds contextual information to pages authored by others. However, meta-documents that are to have a lasting value need to adapt to changes in their components. Thus, paths must address the issue of the unpredictable changing of their fundamental building blocks (Web pages). This is particularly important given the high fluidity of Web pages [3], where authors are prone to change their pages just to maintain a "fresh" look and feel.

Early in the Walden's Paths project issues related to the fluidity of Web pages rapidly arose. We observed that it is a common occurrence for paths to include Web pages that have moved, changed or are no longer available [18]. As a result the collection of paths needed to be constantly revised and updated.

To date in the project, we have implemented a number of approaches that can help to reduce the effects of fluidity. A simple approach is to cache the pages. This approach is one way to address the fluidity issue and also can expedite delivery of the information, particularly in schools that have slow Internet connections. However, it became evident to us that a single caching approach could not solve the problem, since not all Web page changes are undesirable (some paths include pages that by nature change, such as the current weather or a newspaper Web page). Our approach, implemented through the Walden's Paths Authoring Tool, allows authors to specify the caching strategy for individual pages [12]. We note, however, that issues about versioning and intellectual property rights remain, especially when caching Web pages created by third parties.

A second approach might be to increase the fluidity of the paths themselves—perhaps showing only pages that have not moved or changed. Walden's Paths provides an implementation mechanism for this through ephemeral paths [6]. Ephemeral paths are paths that exist for only a short period of time as the result of some computation. Unfortunately the effectiveness of the technique is limited here since many path authors design their paths to have a rhetorical coherence based on the linearity imposed by the path mechanism. Hence, when some pages of the path are removed the rhetorical structure can be broken, rendering the path ineffective.

More importantly, caching and ephemeral paths only provide *mechanisms* for implementing ways of reacting to change. Humans must determine when significant change has occurred. Making this determination requires that the nature of the change be considered within the context of the maintainer's goals (i.e., that its *relevance* be considered). For instance, even though a page may change in appearance or wording, it might remain conceptually the same. For many applications, this would be an insignificant change.

Before proceeding, and in order to avoid possible confusion, it is important to define the different people and components involved in the problem. While the issues addressed in the present work relate to any meta-document application, our interest originated from research conducted in the Walden's Paths project. Therefore meta-documents are defined as "paths" even though this is not the only possible meta-document. Similarly, construction blocks are referred to as "Web pages" or "pages". The first group involved is the people that originally create and publish the Web pages. They are identified to as "page creators". Another group is the people that create the meta-documents or paths. This group is denoted in this document as "path authors" or just "authors". Additionally there is the group of people accessing or using the paths, which is designated as "readers". Finally there is the group of people who has to manage the collection of paths and are referred as "administrators". It is important to consider that any given person might act in any or multiple of these roles at any given time.

2.1 Assessing Change Relevance

We are pursuing approaches to assist path authors and administrators in managing continuously occurring changes in their selected collection of Web pages. While our emphasis is on paths, our approach will be equally applicable to other Web-based domains that need to assess the relevance of changes to Web pages authored by third parties. We would expect, for example, that simple modifications to the Path Manager would enable the maintenance of personal bookmark lists or other collections of resource links.

Simply detecting changes in a page is easy. The simplest implementation would track the "last modified" date returned by the HTTP server. A more accurate mechanism would retain cached copies of Web pages. It is easy to compare a document with a previously cached version and determine if there has been any change. In a similar way, it would be possible to check for changes within specific sections of the document. These cases would return a Boolean "yes/no" indication of change.

As mentioned before, a difficulty arises when attempting to determine the relevance of changes. The obvious approach would be to assess the magnitude or amount of change. Unfortunately, there is not an easily computed metric that provides a direct correlation between syntactic and semantic changes in a Web page

For instance, there is no clear relationship between the number of bytes changed and the relevance of the change to the reader. A large number of bytes changed might result from a page creator who restructures the spacing of a page's source encoding while maintaining the same content from a semantic and rhetorical point of view. Similarly a small number of bytes changed might result from the insertion of a few negations in the text that causes a complete reversal in meaning. On the other hand, we can think of other situations where a large number of byte changes correspond to the creation of a completely different page, and situations where a few bytes are changed when the page creator simply corrects minor spelling and grammatical errors.

The goal would be to efficiently obtain a measure of the semantic distance between two versions of a document. However, at the moment the best we could attempt is to infer the semantic distance based on the syntactic characteristics of the document. In order to accomplish this, some heuristic actions are reasonable. For example the document can be partitioned based on heuristics such as:

- Paragraphs tend to encapsulate concepts
- Different paragraphs tend to encapsulate different concepts

Another way to partition the document is to analyze the document encoding. Markup languages such as SGML and XML already specify structure rather than presentation.

Even though Web pages are encoded in HTML, analyzing the changes is not an easily automated task. While HTML provides some information about the structure of the document, it is commonly used for specifying presentation rather than structure. Adding to that, page creators use HTML in extremely different ways. Paragraphs, headings and other tags are used quite diversely. What conceptually constitutes a paragraph for one page creator might constitute several pages for another. This does not mean that HTML tags are useless for analyzing change. On the contrary, HTML tags and other features such as keywords can be used in order to infer the relevance of changes.

2.2 Related Work

Measuring the magnitude or relevance of changes in an automated fashion is not an easy task. Researchers have encountered this problem in a variety of contexts and their approaches have informed our work.

In his doctoral dissertation David Johnson created a system for authoring and delivering tutorials over the Web [10], [11]. Johnson designed a signature-based approach when he also faced the issue of ever-changing Web pages. Johnson's approach computed a "distance" between two versions of a document by employing weighted values for additions and deletions of paragraphs, headings and keywords. Based on this measurement, Johnson created a mechanism that notifies readers and even blocks Web pages from showing when the distance between the two versions reaches predetermined trigger levels. While his testing was satisfactory with a particular collection of Web pages there are a couple of issues that hinder exporting the approach to the World Wide Web. The first issue is the dependence of the distance measure on the arbitrary determination of the weights assigned to the addition and deletions. The use of these weights might not work for a different collection of Web pages. The second issue is the asymmetric nature of the distance measurement and its lack of normalization. That is, the distance from document A to document B might be different from the distance from document B to document A. In addition there are no normalized values for the distance. Therefore the distance between two documents could be a number like 0.5 or 20. This makes setting trigger levels an arbitrary matter.

There are other approaches that monitor features of Web pages at a fine-grained level. Currently it is possible to find Web-based systems that provide fine-grained monitoring of changes such as AIDE[4], URL-Minder [22] and WatzNew [23]. Systems like these allow monitoring text, links, images and keywords of a given Web page. However a typical critique is that cosmetic changes are reported to be as relevant as substantive content changes. In contrast, the Path Manager evaluates the relevance of the change with regard to the whole page.

The goal of the Path Manager refers to identifying "interesting" or relevant changes to Web pages. As such, its design has been informed by other research work dealing with identifying "interesting" Web pages. In particular, research in page relevance has helped point out possible features of a page that should be monitored more attentively.

Researchers at the University of California at Irvine have investigated the issues of identifying readers' interests and dealing with changing Web pages. They have developed systems such as Syskill and Webert [17] and the Do-I-Care-Agent (DICA) [20], [21]. Syskill and Webert is an agent designed to discover interesting pages on the Web. The approach in this system is to aid readers with long-term information seeking goals. DICA is an agent designed to monitor reader-specified Web pages and notify the reader of important changes to these pages. Both systems rely on user profiles intended to model their user's interests. By interacting with the readers, the agents learn more about the reader's interests. However Walden's Paths is designed for a different environment, more specifically an educational environment. In this case the paths are used as artifacts that provide guidance and direction to the readers. Thus, Web pages must maintain consistency with regard to the semantic and rhetoric composition of the path.

There are two other systems that, although slightly tangential to the present work, have influenced the design of the Path Manager. These systems are WebWatcher [1], [9] and Letizia [14]. Like URL-Minder and WatzNew, WebWatcher notifies the user whenever specified pages change. In addition, WebWatcher attempts to evaluate "how interesting" a given Web page would be for a given reader and provides navigation suggestions in real time, by annotating or adding links to the page. This approach explores the use of the knowledge embedded in the links and the text around the links in order to infer relevance. While Johnson rejected links for page distance analysis, the arguments forwarded by WebWatcher prompted the consideration of them as a metric in the Path Manager.

Letizia is an autonomous interface agent [15] that aids a reader in browsing the Web. Letizia uses a knowledge-based approach to respond to personal interests. Like WebWatcher, it also attempts to evaluate how interesting a given Web page would be for a given reader. While the reader is reading a Web page in Netscape, Letizia traverses the links in the page, retrieves the pages and analyzes them. Then it ranks the links based on how interesting the destination Web pages might be for the reader. Letizia's inference takes place on the client as opposed to WebWatcher where the process is conducted on the server.

3. APPROACH

As described, determining the relevance of changes to a document can only be done in the context of how that document is to be used. As this knowledge will not be available to the system supporting relevance assessment, our goal is to provide a variety of information about changes in a relatively concise interface. To do so, we use a variety of document signatures to recognize different types of change.

3.1 Kinds of Change

In order to infer change relevance, it is important to classify the nature of the change. There are four categories of change that we distinguish between: content or semantic, presentation, structural, and behavioral.

Content changes refer to modifications of the page contents from the reader's point of view. For example, a page created for a soccer tournament might be continuously updated as the tournament progresses. After the tournament has ended, the page might change to a presentation about the tournament results and sports injuries.

Presentation changes are changes related to the document representation that do not reflect changes in the topic presented in the document. For instance, changes to HTML tags can modify the appearance of a Web page while it otherwise remains the same.

Structural changes refer to the underlying connection of the document to other documents. As an example, consider changes in the link destinations of a "Weekly Hot Links" Web page. While this page might be conceptually the same, the fact that the destination of the links have changed might be relevant, even if the text of the links has not. Structural changes are also important to detect, as they often might not be visually perceptible.

Behavioral changes refer to modifications to the active components of a document. For Web pages this includes scripts, plug-ins and applets. The consequences of these changes are harder to predict, especially since many pages hide the script code in other files.

3.2 Document Signatures

To represent the different characteristics of a Web-based document, we use a set of document signatures to infer and compute similarities between two Web pages. In the context of this paper, the term of "document signatures" is not equivalent to "signature files" (as in Witten, et al. [24]), which generally refers to strings of bits created by hashing all the terms in the document. In the present work, the signature approach relies on identifying page features and characteristics that not only identify the Web page, but also allow quantifying the change magnitude based on a comparison with a previously computed signature. As of now, the approach considers four Web page features:

Paragraph Checksums. This metric is used to determine content changes. By recording a checksum for each paragraph, this approach has a finer granularity than is possible with a page checksum. While they provide an idea about the degree of change to the page content, they also provide an idea of which pieces of text changed.

Headings. This metric is used to determine content and presentation changes. Headings typically highlight important text and titles. Changes to headings may indicate changes to the focus or perspective of a page. However, they may also reflect on how the document is divided and the information grouped and presented to the reader. Thus they provide clues about presentation changes.

Links. This metric is used to determine structural changes of the Web page. Since the value of hypertext documents depends not only on the document's contents, but also on the navigation provided by its links, it is important to analyze this connectivity. There are two components to links. On one hand there is an invisible component, namely the documents accessible through the links. On the other hand the visual text or image of the link anchors provides information about the contents of the destination pages. While the page might appear visually the same, the links might have changed to point to different places rendering the page inappropriate.

Keywords. We use reader-provided keywords in order to determine content changes of the Web page. In the current version, keyword presence is used as the feature to be identified. In future versions the effectiveness of more complex techniques such as TFIDF (Term Frequency Inverse Document Frequency) could be explored to determine the degree of change. Additionally, we are considering (but have not yet implemented) evaluating different algorithms for automatic keyword generation. Keywords provided by users are good at distinguishing relevant pages from non-relevant pages [16]. The question remains as to whether users will provide them.

The approach also records a *global checksum* for the whole page in order to diminish the possibility of false negatives. There are some changes to Web pages that do not affect any of the previous metrics. For example, a change to an image source would not be reflected in any of the current metrics as it is not part of the text, headings, links or keywords of the document. In this case the global checksum provides a last resort to point out changes.

Combining the results of the various document signatures into an single metric of change is difficult. As previously mentioned, some Web page changes are easily perceived while others are not. Some changes might alter the visual appearance of a page while maintaining the same structure and content. Other changes might change the links while maintaining exactly the same appearance. As already acknowledged, the relevance of change is situation dependent and no single metric will match all situations. We will return to the particular algorithms explored to compute this overall notion of change after a description of the system and interface being developed.

4. THE SYSTEM

The Walden's Paths Path Manager is a system implemented in Java capable of checking a list of Web pages for relevant changes. It takes a path file as input and checks all the pages specified in the path. Alternatively, an HTML file can be specified as input, whether it is located locally (like a bookmark list) or remotely (specified by a URL). The Path Manager interprets links in the HTML file as a list of URLs to check. In order to detect and assess the possible changes, the system retrieves all the pages from the Web, then parses their contents and creates a new signature for each page. The page signatures obtained are compared with the previously computed signatures. Finally, based on the comparison results, the system presents users with an assessment of the relevance of the overall change of each Web page. The user (the path's administrator) can then review each page individually and if there are no relevant changes, the user can validate the current state of the pages.

In order to compute the magnitude of the change, or the distance between two documents, a comparison is made between the signatures of the current version and those previously stored. Currently the Path Manager records three versions of the signature—the original, the last valid, and the latest time the path was checked. The original signature is the first signature obtained. It is used to give a sense of the total amount of change since the page was selected. The last valid signature corresponds to the last time that the user reviewed the changes and validated the pages. As time passes the user might update the valid signature as the pages change. This signature represents functional state of the page. Finally, the latest signature corresponds to the last time that the Path Manager retrieved and analyzed the pages, regardless if the user validated their state or not. Using these three signatures the Path Manager can indicate the amount of recent and long-term change for a page.

4.1 Scenario of Use

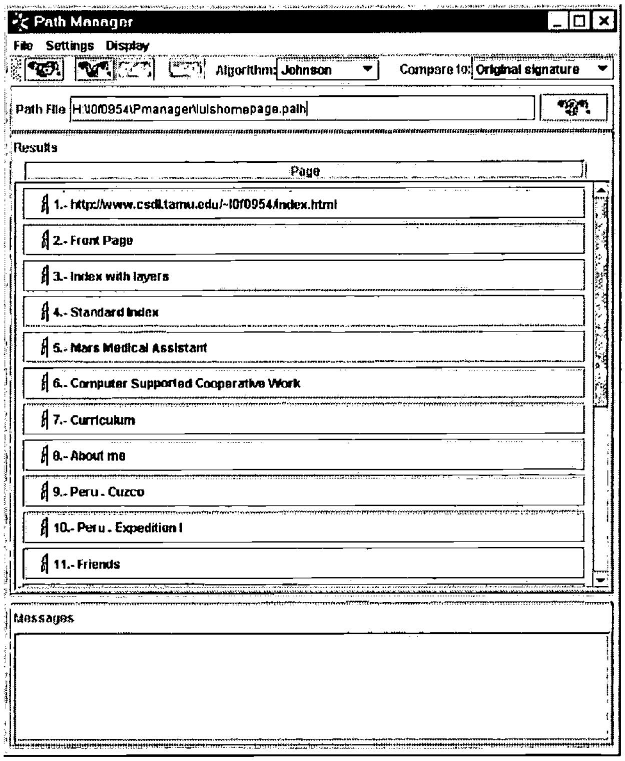
Figure 1 shows the initial state of the interface just after selecting a path to check.

The Path Manger extracts the pages from the path file and presents them to the user. Pages are identified by either their title or, when the page title is not available, their URL. Colors are used to represent the magnitude of the change. In this case pages are shown in blue, meaning that the current state of the page is unknown and needs to be checked. The flag at the left is used to indicate if any change has occurred. Since at this point that is still unknown, the flag is shown as hanging.

At this point the user can specify what algorithm and signature to use in order to check the pages in the list. There are two algorithms implemented, Johnson's algorithm and the proportional algorithm. As for the signatures, the Path Manager maintains three signatures: the original signature, the valid signature, and the latest signature.

Figure 2 shows the Path Manager relevance assessment of the changes in the pages using the proportional algorithm and comparing the signature with the last valid signature.

For each page a red flag is shown fluttering whenever the global checksum has changed. If the global checksum is the same, then a blue hanging flag is shown next to the page title. The flags work as a boolean detection of even the slightest changes. In turn, the degree of relevance of the change is encoded by the color of the page title or URL. In particular, black means that there is no relevant change based on the algorithm chosen. Green, yellow, and red text indicate low, medium, and high degrees of changes considered relevant by the algorithm. The coloring scheme is based on user-defined trigger levels, which in turn could be used for other purposes such as not showing in the path a page that has changed more than the medium level.



Algorithm: Johnson

Compare to: Original signature

Path File H:\I0f0954\Pmanager\luishomepage.path

1.- http://www.csdl.tamu.edu/~l0f0954/index.html

2.- Front Page

3.- Index with layers

4.- Standard Index

5.- Mars Medical Assistant

6.- Computer Supported Cooperative Work

7.- Curriculum

8.- About me

9.- Peru - Cuzco

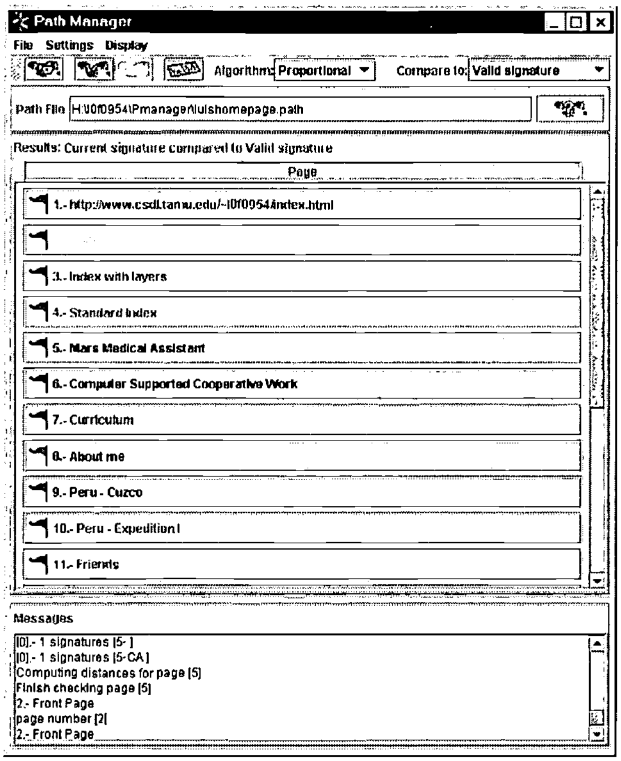
10.- Peru - Expedition I

11.- Friends

Figure 1. Initial State of the Interface.

At this point the user might choose to validate the state of the pages by clicking on the red "Valid" button. Alternatively the user might wonder about what kinds of changes prompt the system to assess the overall relevance of the changes. In order to support the more inquisitive user, the Path Manager can display the amount of change to the particular document signatures used in the relevance assessment. Figure 3 shows the view of the different metrics used in the relevance assessment.

The specific change metrics are presented to the right of the page identification. In addition, the user can get more information on each Web page by selecting the page identification. Figure 4 shows the detailed view of the change metrics for a particular page. In this case the bottom page from Figure 3 was chosen (page 11 in the figure). This view provides the assessment of change to the page at the bottom and information about the use of the page above. For the user to be able to assess the relevance of the change, they may need to see both the content of the page and the context of its use. The top of the detailed view for a page presents properties of the page within the path, such as the cache strategy, annotations and visibility of the page.



Algorithm: Proportional

Compare to: Valid signature

Path File H:\l0f0954\Pmanager\luishomepage.path

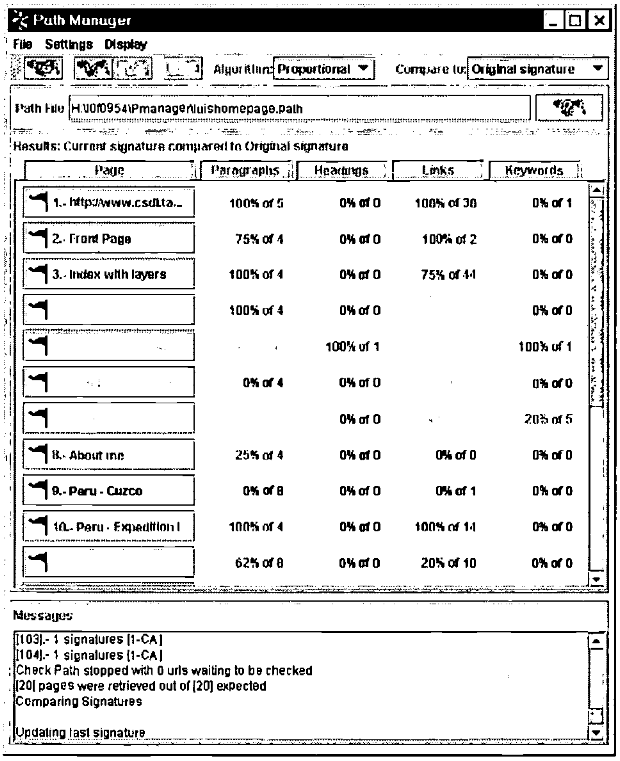
Results: Current signature compared to Valid signature

Page
1.- http://www.csdl.tamu.edu/~l0f0954/index.html
3.- Index with layers
4.- Standard Index
5.- Mars Medical Assistant
6.- Computer Supported Cooperative Work
7.- Curriculum
8.- About me
9.- Peru - Cuzco
10.- Peru - Expedition I
11.- Friends

Messages

```
[0].- 1 signatures [5- ]
[0].- 1 signatures [5-CA]
Computing distances for page [5]
Finish checking page [5]
2.- Front Page
page number [2[
2.- Front Page
```

Figure 2. Overall Change Relevance Assessments.



Algorithm: Proportional

Compare to: Original signature

Path File H:\l0f0954\Pmanager\luishomepage.path

Results: Current signature compared to Original signature

Page	Paragraphs	Headings	Links	Keywords
1.- http://www.csdl.ta...	100% of 5	0% of 0	100% of 30	0% of 1
2.- Front Page	75% of 4	0% of 0	100% of 2	0% of 0
3.- Index with layers	100% of 4	0% of 0	75% of 44	0% of 0
	100% of 4	0% of 0		0% of 0
		100% of 1		100% of 1
	0% of 4	0% of 0		0% of 0
		0% of 0		20% of 5
8.- About me	25% of 4	0% of 0	0% of 0	0% of 0
9.- Peru - Cuzco	0% of 8	0% of 0	0% of 1	0% of 0
10.- Peru - Expedition I	100% of 4	0% of 0	100% of 14	0% of 0
	62% of 8	0% of 0	20% of 10	0% of 0

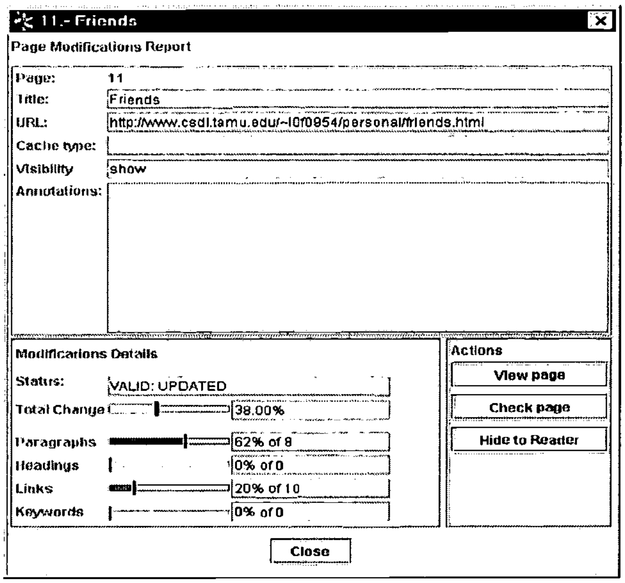
Messages

```
[103].- 1 signatures [1-CA]
[104].- 1 signatures [1-CA]
Check Path stopped with 0 urls waiting to be checked
[20[ pages were retrieved out of [20] expected
Comparing Signatures

Updating last signature
```

Figure 3. View of Change Metrics.

The system will also display the page in a Web browser at the user's request. At this point the user, who might not be the same person who authored the path, might judge the page inappropriate and choose to hide the page from the viewers. While this action prevents the material from being shown to readers of the path, it does not delete the page from the path, leaving that decision to the path author. In case there have been connection or retrieval problems, the manager can prompt the system to check this page again.



Page Modifications Report

Page: 11

Title: Friends

URL: http://www.csdl.tamu.edu/~l0f0954/personal/friends.html

Cache type:

Visibility show

Annotations:

Modifications Details

Status: VALID: UPDATED

Total Change 38.00%

Paragraphs 62% of 8

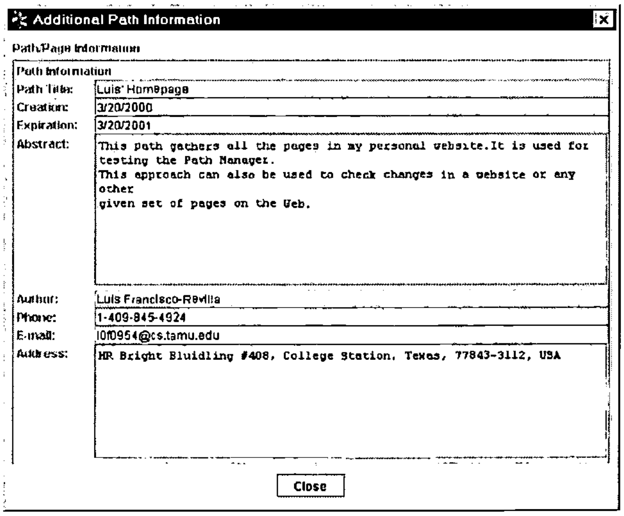
Headings 0% of 0

Links 20% of 10

Keywords 0% of 0

Figure 4. Detailed View of Page Metrics.

In addition the user asks for metadata about the path as a whole by selecting the information button, located to the right of the path file in the main interface (Figures 1-3). Figure 5 shows the path metadata presented to the user.



Path/Page Information

Path Information

Path Title: Luis' Homepage

Creation: 3/20/2000

Expiration: 3/20/2001

Abstract:

```
This path gathers all the pages in my personal website.It is used for
testing the Path Manager.
This approach can also be used to check changes in a website or any
other
given set of pages on the Web.
```

Author: Luis Francisco-Revilla

Phone: 1-409-845-4924

E-mail: l0f0954@cs.tamu.edu

Address: HR Bright Bluidling #408, College Station, Texas, 77843-3112, USA

Figure 5. Path information.

Given this information, the user of the Path Manager can contact the path author in order to inform him/her of the changes or to perform and coordinate corrective measures in order to re-establish the desired function of the path.

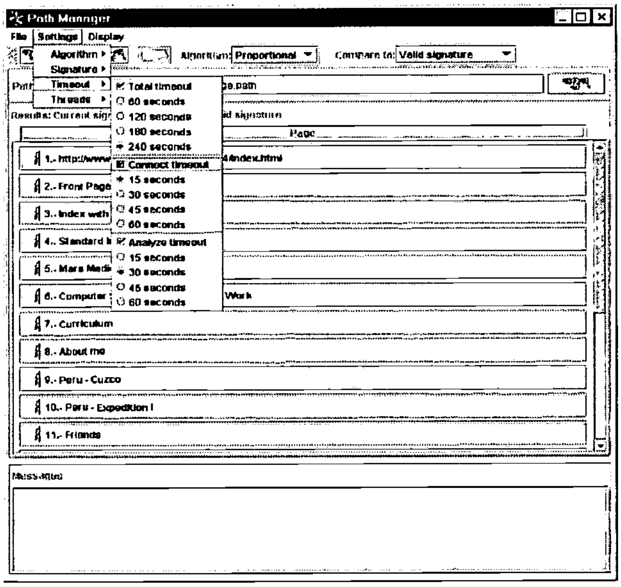
4.2 Web Page Retrieval and Connectivity

An issue for the Path Manager is the connectivity required to retrieve a set of Web pages. Connection times are never constant, as they depend on many variables out of the system or users' control. Some pages are returned rather quickly, while others take longer to return. Also varying connection times means that, sometimes when a Web page seems to take too long to load, canceling the retrieval and immediately restarting the process results in the Web page being loaded faster.

Were the Web pages in a path to be checked sequentially, retrieval problems of a single page would block the retrieval of all consequent pages. Therefore the Path Manager architecture has been designed as a multi-threaded process in order to avoid possible blocks and expedite retrieval. In this scheme, each page is retrieved and analyzed in an independent thread. The user controls the maximum number of simultaneous threads.

Even using independent threads, there are many possible problems when retrieving pages from the Web such as no response, slow connections, and very long pages. In order to deal with these situations the system recognizes three states within each individual retrieval thread:

1. *Connection state*: the system is attempting to contact the Web server hosting the specified Web page.
2. *Retrieval state*: during this state the Web page has been located and its contents are being downloaded.
3. *Analysis state*: all contents have been retrieved and now the system is parsing and analyzing the contents.



Algorithm: Proportional

Compare to: Valid signature

Results: Current sig... ...id signature

1.- http://www... ...4/index.html
2.- Front Page
3.- Index with
4.- Standard h
5.- Mars Medi
6.- Computer ... Work
7.- Curriculum
8.- About me
9.- Peru - Cuzco
10.- Peru - Expedition I
11.- Friends

Figure 6. Selecting the timeouts.

The user can set different timeouts for the connection and retrieval states, which every thread must finish before they expire. The whole path checking process can also be interrupted by a general timeout or by the user clicking the stop button. Figure 6 shows the selection of the timeouts.

Each thread evaluates one page at a time. Once it either successfully or unsuccessfully evaluates the change to that page, the thread will pick another from the set to be checked until all pages have been successfully checked or a general timeout has occurred.

In any case that the Path Manager cannot assess the relevance of the changes, the page titles are shown in blue. Four different shades of blue are used to denote different reasons that the relevance assessment was not successful:

- The page is yet to be checked, i.e., no connection has been attempted.
- There was a network problem, typically a timeout during the connection or retrieval states.
- The general timeout expired before the analysis was completed.
- There were problems that could not be identified.

Because long paths encounter temporary connectivity problems that are resolved quickly, the user can also tell the Path Manager to check only those pages where the assessment of change was not successful.

4.3 Algorithms

We are investigating two different algorithms for categorizing and combining the changes found by the document signatures: one based on Johnson's work [10], and another we call the proportional algorithm.

4.3.1 Johnson's Algorithm

We first implemented a variation on Johnson's algorithm [10], [11] to compute the distance between two documents. As previously mentioned, Johnson only used paragraphs, headings and keywords. In order to take into account structural changes, in our implementation of his approach, we included links as an additional metric. For each signature, additions, deletions and modifications are identified. The distance metric is computed as follows.

$$D = PD + HD + LD + KD$$

Where:

D	– Distance
PD	– Paragraph Distance
HD	– Headings Distance
LD	– Links Distance
KD	– Keywords Distance

In turn

$$PD=\frac{PW * [(\#Pmod*PWmod)+(\#Padd*PWadd)+(\#Pdelete*PWdelete)]}{\#P}$$

Where:

#P	– Number of Paragraphs
#Pmod	– Number of Paragraph Modifications
#Padd	– Number of Paragraph Additions
#Pdelete	– Number of Paragraph Deletions
PW	– Paragraphs Weight
PWmod	– Paragraph Modifications Weight
PWadd	– Paragraph Additions Weight
PWdelete	– Paragraph Deletions Weight

Similarly:

$$HD=\frac{HW*[(\#Hmod*HWmod)+(\#Hadd*HWadd)+(\#Hdelete*HWdelete)]}{\#H}$$

$$LD=\frac{LW*[(\#Lmod*LWmod)+(\#Ladd*LWadd)+(\#Ldelete*LWdelete)]}{\#L}$$

The keyword computation is slightly simpler since it is only important to detect if a keyword is missing.

$$KD = \frac{KW * [\ Kdelete * KWdelete\ ]}{\#K}$$

Johnson developed his algorithm to support Web-based tutorials. In his application, the results of the algorithm were used by the system in deciding whether a page should be displayed to the student—i.e., whether it continued to resemble the page selected by the tutorial's author. Johnson's algorithm includes the ability to weight modifications, additions, and deletions independently. However, since it is supporting a computer system and not a human user, it does not provide normalized results; this makes the results more difficult for a user to interpret. In particular, this makes the selection of cutoff values and the visualization of change somewhat unintuitive.

4.3.2 Proportional Algorithm

This signature-based approach provides a simpler computation of the distance. The change to each individual signature is computed as follows:

$$\begin{aligned}
PD &= (\#Pchanges / \#P) * 100 \\
\#Pchanges &= \#Pmods + \#Padds + \#Pdeletes \\
HD &= (\#Hchanges / \#H) * 100 \\
\#Hchanges &= \#Hmods + \#Hadds + \#Hdeletes \\
LD &= (\#Lchanges / \#L) * 100 \\
\#Lchanges &= \#Lmods + \#Ladds + \#Ldeletes \\
KD &= (\#Kchanges / \#K) * 100 \\
\#Kchanges &= \#Kmods + \#Kadds + \#Kdeletes
\end{aligned}$$

The overall page distance is:

$$\begin{aligned}
D &= (\#Tchanges / \#T) * 100 \\
\#T &= \#P + \#H + \#L + \#K \\
\#Tchanges &= \#Pchanges + \#Hchanges + \#Lchanges + \#Kchanges
\end{aligned}$$

Throughout the proportional algorithm, the number of paragraphs, headings, links, or keywords is taken to be the maximum of the two signatures being compared. Thus, whether a page goes from one to four paragraphs or from four paragraphs to one, #P = 4.

The proportional algorithm provides a normalized and symmetric distance that is easier to use for different sets of Web pages. The normalized and symmetric properties of the distance measurement facilitate providing the user with a visualization or listing of the different changes. This, in turn, allows the user to more effectively evaluate the page changes without having to actually review all the pages in the path.

5. THE PERCEPTION OF CHANGE

In order to inform and evaluate the approach and design of systems that aid in the automatic assessment of change relevance, we conducted a study to observe how humans perceive changes of Web pages. (We give a brief overview of the study here. For details consult the companion paper [5].)

This study covered three kinds of change: content, structure and presentation. In particular, content changes included modifications to the text presented, structure changes referred to modifications of URLs and the anchor text for links, and presentation changes included modifications to colors, fonts, backgrounds, spatial positioning of information, or combinations of these. Web pages were selected from paths previously created by teachers and from personal bookmarks. Each page was modified to reflect a single type of change, and was classified by the magnitude and the type of change.

5.1 Methodology

The experiment consisted of presenting the participant with two versions of a Web page, the original version and a possibly modified version (some were unmodified copies of the original). The person was then asked to evaluate the magnitude of change.

The goal of the study was to address the following three questions:

1. Do people view the same changes in a different way when given different amounts of time to analyze the pages?
2. What kinds of changes are easily perceived?
3. Of what kind of changes do users want to be notified?

Before the evaluation we familiarized the participants with the testing software and the kinds of change. Additionally we provided a context for evaluating the changes in the Web pages by providing a scenario. The participant was asked to act the role of the Information Facilitator in a K-12 school (Kindergarten to High School). Teachers have chosen pages from the Web to teach their classes and it is the participant's responsibility to check for changes.

The first task in the study was to fill in a pre-evaluation questionnaire collecting demographic data about the participants and their computer literacy. The study was then divided into three phases:

1. In *phase I*, the person was given 60 seconds to view each pair of Web pages. In this phase the system presented the participant 8 cases, which included examples from all of the different kinds of change. The system also identified the changes to the participant. The objective of this phase was to provide training, and the answers to the question allowed us to address the third question.
2. In *phase II*, the person was only allowed to view the page pairs for 15 seconds before evaluating them. There were 31 pairs in this phase. This phase addressed the first and second questions.
3. In *phase III*, the person was given 60 seconds to view the Web pages. There were 31 pages in this phase, which also addressed the first and second questions.

In each phase, once the person viewed a Web page pair, s/he was presented with five questions:

1. From the Content perspective, the degree of change is:
2. From the Structure perspective, the degree of change is:
3. From the Presentation perspective, the degree of change is:
4. Overall, how significant are the changes?
5. If this page were in my bookmark list, I would like to be notified when changes like these occur.

Questions 1-4 were answered in a 7-stop scale ranging from "none" to "moderate" to "drastic". Question 5 was answered in a 7-stop scale ranging from strongly disagree to strongly agree.

The final task in the study was to fill a post-evaluation questionnaire that gathered the participant's general comments about how easy it was to assess the magnitude of changes in Web pages, and about the evaluation software.

In all phases the testing software controlled the presentation of the Web pages directly. The time required to record the participant answer was not controlled. The total time for the study varied from 1 to 2 hours per person, depending on the time the participants took in filling in the answers, and whether they decided to rest between phases.

5.2 Target Population

In order to conduct the study, we recruited adults, specifically students at Texas A&M University. Subjects were divided into two groups. Pages in Phase II for one group were used as the pages in Phase III for the other group. This allows comparing the results for the same page when given 15 or 60 seconds for observation.

5.3 Study Results

The results of Phase I, when the subjects were provided textual descriptions of what had changed, give the clearest indication of what people consider change and of what they would like to be notified. For content changes, as the percent of paragraphs changed, the perception of overall change also increased, as did the subjects' desire to be notified of the change. Interestingly, as the degree of structural changes increased, the perception of the overall change did not increase but the desire to be notified of the change did. For similar percentages of content and structure change, subjects rated the content changes higher in overall change but lower in desire to be notified.

In Phases II and III, time did not alter the perception of content change but was seen to play a large role in the identification of structural change. This is likely due to the visibility of content changes and the invisibility of structure changes—subjects commented in the exit survey about how difficult and time-consuming it was to detect structure change. As with the results of Phase I, subjects desire to be notified of perceived changes in structure were higher than their desire to be notified for similarly perceived changes in content.

In Phases II and III, subjects rated low and medium changes in content similarly, but rated those pages that had drastically changed quite a bit higher for all questions. Structural changes saw a similar but less extreme jump in ratings when almost all the links had been changed.

Presentation changes were not considered by amount of change but by type. Subjects did not seem to notice changes to fonts, while changes to background color and navigational images were noticed but rated low as contributing to overall change or desire to be notified. Changes in the arrangement of material on a page was noticed by subjects but interpreted differently depending on the amount of time they had to look at the page. With only 15 seconds, the rearrangement was considered a content change and there was a large desire to be notified while with 60 seconds the subjects recognized that the material was the same but moved and they had a lower desire to be notified. Finally, when many presentation features were changed at once, subjects rated the change as high on all metrics and wanted to be notified.

5.4 Implications for the Path Manager

The results of the study indicate that including links in metrics for overall change will better match our subject's evaluation of change and desire to be notified. The results also show that presentation changes were viewed as largely unrelated to the function of the page, except in extreme cases. Finally, a more detailed analysis of the results may provide weightings for combining the results of the four signatures into an overall view of change.

6. CHALLENGES AND LIMITATIONS

An issue with using document signatures based on HTML tags is that not every page creator or Web page authoring tool uses the HTML tags consistently. In the case of headings, page creators often choose to modify the visual appearance of the text by using tags such as FONT or resorting to images. This imposes the requirement of implementing smarter parsers. We are currently attempting to address some of these by augmenting the system to infer what is conceptually a heading and where paragraphs begin and end, and to identify different types of changes to links.

A limitation of the current Path Manager is that no indirection is supported. When faced with Web pages containing frames, the Path Manager does not check the pages contained in the frames. The same is true for other tags such as images or links. There is no retrieval of the pages specified in the SRC or HREF fields unless they are also included explicitly in the path.

Another limitation of the Path Manager is that it does not monitor any JavaScript or other page behaviors. This is in part due to the complexity of the required parser and to the fact that this remains a moving target, where specifications and support vary constantly over time and browser type.

Finally, while the Path Manager can parse dynamically generated pages returned by CGIs, it does not recognize that they are dynamic and therefore variable by nature. Augmenting the system to recognize these, could enable separate treatment for such pages.

7. CONCLUSIONS

Maintenance of distributed collections of documents remains a challenging and time-consuming task. People must monitor the documents for change and then interpret changes to the documents in the context of the collection's goals.

The Walden's Paths Path Manager supports the maintenance of collections of Web pages by recognizing, evaluating, and informing the user of changes. The evaluation of change is based on document signatures of the paragraphs, headings, links, and keywords. The Path Manager keeps track of original, last valid, and last collected signatures so users can determine both long-term and short-term change depending on their particular concern.

The Path Manager has been designed to work in a distributed environment where connectivity to documents is unpredictable. Its architecture and instantiation provide users with control over system resources consumed. Also, the system provides feedback about documents that are not successfully evaluated.

Particular to uses for Walden's Paths, the Path Manager provides access to information about the use of the documents in a path and to metadata about the path that may be important to determining the relevance of particular changes. Users may also mark pages so that they remain in the path but the Walden's Paths server will not display them to readers of the path.

A study of the perception of changes to Web pages indicated the desire for structural changes to be included in the determination of overall change. The study also showed that presentation changes were largely considered irrelevant. Current work on the Path Manager aims to overcome difficulties with the inconsistencies and indirection found in Web documents.

8. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Numbers IIS-9812040 and DUE-0085798.

9. REFERENCES

[1] Armstrong, R., Freitag, D.M, Jachims, T., & Mitchell T. WebWatcher: A Learning Apprentice for the World Wide Web, in Working Notes of AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments (Stanford University CA, March 1995) AAAI Press, 6-12.

[2] Bush, V. As We May Think. The Atlantic Monthly, (August 1945), 101-108.

[3] Brewington, B. & Cybenko, G. How Dynamic is the Web, in Proc. of WWW9—9th International World Wide Web Conference, IW3C2, (2000) 264-296.

[4] Douglis, F., Ball, T., Chen, Y., and Koutsofios, E. The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web, in World Wide Web, 1(1) 27-44 (January 1998).

[5] Francisco-Revilla, L., Shipman, F.M., Karadkar, U., Furuta, R., and Arora, A. Changes to Web Pages: Perception and Evaluation. To appear in Proc. Hypertext 2001, (Åarhus, Denmark, August 2001).

[6] Furuta, R., Shipman, F., Francisco-Revilla, L., Karadkar, U., & Hu, S. Ephemeral Paths on the WWW: The Walden's Paths Lightweight Path Mechanism. WebNet (1999), 409-414.

[7] Furuta, R., Shipman, F., Marshall, C., Brenner, .D., & Hsieh, H. Hypertext Paths and the World-Wide Web: Experiences with Walden's Paths, in Proc. of Hypertext'97 (Southampton U.K., April 1997). ACM Press, 167-176.

[8] Giles, L., Bollacker, K., & Lawrence, L. CiteSeer: An Automatic Citation Indexing System, in Proc. of DL'98 (Pittsburgh PA, June 1998). ACM Press, 89-98.

[9] Joachims, T., Mitchell, T., Freitag, D., & Armstrong, R. WebWatcher: Machine Learning and Hypertext, Fachgruppenterffen Maschinelles Lernen. Dortmund, Germany, August 1995.

[10] Johnson, D.B. Enabling the Reuse of World Wide Web Documents in Tutorials. PhD. Dissertation, Dept. of computer Science and Engineering. University of Washington, Seattle, WA. 1997

[11] Johnson, D.B., & Tanimoto, S.L. Reusing Web Documents in Tutorials with the Current-Documents Assumption: Automatic Validation of Updates, in Proc. of EDMEDIA'99 (Seattle WA, June 1999). AACE, 74-79.

[12] Karadkar, U., Francisco-Revilla, L., Furuta, R., Hsieh, H., & Shipman, F. Evolution of the Walden's Paths Authoring Tools, in Proceedings of WebNet 2000--World Conference on the WWW and Internet (San Antonio, TX, October 30--November 4, 2000) AACE, 299-304.

[13] Levy, D.M. Fixed or Fluid? Document Stability and new Media, in Proc. of the European Conference on Hypertext Technology '94 (Edinburgh Scotland, September 1994). ACM Press, 24-41.

[14] Lieberman, H. Letizia: An Agent That Assists Web Browsing. International Joint Conference on Artificial Intelligence (Montreal Canada, August 1995). Morgan Kaufman, 924-929.

[15] Lieberman, H. Autonomous Interface Agents, in Proc. of CHI'97 (Atlanta GA, March 1997). ACM Press, 67-74.

[16] Pazzani, M., & Billsus, D. Learning and Revising Reader Profiles: The Identification of Interesting Web Sites. Machine Learning 27 (1997), Kluwer Academic Publishers, 313-331.

[17] Pazzani, M., Muramatsu, J., & Billsus, D. Syskill and Webert: Identifying interesting Web sites, in Proc. of AAAI'96 (Portland Oregon, August 1996). American Association for Artificial Intelligence, 54-59

[18] Shipman, F., Furuta, R., Brenner, .D., Chung, C., & Hsieh, H. Using Paths in the Classroom: Experiences and Adaptations, in Proc. Hypertext'98 (Pittsburgh PA, June 1998). ACM Press, 167-176.

[19] Shipman, F., Marshall, C., Furuta, R., Brenner, .D., Hsieh, H., & Kumar, V. Creating Educational Guided Paths over the World-Wide Web, in Proc. of ED-TELECOM'96 (Boston MA, June 1996), AACE, 326-331.

[20] Starr, B., Ackerman, M.S., & Pazzani, M. Do-I-Care: a collaborative Web agent, in Proc. of CHI'96, (Vancouver Canada, April 1996), ACM Press, 273-274.

[21] Starr, B., Ackerman, M.S., & Pazzani, M. Do-I-Care: Tell Me What's Changed on the Web, in Proc. of the AAAI Spring Symposium on Machine Learning in Information Access (Stanford CA, March 1996).

[22] URL-Minder. Available at http://www.netmind.com/html/url-minder.html

[23] WatzNew. Available at http://www.watznew.com

[24] Witten, I.H., Moffat, A., and Bell, T.C., Managing Gygabytes. Compressing and Indexing Documents and Images, 2nd Edition, Morgan Kaufman, San Francisco, CA, 1999.

Measuring the Reputation of Web Sites: A Preliminary Exploration

Greg Keast, Elaine G. Toms, Joan Cherry
Faculty of Information Studies
University of Toronto
Toronto, Ontario, Canada
gkeast@ica.net, { toms, cherry }@fis.utoronto.ca

ABSTRACT

We describe the preliminary results from a pilot study, which assessed the perceived reputation – authority and trustworthiness – of the output from five WWW indexing/ranking tools. The tools are based on three techniques: external link structures, internal content, or human selection/indexing. Twenty-two participants reviewed the output from each tool and assessed the reputation of the retrieved sites.

Categories and Subject Descriptors

H.3.3 Information Search and Retrieval.

General Terms

Measurement, Performance, Reliability, Experimentation

Keywords

Web sites, Reputation, Authority, Evaluation, TOPIC, Google, Alta Vista, Lycos, Yahoo

1. INTRODUCTION

Multiple techniques have been developed to retrieve and rank web sites on the World Wide Web. To date the evaluation of results has been based primarily on the relevance and 'aboutness' of a site to the query. Equally valuable to the user is the perceived reputation or trustworthiness of the content. TOPIC, developed by Rafei and Mendelzon [4], identifies the topics for which a web page has a good reputation by mining the link structure (see details in [4]). In this pilot study we compare the output from TOPIC with that of several tools to ascertain if certain types of tools are more conducive to providing 'reputable' web sites.

The five tools tested were created from three techniques: those that use the external link structure to make sense of the site; those that use the content of a site as the input to the indexing and ranking; and, those that are human-selected and indexed. In this pilot study, we compare how users perceive the output from each tool. Do certain types of tools yield sites that are perceived to be more reputable -- authoritative and trustworthy – than others?

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

2. BACKGROUND

Reputation is admittedly an elusive construct. The Oxford dictionary defines it as "the general estimation in which a person is held." In essence, it is an external perception about an object (deserved or otherwise). An analogy may be made to web sites. If a site links to another site, then it serves as a recommendation or endorsement of that site. However, the mere fact that a link exists is insufficient to claim an endorsement and thus, a reputation. Links to sites, like citations to journals, can be created for many reasons. But when the site is not only 'well-linked' but also the structure can be analyzed to such an extent that the evidence is overwhelming, then one could conclude that a reputation has been made. TOPIC is built on this foundation.

Source authority is generally considered a key criterion for judging the perceived quality of information and for filtering information [7]. Rieh and Belkin [5] showed that "people depend upon such judgments of source authority and credibility more in the Web environment than in the print environment". The importance of reputation to a company or organization is amply illustrated by the proliferation of consulting firms that provide reputation management services. These distinctive activities illustrate the importance of reputation as an influential factor for clients when choosing institutions, services or products [6].

Kleinberg [3], Chakrabarti et al [2] and Brin & Page [1] have exploited the link structure to confer authority on web sites. TOPIC augments this work by identifying categories in which a page has a good reputation.

3. METHODS

3.1 Participants

Twenty-two participants (13 female, 8 male and 1 unknown) participated in the study. Twenty per cent were under 26 and 30% were over 35. They were a well-educated group, 95% had at least one undergraduate degree. Forty per cent browse the web more than 15 times a week with the remainder browsing the web fewer times. Due to the nature of the task that participants were to be assigned, we also noted that 65% watch one to four movies a month; 10% watched more than 15. All were recruited from within the University of Toronto and were paid $10.00 for their participation.

3.2 Task

Initially, we assessed the results to a query from five different tools, which are based on three types of techniques: 1) Google and TOPIC which use intra-site link analysis, 2) Altavista and Lycos which use automatic indexing, and 3) Yahoo which uses human-assigned categories and assessments. The five search tools were accessed by the first author during the same two-hour period to search for "movie review" sites. We selected movie reviews because it has wide appeal and we knew that we would find likely participants who had some experience in examining web sites on that topic and were thus sufficiently knowledgeable to comment on a site's reputation. The first four sites retrieved by each tool were chosen and a final list of 17 unique sites on movie reviews was compiled. Duplicates were removed, but the source of each duplicate was included in the subsequent analysis.

3.3 Procedures

Testing took place primarily in a lab setting over a two-week period. Each participant was assigned a randomized list of the 17 sites. They visited each in turn and assessed on a scale of one to five the trustworthiness, authoritativeness and aboutness of the site and also of the links leading from the site. In addition, they indicated if they would return to the site and/or recommend the site as a good source of movie reviews to a friend or colleague. Finally, they indicated which of the 17 sites they considered the best source for movie reviews. The entire process took about one hour per participant.

4. RESULTS

Some sites were not available at the time of testing and thus the number of cases in each analysis varied from 17 to 22. All data were analyzed using SPSS's GLM repeated measures. Perception measures were averaged (see Table 1) by participant. Selected data are presented here.

The sites found by the five tools differed significantly in terms of perceived Trustworthiness (F(4,72)=3.314, p=.039), Authoritativeness (F(4,72)=4.165, p=.02) and Aboutness (F(4,72)=11.353, p>.001). In post hoc tests, the sites found by Topic were considered more trustworthy and authoritative than those of Google and Lycos, while Google was less authoritative than Alta Vista. All five differed significantly in Aboutness. Yahoo sites were considered to be more on topic than those of Alta Vista and Topic.

Table 1. Average Rating for Sites by Tool

Tools	Trust	Authority	Aboutness
Alta Vista	3.7	3.7	4.2
Google	3.3	3.3	3.4
Lycos	3.2	3.1	3.7
Topic	3.7	3.8	4.0
Yahoo	3.6	3.6	4.5

Outgoing links from each of the sites were perceived similarly. But there were notable exceptions. There were no significant differences among the five tools according to Trustworthiness and Authoritativeness of outgoing links. Sites retrieved by Google performed the most poorly on Aboutness, while those retrieved by Alta Vusta, Topic and Yahoo were rated the highest.

5. DISCUSSION & CONCLUSIONS

This reports on results from one test with a single topic and thus results are tentative, although noteworthy. (Because of the human effort required in the conduct of this study, only one topic, i.e., subject matter of site, was selected for the pilot study.) TOPIC performed on average as well as Altavista and outperformed Google. Surprisingly, it matched or outperformed the human-selected-and-indexed Yahoo. In general the two non-human techniques performed on par with the human-generated tool.

Overall the scores were relatively low with 3.5 being the average rating assigned on almost all variables. Perhaps this is indicative of the status of movie review sites. Of the 17 sites that were retrieved from these five tools, less than a third were considered worth recommending or returning to or identified as the best sources of movie reviews.

Predicting the reputation of a web site is a difficult problem. TOPIC's approach to measuring reputation is novel. Findings from this work are promising, justifying the need for further assessment using multiple topics from diverse domains with different users.

6. ACKNOWLEDGMENTS

Our thanks to the Faculty of Information Studies for the use of its facilities, to the 22 participants and to Alberto Mendelzon.

7. REFERENCES

[1] Brin, S. & Page, L. (1998). *The anatomy of a large-scale hypertextual web search engine.* http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm

[2] Chakrabarti, S., van den Berg, M., & Dom, B. (1999). Focused crawling: a new approach to topic specific resource discovery. http://www.cs.berkeley.edu/~soumen/doc/www1999f/html/

[3] Kleinberg, J. M. *Authoritative sources in a hyperlinked environment. ACM*, 46 (5 1998), 604-632.)

[4] Rafiei, D. & Mendelzon, A. O. (2000). *What is this page known for? Computing Web page reputations.* Toronto, Ont.: University of Toronto. Retrieved September 6, 2000, from the World Wide Web: ftp://ftp.db.toronto.edu/pub/papers/www9.ps.gz

[5] Rieh, S. Y. & Belkin, N. J. (1998). Understanding judgment of information quality and cognitive authority in the WWW. *Proceedings of the ASIS Annual Meeting*, 35 (1998), 279-289.

[6] Tapp, L. (2000). *Reputation is key in picking the best business school.* The Globe and Mail, Sept. 6, 2000

[7] Wilson, P. (1983). Second-Hand Knowledge: an Inquiry into Cognitive Authority. Westport, CT: Greenwood Press.

Personalized Spiders for Web Search and Analysis

Michael Chau
Dept of Management Info. Sys.
The University of Arizona
Tucson, Arizona 85721, USA
1-520-626-9239
mchau@bpa.arizona.edu

Daniel Zeng
Dept of Management Info. Sys.
The University of Arizona
Tucson, Arizona 85721, USA
1-520-621-4614
zeng@bpa.arizona.edu

Hsinchun Chen
Dept of Management Info. Sys.
The University of Arizona
Tucson, Arizona 85721, USA
1-520-621-2748
hchen@bpa.arizona.edu

ABSTRACT

Searching for useful information on the World Wide Web has become increasingly difficult. While Internet search engines have been helping people to search on the web, low recall rate and outdated indexes have become more and more problematic as the web grows. In addition, search tools usually present to the user only a list of search results, failing to provide further personalized analysis which could help users identify useful information and comprehend these results. To alleviate these problems, we propose a client-based architecture that incorporates noun phrasing and self-organizing map techniques. Two systems, namely CI Spider and Meta Spider, have been built based on this architecture. User evaluation studies have been conducted and the findings suggest that the proposed architecture can effectively facilitate web search and analysis.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, information filtering, search process.*

General Terms

Design, Experimentation.

Keywords

Information retrieval, Internet spider, Internet searching and browsing, noun-phrasing, self-organizing map, personalization, user evaluation.

1. INTRODUCTION

The World Wide Web has become the biggest digital library available, with more than 1 billion unique indexable web pages [9]. However, it has become increasingly difficult to search for useful information on it, due to its dynamic, unstructured nature and its fast growth rate. Although development of web search and analysis tools such as search engines has alleviated the problem to a great extent, exponential growth of the web is making it impossible to collect and index all the web pages and refresh the index frequently enough to keep it up-to-date. Most search engines present search results to users that are incomplete and outdated, usually leaving users confused and frustrated.

A second problem that Internet users encounter is the difficulty in searching information on a particular website, e.g., looking for information related to a certain topic in the website www.phoenix.com. Among the popular commercial search engines, only a few offer the search option to limit a search session to a specified website. Because most search engines only index a certain portion of each website, the recall rate of these searches is very low, and sometimes even no documents are returned. Although most large websites nowadays have their built-in internal search engines, these engines index the information based on different schemes and policies and users may have difficulty in uncovering useful information. In addition, most of the websites on the Internet are small sites that do not have an internal search feature.

A third problem is the poor retrieval rate when only a single search engine is used. It has been estimated that none of the search engines available indexes more than 16% of the total web that could be indexed [12]. Even worse, each search engine maintains its own searching and ranking algorithm as well as query formation and freshness standard. Unless the different features of each search engine are known, searches will be inefficient and ineffective. From the user's point of view, dealing with an array of different interfaces and understanding the idiosyncrasies of each search engine is too burdensome. The development of meta-search engines has alleviated this problem. However, how the different results are combined and presented to the user greatly affects the effectiveness of these tools.

In addition, given the huge number of daily hits, most search engines are not able to provide enough computational power to satisfy each user's information need. Analysis of search results, such as verifying that the web pages retrieved still exist or clustering of web pages into different categories, are not available in most search engines. Search results are usually presented in a ranked list fashion; users cannot get a whole picture of what the web pages are about until they click on every page and read the contents. This can be time-consuming and frustrating in a dynamic, fast-changing electronic information environment.

In order to alleviate the above problems, we propose a personalized and integrated approach to web search. In this paper, we present a client-side web search tool that applies various artificial intelligence techniques. We believe that a search tool that is more customizable would help users locate useful

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

information on the web more effectively. The client-based architecture also allows for greater computation power and resources to provide better searching and analysis performance. We have conducted two experiments to evaluate the performance of different prototypes built according to this architecture.

2. RELATED WORK

In order to address the information overload problem on the web, research has been conducted in developing techniques and tools to analyze, categorize and visualize large collections of web pages, among other text documents. A variety of tools have been developed to assist searching, gathering, monitoring and analyzing information on the Internet.

2.1 Web Search Engines and Spiders

Many different search engines are available on the Internet. Each has its own characteristics and employs its preferred algorithm in indexing, ranking and visualizing web documents. For example, AltaVista (www.altavista.com) and Google (www.google.com) allow users to submit queries and present web pages in a ranked order, while Yahoo! (www.yahoo.com) groups websites into categories, creating a hierarchical directory of a subset of the Internet.

Another type of search engine is comprised of meta-search engines, such as MetaCrawler (www.metacrawler.com) and Dogpile (www.dogpile.com). These search engines connect to multiple search engines and integrate the results returned. As each search engine covers different portion of the Internet, meta-search engines are useful when the user needs to get as much of the Internet as possible. There are also special-purpose topic-specific search engines [4]. For example, BuildingOnline (www.buildlingonline.com) specializes in searching in the building industry domain on the web, and LawCrawler (www.lawcrawler.com) specializes in searching for legal information on the Internet.

Internet spiders (a.k.a. crawlers), have been used as the main program in the backend of most search engines. These are programs that collect Internet pages and explore outgoing links in each page to continue the process. Examples include the World Wide Web Worm [16], the Harvest Information Discovery and Access System [1], and the PageRank-based Crawler [5].

In recent years, many client-side web spiders have been developed. Because the software runs on the client machine, more CPU time and memory can be allocated to the search process and more functionalities are possible. Also, these tools allow users to have more control and personalization options during the search process. For example, Blue Squirrel's WebSeeker (www.bluesquirrel.com) and Copernic 2000 (www.copernic.com) connect with different search engines, monitor web pages for any changes, and schedule automatic search. Focused Crawler [2] locates web pages relevant to a pre-defined set of topics based on example pages provided by the user. In addition, it also analyzes the link structures among the web pages collected.

2.2 Monitoring and Filtering

Because of the fast changing nature of the Internet, different tools have been developed to monitor websites for changes and filter out unwanted information. Push Technology is one of the emerging technologies in this area. The user first needs to specify some areas of interest. The tool will then automatically push related information to the user. Ewatch (www.ewatch.com) is one such example. It monitors information not only from web pages but also from Internet Usenet groups, electronic mailing lists, discussion areas and bulletin boards to look for changes and alert the user.

Another popular technique used for monitoring and filtering employs a software agent, or intelligent agent [15]. Personalized agents can monitor websites and filter information according to particular user needs. Machine learning algorithms, such as an artificial neural network, are usually implemented in agents to learn the user's preferences.

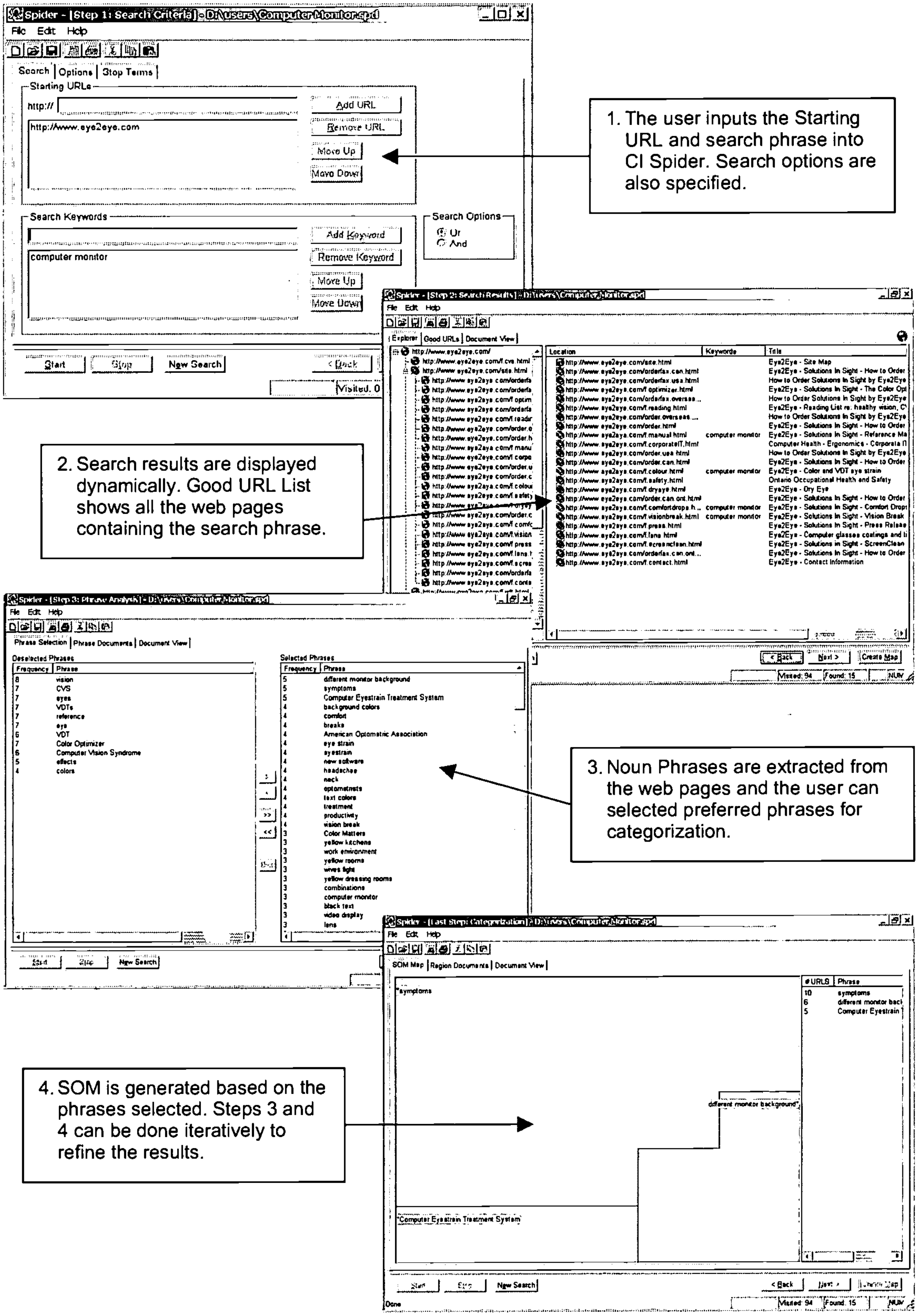
2.3 Indexing and Categorization

There have been many studies in textual information analysis of information retrieval and natural language processing. In order to retrieve documents based on given concepts, the documents have to be indexed. Automatic indexing algorithms have been used widely to extract key concepts from textual data. It having been shown that automatic indexing is as effective as human indexing [18], many proven techniques have been developed. Linguistics approaches such as noun phrasing also have been applied to perform indexing for phrases rather than just words [21]. These techniques are useful in extracting meaningful terms from text documents not only for document retrieval but also for further analysis.

Another type of analysis tool is categorization. These tools allow a user to classify documents into different categories. Some categorization tools facilitate the human categorization process by simply providing a user-friendly interface. Tools that are more powerful categorize documents automatically, allowing users to quickly identify the key topics involved in a large collection of documents [e.g., 8, 17, 23].

In document clustering, there are in general two approaches. In the first approach, documents are categorized based on individual document attributes. An attribute might be the query term's frequency in each document [7, 22]. NorthernLight, a commercial search engine, is another example of this approach. The retrieved documents are organized based on the size, source, topic or author of each document. Other examples include Envision [6] and GRIDL [19].

In the second approach, documents are classified based on inter-document similarities. This approach usually includes some kind of machine learning algorithms. For example, the Self-Organizing Map (SOM) approach classifies documents into different categories which are defined during the process, using neural network algorithm [10]. Based on this algorithm, the SOM technique automatically categorizes documents into different regions based on the similarity of the documents. It produces a data map consisting of different regions, where each region contains similar documents. Regions that are similar are located close to each other. Several systems utilizing this technique have been built [3, 11, 14].



Spider - [Step 1: Search Criteria] - D:\users\Computer Monitor.spd

Starting URLs: http://www.eye2eye.com

Search Keywords: computer monitor

Search Options: ● Or ○ And

1. The user inputs the Starting URL and search phrase into CI Spider. Search options are also specified.

Spider - [Step 2: Search Results] - D:\users\Computer Monitor.spd

Location	Keywords	Title
http://www.eye2eye.com/site.html		Eye2Eye - Site Map
http://www.eye2eye.com/orderfax.can.html		Eye2Eye - Solutions In Sight - How to Order
http://www.eye2eye.com/orderfax.usa.html		How to Order Solutions In Sight by Eye2Eye
http://www.eye2eye.com/f.optimizer.html		Eye2Eye - Solutions In Sight - The Color Opt
http://www.eye2eye.com/orderfax.overseas...		How to Order Solutions In Sight by Eye2Eye
http://www.eye2eye.com/f.reading.html		Eye2Eye - Reading List re: healthy vision, C
http://www.eye2eye.com/order.overseas....		How to Order Solutions In Sight by Eye2Eye
http://www.eye2eye.com/order.html		Eye2Eye - Solutions In Sight - How to Order
http://www.eye2eye.com/f.manual.html	computer monitor	Eye2Eye - Solutions In Sight - Reference Ma
http://www.eye2eye.com/f.corporateIT.html		Computer Health - Ergonomics - Corporate I
http://www.eye2eye.com/order.usa.html		How to Order Solutions In Sight by Eye2Eye
http://www.eye2eye.com/order.can.html		Eye2Eye - Solutions In Sight - How to Order
http://www.eye2eye.com/f.colour.html	computer monitor	Eye2Eye - Color and VDT eye strain
http://www.eye2eye.com/f.safety.html		Ontario Occupational Health and Safety
http://www.eye2eye.com/f.dryeye.html		Eye2Eye - Dry Eye
http://www.eye2eye.com/order.can.ont.html		Eye2Eye - Solutions In Sight - How to Order
http://www.eye2eye.com/f.comfortdrops.h...	computer monitor	Eye2Eye - Solutions In Sight - Comfort Drops
http://www.eye2eye.com/f.visionbreak.html	computer monitor	Eye2Eye - Solutions In Sight - Vision Break
http://www.eye2eye.com/f.press.html		Eye2Eye - Solutions In Sight - Press Release
http://www.eye2eye.com/f.lens.html		Eye2Eye - Computer glasses coatings and li
http://www.eye2eye.com/f.screenclean.html		Eye2Eye - Solutions in Sight - ScreenClean
http://www.eye2eye.com/orderfax.can.ont...		Eye2Eye - Solutions In Sight - How to Order
http://www.eye2eye.com/f.contact.html		Eye2Eye - Contact Information

Visited: 94 Found: 15

2. Search results are displayed dynamically. Good URL List shows all the web pages containing the search phrase.

Spider - [Step 3: Phrase Analysis] - D:\users\Computer Monitor.spd

Deselected Phrases

Frequency	Phrase
8	vision
7	CVS
7	eyes
7	VDTs
7	reference
7	eye
6	VDT
7	Color Optimizer
6	Computer Vision Syndrome
5	effects
4	colors

Selected Phrases

Frequency	Phrase
5	different monitor background
5	symptoms
5	Computer Eyestrain Treatment System
4	background colors
4	comfort
4	breaks
4	American Optometric Association
4	eye strain
4	eyestrain
4	new software
4	headaches
4	neck
4	optometrists
4	text colors
4	treatment
4	productivity
4	vision break
3	Color Matters
3	yellow kitchens
3	work environment
3	yellow rooms
3	wives fight
3	yellow dressing rooms
3	combinations
3	computer monitor
3	black text
3	video display
3	lens

3. Noun Phrases are extracted from the web pages and the user can selected preferred phrases for categorization.

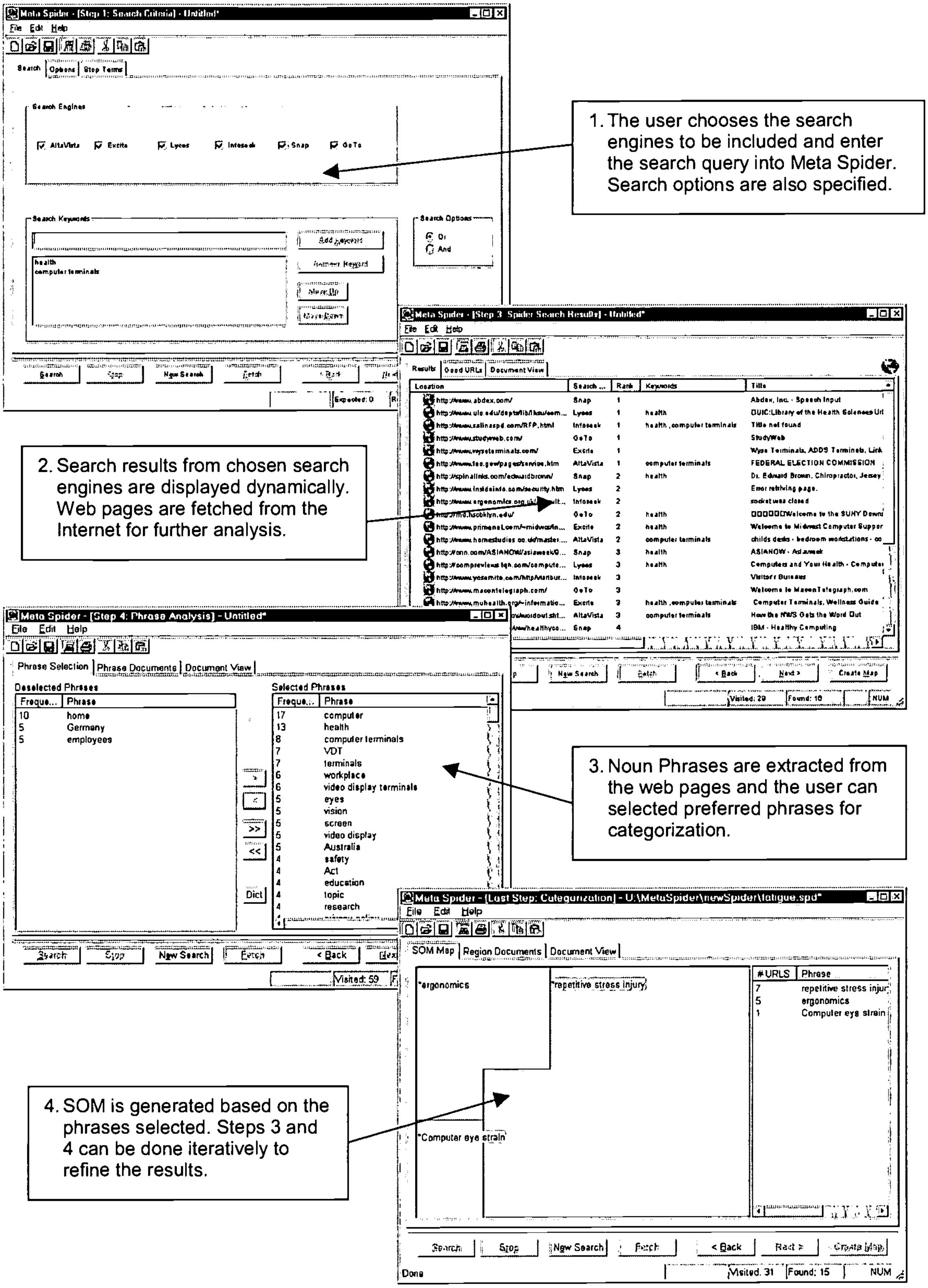
Spider - [Last Step: Categorization] - D:\users\Computer Monitor.spd

#URLS	Phrase
10	symptoms
6	different monitor back
5	Computer Eyestrain

File Edit Help
SOM Map | Region Documents | Document View
"symptoms
different monitor background"
"Computer Eyestrain Treatment System
Start
Stop
New Search
< Back
Next >
Create Map
Done
Visited: 94 Found: 15
NUM

4. SOM is generated based on the phrases selected. Steps 3 and 4 can be done iteratively to refine the results.

Figure 1. Example of a User Session with CI Spider



1. The user chooses the search engines to be included and enter the search query into Meta Spider. Search options are also specified.

Meta Spider - [Step 1: Search Criteria] - Untitled*

Search Engines: ☑ AltaVista ☑ Excite ☑ Lycos ☑ Infoseek ☑ Snap ☑ GoTo

Search Keywords: health, computer terminals

Search Options: ● Or ○ And

2. Search results from chosen search engines are displayed dynamically. Web pages are fetched from the Internet for further analysis.

Meta Spider - [Step 3 Spider Search Results] - Untitled*

Location	Search ...	Rank	Keywords	Title
http://www.abdex.com/	Snap	1		Abdex, Inc. - Speech Input
http://www.uic.edu/depts/lib/lhsu/com...	Lycos	1	health	UIC:Library of the Health Sciences Ur
http://www.salinaspd.com/RFP.html	Infoseek	1	health ,computer terminals	Title not found
http://www.studyweb.com/	GoTo	1		StudyWeb
http://www.wyseterminals.com/	Excite	1		Wyse Terminals, ADDS Terminals, Link
http://www.fec.gov/pages/service.htm	AltaVista	1	computer terminals	FEDERAL ELECTION COMMISSION
http://spinalinks.com/edwardbrown/	Snap	2	health	Dr. Edward Brown, Chiropractor, Jersey
http://www.insideinfo.com/security.htm	Lycos	2		Error retrieving page.
http://www.ergonomics.org.uk/...ult...	Infoseek	2		socket was closed
http://md.hscbklyn.edu/	GoTo	2	health	□□□□□□Welcome to the SUNY Down
http://www.primenet.com/~midwcsfin...	Excite	2	health	Welcome to Midwest Computer Suppor
http://www.homestudies.co.uk/master....	AltaVista	2	computer terminals	childs desks - bedroom workstations - co
http://cnn.com/ASIANOW/asiaweek/0...	Snap	3	health	ASIANOW - Asiaweek
http://compreviews.tgn.com/compute...	Lycos	3	health	Computers and Your Health - Computer
http://www.yosemite.com/http/marbur...	Infoseek	3		Visitors Bureaus
http://www.macontelegraph.com/	GoTo	3		Welcome to MaconTelegraph.com
http://www.muhealth.org/~informatio...	Excite	3	health ,computer terminals	Computer Terminals, Wellness Guide
...ov/wordout.sht...	AltaVista	3	computer terminals	How the NWS Gets the Word Out
.../healthyco...	Snap	4		IBM - Healthy Computing

Visited: 29 Found: 10

3. Noun Phrases are extracted from the web pages and the user can selected preferred phrases for categorization.

Meta Spider - [Step 4: Phrase Analysis] - Untitled*

Deselected Phrases

Freque...	Phrase
10	home
5	Germany
5	employees

Selected Phrases

Freque...	Phrase
17	computer
13	health
8	computer terminals
7	VDT
7	terminals
6	workplace
6	video display terminals
5	eyes
5	vision
5	screen
5	video display
5	Australia
4	safety
4	Act
4	education
4	topic
4	research

Visited: 59

4. SOM is generated based on the phrases selected. Steps 3 and 4 can be done iteratively to refine the results.

Meta Spider - [Last Step: Categorization] - U:\MetaSpider\newSpider\fatigue.spd*

SOM Map regions: *ergonomics, *repetitive stress injury, *Computer eye strain

#URLS	Phrase
7	repetitive stress injur
5	ergonomics
1	Computer eye strain

Done Visited: 31 Found: 15 NUM

Figure 2. Example of a User Session with Meta Spider

BEST COPY AVAILABLE

3. SYSTEM DESIGN

Two different prototypes based on the proposed architecture have been built. Competitive Intelligence Spider, or CI Spider, collects web pages on a real-time basis from websites specified by the user and performs indexing and categorization analysis on them, to provide the user with a comprehensive view of the websites of interest. A sample user session with CI Spider is shown in Figure 1. The second tool, Meta Spider, has similar functionalities as the CI Spider, but instead of performing breadth-first search on a particular website, connects to different search engines on the Internet and integrates the results. A sample user session with Meta Spider is shown in Figure 2.

The architecture of CI Spider and Meta Spider is shown in Figure 3. There are 4 main components, namely (1) User Interface, (2) Internet Spiders, (3) Noun Phraser, and (4) Self-Organizing Map (SOM). These components work together as a unit to perform web search and analysis.

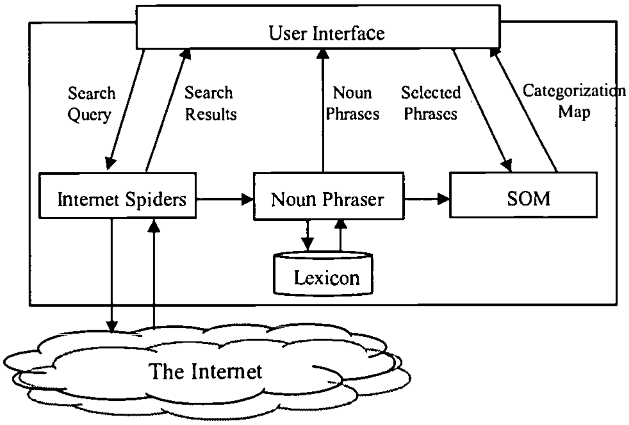


Figure 3. System Architecture

3.1 Internet Spiders

In CI Spider, the Internet Spiders are Java spiders that start from the URLs specified by the user and follow the outgoing links to search for the given keywords, until the number of web pages collected reaches a user-specified target. The spiders run in multi-thread such that the fetching process will not be affected by slow server response time. Robots exclusion protocol is also implemented such that the spiders will not access sites where the web master has placed a text file in a host or a meta-tag in a web page, indicating that robots are not welcome to these sites.

In the case of Meta Spider, the Internet Spiders first send the search queries to the search engines chosen. After the results are obtained, the Internet Spiders attempt to fetch every result page. Deadlinks and pages which do not contain the search keyword are discarded.

Whenever a page is collected during the search, the link to that page is displayed dynamically. The user can click on any link displayed and read its full content without having to wait for the whole search to be completed. The user can also switch to the Good URL List to browse only the pages that contain the search keyword. When the number of web pages collected meets the amount specified by the user, the spiders will stop and the results will be sent to the Noun Phraser for analysis.

3.2 Noun Phraser

The Arizona Noun Phraser developed at the University of Arizona is the indexing tool used to index the key phrases that appear in each document collected from the Internet by the Internet Spiders. It extracts all the noun phrases from each document based on part-of-speech tagging and linguistic rules [21]. The Arizona Noun Phraser has three components. The tokenizer takes web pages as text input and creates output that conforms to the UPenn Treebank word tokenization rules by separating all punctuation and symbols from text without interfering with textual content. The tagger module assigns a part-of-speech to every word in the document. The last module, called the phrase generation module, converts the words and associated part-of-speech tags into noun phrases by matching tag patterns to a noun phrase pattern given by linguistic rules. Readers are referred to [21] for more detailed discussion. The frequency of every phrase is recorded and sent to the User Interface. The user can view the document frequency of each phrase and link to the documents containing that phrase. After all documents are indexed, the data are aggregated and sent to the Self-Organizing Map for categorization.

3.3 Self-Organizing Map (SOM)

In order to give users an overview of the set of documents collected, the Kohonen SOM employs an artificial neural network algorithm to automatically cluster the web pages collected into different regions on a 2-D map [10]. Each document is represented as an input vector of keywords and a two-dimensional grid of output nodes is created. After the network is trained, the documents are submitted to the network and clustered into different regions. Each region is labeled by the phrase which is the key concept that most accurately represents the cluster of documents in that region. More important concepts occupy larger regions, and similar concepts are grouped in a neighborhood [13]. The map is displayed through the User Interface and the user can view the documents in each region by clicking on it.

3.4 Personalization Features

Because both CI Spider and Meta Spider have been designed for personalized web search and analysis, a user has been given more control during the search process.

In the Options Panel, the user can specify how the search is to be performed. This is similar to the "Advanced Search" feature of some commercial search engines. The user can specify number of web pages to be retrieved, domains (e.g. .gov, .edu or .com) to be included in the search results, number of Internet Spiders to be used, and so on. In CI Spider, the user can also choose either Breadth-First Search or Best-First Search to be the algorithm used by the Internet Spiders.

The SOM also is highly customizable in the sense that the user can select and deselect phrases for inclusion in the analysis and produce a new map at any time. If the user is not satisfied with the map produced, he can always go back to the previous step to discard some phrases that are irrelevant or too general and generate a new map within seconds. The systems also let each user store a personalized "dictionary" which contains the terms that the user does not want to be included in the results of the Arizona Noun Phraser and the SOM.

Another important functionality incorporated in the system is the Save function. The user can save a completed search session and open it at a later time. This feature allows the user to perform a web search and review it in the future. This also helps users who want to monitor web pages on a particular topic or website.

4. EVALUATION METHODOLOGIES

Two separate experiments have been conducted to evaluate CI Spider and Meta Spider. Because we designed the two spider systems to facilitate both document retrieval and document categorization tasks, traditional evaluation methodologies would not have been appropriate. These methodologies treat document retrieval and document categorization separately. In our experiments, the experimental task was therefore so designed as to permit evaluation of the performance of a combination of their functionalities in identifying the major themes related to a certain topic being searched.

4.1 Evaluation of CI Spider

In our experiment, CI Spider was compared with the usual methods that Internet users use to search for information on the Internet. General users usually use popular commercial search engines to collect data on the Internet, or they simply explore the Internet manually. Therefore, these two search methods were compared with the CI Spider. The first method evaluated was Lycos, chosen because it is one of the few popular search engines that offer the functionality to search for a certain keyword in a given web domain. The second method was "within-site" browsing and searching. In this method the subject was allowed to freely explore the contents in the given website using an Internet browser. When using CI Spider, the subject was allowed to use all the components including Noun Phraser and SOM.

Each subject first tried to locate the pages containing the given topic within the given web host using the different search methods described above. The subject was required to comprehend the contents of all the web pages relevant to that keyword, and to summarize the findings as a number of themes. In our experiment, a theme was defined as "a short phrase which describes a certain topic." Phrases like "success of the 9840 tape drive in the market" and "business transformation services" are examples of themes in our experiment. By examining the themes that the subjects came up with using different search methods, we were able to evaluate how effectively and efficiently each method helped a user locate a collection of documents and gain a general understanding of the response to a given search query on a certain website. Websites with different sizes, ranging from small sites such as www.eye2eye.com to large sites such as www.ibm.com were chosen for the experiments.

Six search queries were designed for the experiment, based on suggestions given by professionals working in the field of competitive intelligence. For example, one of our search tasks was to locate and summarize the information related to "merger" on the website of a company called Phoenix Technologies (www.phoenix.com). Two pilot studies were conducted in order to refine the search tasks and experiment design. During the real experiment, thirty subjects, mostly information systems management students, were recruited and each subject was required to perform three out of the six different searches using the three different search methods. At the beginning of each experiment session, the subject was trained in using these search methods. Each subject performed at least one complete search session for each of the 3 search methods until he felt comfortable with each method. Rotation was applied such that the order of search methods and search tasks tested would not bias our results.

4.2 Evaluation of Meta Spider

Meta Spider was compared with MetaCrawler and NorthernLight. MetaCrawler (www.metacrawler.com) is a renowned, popular meta-search engine and has been recognized for its adaptability, portability and scalability [20]. NorthernLight (www.northernlight.com), being one of the largest search engines on the web, provides clustering functionality to classify search results into different categories. When using Meta Spider, the subject was allowed to use all the components including Noun Phraser and SOM.

Each subject was required to use the different search tools to collect information related to the given topic. As in the CI Spider experiment, each subject was required to summarize the web pages collected as a number of themes. The search topics were chosen from TREC 6 topics. Because the TREC topics were not especially designed for web document retrieval, care was taken to make sure each search topic was valid and retrievable on the Internet. Thirty undergraduate students from an MIS class at The University of Arizona were recruited to undertake the experiment. Training and rotation similar to those used in the CI Spider experiment were applied.

5. EXPERIMENT RESULTS AND DISCUSSION

Two graduate students majoring in library science were recruited as experts for each experiment. They employed the different search methods and tools being evaluated and came up with a comprehensive set of themes for each search task. Their results were then aggregated to form the basis for evaluation. Precision and recall rates for themes were used to measure the effectiveness of each search method.

The time spent for each experiment, including the system response time and the user browsing time, was recorded in order to evaluate the efficiency of the 3 search methods in each experiment. During the studies, we encouraged our subjects to tell us about the search method used and their comments were recorded. Finally, each subject filled out a questionnaire to record further comments about the 3 different methods.

5.1 Experiment Results of CI Spider

The quantitative results of the CI Spider experiment are summarized in Table 1. Four main variables for each subject have been computed for comparison: precision, recall, time, and ease of use. Precision rate and recall rate were calculated as follows:

$$precision = \frac{\textit{number of correct themes identified by the subject}}{\textit{number of all themes identified by the subject}}$$

$$recall = \frac{\textit{number of correct themes identified by the subject}}{\textit{number of all themes identified by the expert judges}}$$

The time recorded was the total duration of the search task, including both response time of the system and the browsing time of the subject. Usability was calculated based on subjects' responses to the question "How easy/difficult is it to locate useful information using [that search method]?" Subjects were required to choose a level from a scale of 1 to 5, with 1 being the most difficult and 5 being the easiest.

In order to see whether the differences between the values were statistically significant, *t*-tests were performed on the experimental data. The results are summarized in Table 2. As can be seen, the precision and recall rates for CI Spider both were significantly higher than those of Lycos at a 5% significant level. CI Spider also was given a statistically higher value than Lycos and within-site browsing and searching in usability.

Table 1: Experiment results of CI Spider

		CI Spider	Lycos	Within-Site Browsing/ Searching
Precision:	Mean	**0.708**	0.477	0.576
	Variance	0.120	0.197	0.150
Recall:	Mean	**0.273**	0.163	0.239
	Variance	0.027	0.026	0.033
Time(min):	Mean	10.02	9.23	**8.60**
	Variance	11.86	44.82	36.94
Usability*:	Mean	**3.97**	3.33	3.23
	Variance	1.34	1.13	1.29

*Based on a scale of 1 to 5, where 1 being the most difficult to use and 5 being the easiest.

Table 2: *t*-test results of CI Spider Experiment

	CI Spider vs Lycos	CI Spider vs Within-Site B/S	Lycos vs Within-Site B/S
Precision	***0.029**	0.169	0.365
Recall	***0.012**	0.459	0.087
Time	0.563	0.255	0.688
Usability	***0.031**	***0.016**	0.126

* The mean difference is significant at the 0.05 level.

5.2 Experiment Results of Meta Spider

Three variables, namely precision, recall, and time, have been computed for comparison in the Meta Spider experiment and the results are summarized in Table 3. The *t*-test results are summarized in Table 4. In terms of precision, Meta Spider performed better than MetaCrawler and NorthernLight, and the difference with NorthernLight was statistically significant. For recall rate, Meta Spider was comparable to MetaCrawler and better than NorthernLight.

Table 3: Experiment results of Meta Spider

		Meta Spider	Meta-Crawler	Northern-Light
Precision:	Mean	**0.815**	0.697	0.561
	Variance	0.281	0.315	0.402
Recall:	Mean	0.308	**0.331**	0.203
	Variance	0.331	0.291	0.181
Time(min):	Mean	**10.93**	11.13	11.00
	Variance	4.04	4.72	5.23

Table 4: *t*-test results of Meta Spider Experiment

	Meta Spider vs Meta-Crawler	Meta Spider vs Northern-Light	Meta-Crawler vs Northern-Light
Precision	0.540	***0.013**	0.360
Recall	1.000	0.304	0.139
Time	1.000	1.000	1.000

* The mean difference is significant at the 0.05 level.

5.3 Strength and Weakness Analysis

5.3.1 Precision and Recall

The *t*-test results show that CI Spider performed statistically better in both precision and recall than Lycos, and Meta Spider performed better than NorthernLight in precision. In terms of precision, we suggest that the main reason for the high precision rate of CI Spider and Meta Spider is their ability to fetch and verify the content of each web page in real time. That means our Spiders can ensure that every page shown to the user contains the keyword being searched. On the other hand, we found that indexes in Lycos and NorthernLight, like most other search engines, were often outdated. A number of URLs returned by these two search engines were irrelevant or dead links, resulting in low precision. Subjects also reported that in some cases two or more URLs returned by Lycos pointed to the same page, which led to wasted time verifying the validity of each page.

The high recall rate of CI Spider is mainly attributable to the exhaustive searching nature of the spiders. Lycos has the lowest recall rate because, like most other commercial search engines, it samples only a number of web pages in each website, thereby missing other pages that contain the keyword. For within-site browsing and searching, a user is more likely to miss some important pages because the process is mentally exhausting.

5.3.2 Display and Analysis of Web Pages

In the CI Spider study, subjects believed it was easier to find useful information using CI Spider (with a score of 3.97/5.00) than using Lycos domain search (3.33) or manual within-site browsing and searching (3.23). Three main reasons may account for this. The first is the high precision and recall discussed above. The high quality of data saved users considerable time and mental effort. Second, the intuitive and useful interface design helped subjects locate information they needed more easily. Third, the analysis tools helped subjects form an overview of all the relevant web pages collected. The Arizona Noun Phraser allowed subjects to narrow and refine their searches as well as provided a list of key phrases that represented the collection. The Self-Organizing Map generated a

2-D map display on which subjects could click to view the documents related to a particular theme of the collection.

In our post-test questionnaires in the CI Spider experiment, we found that 77% of subjects found the Good URL List useful for their analyses, while 40% of subjects found either the Noun Phraser or the SOM useful. This suggests that while many subjects preferred traditional search result list, a significant portion of subjects were able to gain from the use of advanced analysis tools. Similar results were obtained in the Meta Spider experiment, in which 77% of subjects found the list display useful and 45% found either the Noun Phraser or the SOM useful.

5.3.3 Speed

The *t*-test results demonstrated that the three search methods in each experiment did not differ significantly in time requirements. As discussed in the previous section, the time used for comparison is total searching time and browsing time. Real-time indexing and fetching time, which usually takes more than 3 minutes, also was included in the total time for CI Spider and Meta Spider. Therefore, we anticipate that the two Spiders can let users spend less time and effort in the whole search process, because the users only need to browse the verified results.

6. CONCLUSION

The results of the two studies are encouraging. They indicate that the use of CI Spider and Meta Spider can potentially facilitate the web searching process for Internet users with different needs by using a personalized approach. The results also demonstrated that powerful AI techniques such as noun phrasing and SOM can be processed on the user's personal computer to perform further analysis on web search results, which allows the user to understand the search topic more correctly and more completely. We believe that many other powerful techniques can possibly be implemented on client-side search tools to improve efficiency and effectiveness in web search as well as other information retrieval applications.

7. ACKNOWLEDGMENTS

We would like to express our gratitude to the following agencies for supporting this project:

- NSF Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management", IIS-9817473, April 1999-March 2002.
- NSF/CISE/CSS, "An Intelligent CSCW Workbench: Personalized Analysis and Visualization", IIS-9800696, June 1998-June 2001.

We would also like to thank all the members of the Artificial Intelligence Lab at the University of Arizona who have contributed in developing the two search tools, in particular Wojciech Wyzga, Harry Li, Andy Clements, Ye Fang, David Hendriawan, Hadi Bunnalim, Ming Yin, Haiyan Fan, Bill Oliver and Esther Chou.

8. REFERENCES

[1] Bowman, C., Danzig, P., Manber, U. and Schwartz, F. Scalable Internet Resource Discovery: Research Problems and Approaches, Communications of the ACM 37(8): 98-107 (1994).

[2] Chakrabarti, S., van der Berg, M. and Dom, B. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery, in Proceedings of the 8th International World Wide Web Conference (Toronto, Canada, 1999).

[3] Chen, H., Schufels, C. and Orwig, R. Internet Categorization and Search: A Self-Organizing Approach, Journal of Visual Communication and Image Representation, 7(1): 88-102 (1996).

[4] Chignell, M. H., Gwizdka, J. and Bodner, R. C. Discriminating Meta-Search: A Framework for Evaluation. Information Processing and Management, 35 (1999).

[5] Cho, J., Garcia-Molina, H. and Page, L. Efficient Crawling Through URL Ordering, in Proceedings of the 7th World Wide Web Conference (Brisbane, Australia, Apr 1998).

[6] Fox, E., Hix, D., Nowell, L. T., Brueni, D. J., Wake, W. C., Lenwood, S. H. and Rao, D. Users, User Interfaces, and Objects: Envision, A Digital Library. Journal of the American Society for Information Science, 44(8), 480-491 (1993).

[7] Hearst, M. TileBars: Visualization of Term Distribution Information in Full Text Information Access, in Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'95), 59-66 (1995).

[8] Hearst, M. and Pedersen, J. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, in Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96), 76-84 (1996).

[9] Inktomi WebMap, http://www.inktomi.com/webmap/

[10] Kohonen, T. Self-Organizing Maps. Springer-Verlag, Berlin, 1995.

[11] Kohonen, T. Exploration of Very Large Databases by Self-Organizing Maps, in Proceedings of the IEEE International Conference on Neural Networks, 1:1-6 (1997).

[12] Lawrence, S. and Giles, C. L. Accessibility of Information on the Web, Nature, 400 (1999), 107-109.

[13] Lin, C., Chen, H. and Nunamaker J. Verifying the Proximity and Size Hypothesis for Self-Organizing Maps. Journal of Management Information Systems, 16(3) (1999-2000), 61-73.

[14] Lin, X., Soergel, D., and Marchionini, G. A Self-organizing Semantic Map for Information Retrieval, in Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval (1991), 262-269.

[15] Maes, P. Agents that Reduce Work and Information Overload. Communications of the ACM, 37(7) (July 1994), 31-40.

[16] McBryan, O. GENVL and WWW: Tools for Taming the Web, in Proceedings of the 1st International World Wide Web Conference (Geneva, Switzerland, Mar 1994).

[17] Rasmussen, E. Clustering Algorithms. In W. B. Frakes and R. Baeza-Yates (eds.) Information Retrieval Data Structures and Algorithms, Prentice Hall, N. J., 1992.

[18] Salton, G. Another look at automatic text-retrieval systems. Communications of the ACM, 29(7) (1986), 648-656.

[19] Schneiderman, B., Feldman, D., Rose, A. and Grau, X. F. Visualizing Digital Library Search Results with Categorical and Hierarchical Axes, in Proceedings of 5th ACM Conference on ACM 2000 Digital Libraries (San Antonio, Texas USA, 2000).

[20] Selberg, E. and Etzioni, O. The MetaCrawler architecture for resource aggregation on the Web. IEEE Expert (1997).

[21] Tolle, K. M. and Chen, H. Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools. Journal of the American Society for Information Science, 51(4), 352-370 (2000).

[22] Veerasamy, A. and Belkin, N. J., Evaluation of a Tool for Visualization of Information Retrieval Results, in Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96), 85-92 (1996).

[23] Zamir, O. and Etzioni, O. Grouper: A Dynamic Clustering Interface to Web Search Results, in Proceedings of the 8th International World Wide Web Conference (Toronto, Canada, May 1999).

Salticus: Guided Crawling for Personal Digital Libraries

Robin Burke
California State University, Fullerton
Dept. of Information Systems and Decision Sciences
Fullerton, CA 92834
rburke@fullerton.edu

ABSTRACT

In this paper, we describe Salticus, a web crawler that learns from users' web browsing activity. Salticus enables users to build a personal digital library by collecting documents and generalizing over the user's choices.

Keywords

personal digital library, business intelligence, web crawling, document acquisition

1. INTRODUCTION

One solution to the problem of information stability on the World-Wide Web is the creation of a "personal digital library," [1] a personal collection of documents from on-line sources, stored in a local cache and organized for easy access.

The Document Organization aNd Navigation Agent (DONNA) is a project that seeks to build personal digital libraries to support competitive business intelligence. A personal library, by its nature, does not have the problems of scale presented by public libraries. On the other hand, all aspects of a document's life-cycle must be managed by a single individual.

This paper describes our work in the area of document acquisition. The business intelligence professionals with whom we work use the Web for a large proportion of their information gathering activity, and have a set of strategies for seeking out information of different types. Many of these strategies are not accessible to standard link-based web crawlers. For example, the analysts we studied made heavy use of site-based search engines.

Salticus is a document collection agent that observes a user's browsing and document collection behavior and makes predictions about other useful documents. The user can use these predictions to broaden the collection process and automate it for future visits to the same pages.

2. WEB CRAWLING

The standard approach to document collection and indexing on the web is the use of a web crawler. [2] If the Web is viewed as a graph with the nodes as documents and the edges as hyperlinks, a crawler typically performs some type of best-first search through the graph, indexing or collecting all of the pages it finds.

This approach is suitable for building a comprehensive index, as found in search engines such as Google or AltaVista. For a personal digital library, we must be more selective. One approach is to focus the crawler on a particular site and mirror its complete contents. Several commercial tools have this capability, for example, WebReaper.[1]

This is also unsatisfactory for several reasons. First, many of the documents returned in a complete mirror are irrelevant to the analyst's work. Second, many documents available to a human user are not accessible with link-based crawling at all, because they are accessible only through login pages or other form-based access controls. It should be noted that such controls are often put in place precisely so that documents can be made available to human users and not to web crawlers.

3. SALTICUS

3.1 Document collection

Salticus[2] is a part of the DONNA system that acts as a proxy between the user and the web. It observes all of a user's browsing behavior including form submission, authentication interactions and cookie creation. To record browsing behavior, the user initiates an "excursion" onto the web with Salticus observing. The system builds a list of the interactions that occur between the user and the various web servers visited, and caches the documents that are accessed. When the user decides to collect a document, it is transferred to the appropriate collection within DONNA.

With this capability, the system operates much like document collecting systems like iHarvest[3] that also build personal collections of web documents. Where Salticus differs is in its ability to generalize over document collection actions and to predict other documents to gather.

3.2 Structural generalization

There are several ways that one might generalize about a user's document gathering activity. One possibility is to generalize over the content of links or documents downloaded, building a representation of the user's interests. [3, 4] For this approach to work, the crawler must either rely entirely on the text of the hyperlink to the document as evidence of its content, or it must download every document and analyze it to determine its value.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

[1] <URL: http://www.otway.com/webreaper/ >

[2] Salticus = Search Augmented with Learning Techniques In a Customizable User Spider. It is also the name of a genus of spiders, known for its jumping abilities and advanced eyes.

[3] <URL: http://www.iharvest.com/ >

With the text associated with links not typically very informative, and the prospect of processing every document not practical in an interactive system, we chose a third way, which is to focus on the structure of documents. Salticus has a "Collect" mode, which enables the user to select items to collect on a page without navigating to them. As the user selects items to collect, Salticus builds an XPath [5] representation of each selected link, and it generalizes over these structural descriptions to predict what else on the page might be of interest.

We have found that very simple pre-fix/post-fix generalization is sufficient for the tasks we are supporting. For example, suppose the user has selected three links with the following XPath representations:

HTML:BODY:TABLE[1]:TR[1]:TD[1]:A[1]
HTML:BODY:TABLE[1]:TR[2]:TD[1]:A[1]
HTML:BODY:TABLE[1]:TR[3]:TD[1]:A[1]

Salticus predicts that the links the user is interested in will be those that are exhibit the same variation at the user's choices. It selects the largest common prefix from these paths and the largest common postfix, and replaces the varying part with "*" or "don't care" pattern. In this case, the pattern becomes

HTML:BODY:TABLE[1]:TR[*]:TD[1]:A[1]

which is the first link in the first cell of every row of table 1. If the user instead had selected links in table 1 and in table 2, then the generalization would include both the table and the row.

3.3 Example

Suppose a user initiates a Salticus excursion and enters the address of KMWorld, an on-line publication in the area of knowledge management. Once at the site, the user enters the term "workflow" in the site's search box. The next page shows a list of 25 documents containing this term. The user selects "Collect" mode in Salticus and clicks on the first several links. At each click, the associated document is retrieved and collected by the proxy, but the original page is still displayed to the user. Then the user clicks on Salticus's "Predict" function, the system predicts that all of the documents should be collected, and the user accepts the predictions, copying all of the files into the collection.

3.4 Automated operation

Once the user has performed this excursion, Salticus has a record of the steps required to return to the same "place" in the future. Note that with the widespread adoption of database-backed web sites and session-based technologies such as ASP and ColdFusion, the URL has ceased to become a useful identifier for internal pages within a site. A user trying to reissue a URL from a previous session will typically get redirected to the outermost part of the site to login and walk through the application once more.

It is therefore not sufficient to record and replay the URLs that a user has visited. Instead, Salticus makes use of the XPaths recorded for each interaction. When replaying an excursion autonomously, Salticus starts by issuing a URL for the first interaction, but for every subsequent interaction, it follows the XPath to the same location on the page where the user clicked in the previous session. This method is robust in the face of session-based URLs. It is therefore possible for the user to send Salticus to collect "workflow" related documents from KMWorld in the future, as long as the site is not redesigned.

4. FUTURE WORK

Salticus pattern-based prediction of useful documents has worked well in our informal evaluation of the collecting patterns of intelligence analysts. However, it has some significant limitations. It cannot capture more complex patterns of collection, such as "every third link" or "all rows of only tables 2 and 4". We are investigating where our current model fails and how more complex predictions might be made.

The problem of automated revisitation brings to the fore the problem of identity: what constitutes a new page? Or a new version of an already-collected page? In the world of session-based URLs, no page will have the same URL that it did when previously visited, even if its contents are the same. We believe that we will be able to use our path-based representation as additional evidence in determining the novelty of documents.

We are also seeking to identify high-level search behaviors, such as the site search engine strategy in the workflow example above, higher-level collection patterns for Salticus to recognize. This would enable the system to make predictions that go across sites and pages, and provide more possibilities for automation.

5. CONCLUSION

Salticus is a document acquisition agent that assists a user building a personal digital library. Salticus tracks user browsing and document collection and generalizes over the user's collection actions. Salticus's path-based representation enables it to avoid the problems associated with URLs as identifiers.

6. ACKNOWLEDGMENTS

The DONNA project is sponsored by the FileNet Corporation, a leading provider of document management software, and by the University of California's DiMI program. Christie Delzeit and Melissa Hanson implemented Salticus's user interface.

7. REFERENCES

[1] Rasmusson, A & Olsson, T & Hansen, P. 1998. A Virtual Community Library: SICS Digital Library Infrastructure Project. Research and Advanced Technology for Digital Libraries, CDL'98. Lecture Notes in Computer Science, Vol. 1513. pp 677-678. Springer Verlag.

[2] Heydon, A, and Najork, M. A. 1999. A scalable, extensible web crawler. World Wide Web, 2(4):219-229, December 1999.

[3] Chakrabarti, S., van der Berg, M., & Dom, B. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. In Proceedings of WWW8.

[4] Miller, R. C. and Bharat, K. 1998. "SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers." Proceedings of WWW7, pp. 119-130, Brisbane, Australia, April 1998.

[5] World-Wide Web Consortium, 1999. XML Path Language (XPath) Version 1.0. <URL: http://www.w3.org/TR/1999/REC-xpath-19991116>

Power to the People: End-user Building of Digital Library Collections

Ian H. Witten
Dept of Computer Science
The University of Waikato
Hamilton, New Zealand
+64 7 838 4246
ihw@cs.waikato.ac.nz

David Bainbridge
Dept of Computer Science
The University of Waikato
Hamilton, New Zealand
+64 7 838 4407
davidb@cs.waikato.ac.nz

Stefan J. Boddie
Dept of Computer Science
The University of Waikato
Hamilton, New Zealand
+64 7 838 6038
sjboddie@cs.waikato.ac.nz

ABSTRACT

Naturally, digital library systems focus principally on the reader: the consumer of the material that constitutes the library. In contrast, this paper describes an interface that makes it easy for people to build their own library collections. Collections may be built and served locally from the user's own web server, or (given appropriate permissions) remotely on a shared digital library host. End users can easily build new collections styled after existing ones from material on the Web or from their local files—or both, and collections can be updated and new ones brought on-line at any time. The interface, which is intended for non-professional end users, is modeled after widely used commercial software installation packages. Lest one quail at the prospect of end users building their own collections on a shared system, we also describe an interface for the administrative user who is responsible for maintaining a digital library installation.

1. INTRODUCTION

The Greenstone Digital Library Software from the New Zealand Digital Library (NZDL) project provides a new way of organizing information and making it available over the Internet. A *collection* of information is typically comprised of several thousand or several million *documents*, and a uniform interface is provided to all documents in a collection. A library may include many different collections, each organized differently—though there is a strong family resemblance in how they are presented.

Greenstone collections are widely used, with many of them publicly available on the Web. Some have also been written, in precisely the same form, to CD-ROMS that are widely distributed within developing countries (50,000 copies/year). Created on behalf of organizations such as UNESCO, the Pan-American Health Organization, the World Health Organization, and the United Nations University, they cover topics ranging from basic humanitarian needs through environmental concerns to disaster relief. Titles include the *Food and Nutrition Library*, *World Environmental Library*, *Humanity Development Library*, *Medical and Health Library*, *Virtual Disaster Library*, and the *Collection on Critical Global Issues*. Further details can be obtained from *www.nzdl.org*.

A recent enhancement to Greenstone is a facility for what we call "end-user collection building." It was provoked by our work on digital libraries in developing countries, and in particular by the observation that effective human development blossoms from empowerment rather than gifting. Disseminating information originating in the developed world, as the above-mentioned collections do, is very useful for developing countries. But a more effective strategy for sustained long-term development is to disseminate the capability to create information collections rather than the collections themselves [7]. This allows developing countries to participate actively in our information society rather than observing it from outside. It will stimulate the creation of new industry. And it will help ensure that intellectual property remains where it belongs—in the hands of those who produce it.

The end-user collection building facility, which we call the "Collector," is modeled after popular end-user installation software (such as InstallShield[1]). Frequently called a software "wizard"—a term we deprecate because of its appeal to mysticism and connotations of utter inexplicability—this interaction style suits novice users because it simplifies the choices and presents them very clearly.

The core Greenstone software addresses the needs of the reader: the Collector addresses the needs of people who want to build and distribute their own collections. A third class of user, vital in any multi-user Greenstone installation is the system administrator, who is responsible for configuring the software to suit the local environment, enabling different classes of Greenstone user, setting appropriate file permissions, and so on. Greenstone includes an interface, not described in previous papers, through which the administrative user can check the status of the system, and alter it, interactively. Sensitive and flexible administrative support becomes essential when many end users are building collections.

This paper begins with a brief synopsis of the features of Greenstone. To some extent this section overlaps material presented at DL00 [8], but many features are new and others extend what was previously reported. The remainder of the paper is completely new. We examine the new interactive interface for collection building, which will extend Greenstone's domain of application by encouraging end users to create their own digital library collections. The structure of a collection is determined by

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

[1] www.installshield.com

its "collection configuration file," and we briefly examine what can be specified in this file. Next we turn to the administrator's interface and describe the facilities it provides. Finally we discuss the design process and usability evaluation of the system.

2. THE GREENSTONE SOFTWARE

To convey the breadth of coverage provided by Greenstone, we start with a brief overview of its facilities. More detail appears in [8].

Accessible via Web browsers. Collections are accessed through a standard web browser, such as Netscape or Internet Explorer. The browser is used for both local and remote access—whether Greenstone is running on your own personal computer or on a remote central library server.

Runs on Windows and Unix. Collections can be served on either Windows (3.1/3.11, 95/98, NT) or Unix (Linux and SunOS). Any of these systems serve Greenstone collections over the Internet using either an integrated built-in Web server (the "local library" version of Greenstone) or an external server—typically Apache (the "web library" version).

Full-text and fielded search. Users can search the full text of the documents, or choose between indexes built from different parts of the documents. Some collections have an index of full documents, an index of sections, an index of paragraphs, an index of titles, and an index of authors, each of which can be searched for particular words or phrases. Queries can be ranked or Boolean; terms can be stemmed or unstemmed, case-folded or not.

Flexible browsing facilities. The user can browse lists of authors, lists of titles, lists of dates, hierarchical classification structures, and so on. Different collections offer different browsing facilities, and even within a collection a broad variety of browsing interfaces are available. Browsing and searching interfaces are constructed during the building process according to collection configuration information.

Creates access structures automatically. All collections are easy to maintain. Searching and browsing structures are built directly from the documents themselves: no links are inserted by hand. This means that if new documents in the same format become available, they can be merged into the collection automatically. However, existing hypertext links in the original documents, leading both within and outside the collection, are preserved.

Makes use of available metadata. Metadata forms the raw material for browsing indexes: it may be associated with each document or with individual sections within documents. Metadata must be provided explicitly (often in an accompanying spreadsheet) or derivable automatically from the source documents. The Dublin Core scheme is used, however, provision is made for extensions, and other schemes.

Plugins and classifiers extend the system's capabilities. "Plugins" (small modules of PERL code) can be written to accommodate new document types. Existing plugins process plain text documents, HTML documents, Microsoft WORD, PDF, PostScript, and some proprietary formats. So collections can include different source document types, a pipeline of plugins is formed and each document passed down it; the first plugin to recognize the document format processes it. Plugins are also used for generic tasks such as recursively traversing directory structures containing documents. In order to build browsing indexes from metadata, an analogous scheme of "classifiers" is used: classifiers (also written in PERL) create browsing indexes of various kinds based on metadata.

Multiple-language documents. Unicode is used throughout the software, allowing any language to be processed in a consistent manner, and searched properly. To date, collections have been built containing French, Spanish, Maori, Chinese, Arabic and English. On-the-fly conversion is used to convert from Unicode to an alphabet supported by the user's Web browser. A "language identification" plugin allows automatic identification of languages in multilingual collections, so that separate indexes can be built.

Multiple-language user interface. The interface can be presented in multiple languages. Currently it is available in French, German, Spanish, Portuguese, Maori, Chinese, Arabic and English. New languages can be added easily.

Multimedia collections. Greenstone collections can contain text, pictures, audio and even video clips. Most non-textual material is either linked in to textual documents or accompanied by textual descriptions (ranging from figure captions to descriptive paragraphs) to allow full-text searching and browsing. However, the architecture is general enough to permit implementation of plugins and classifiers for non-textual data.

Classifiers allow hierarchical browsing. Hierarchical phrase and keyphrase indexes of text, or indeed any metadata, can be created using standard classifiers. Such interfaces are described by Gutwin *et al.* [3] and Paynter *et al.* [5].

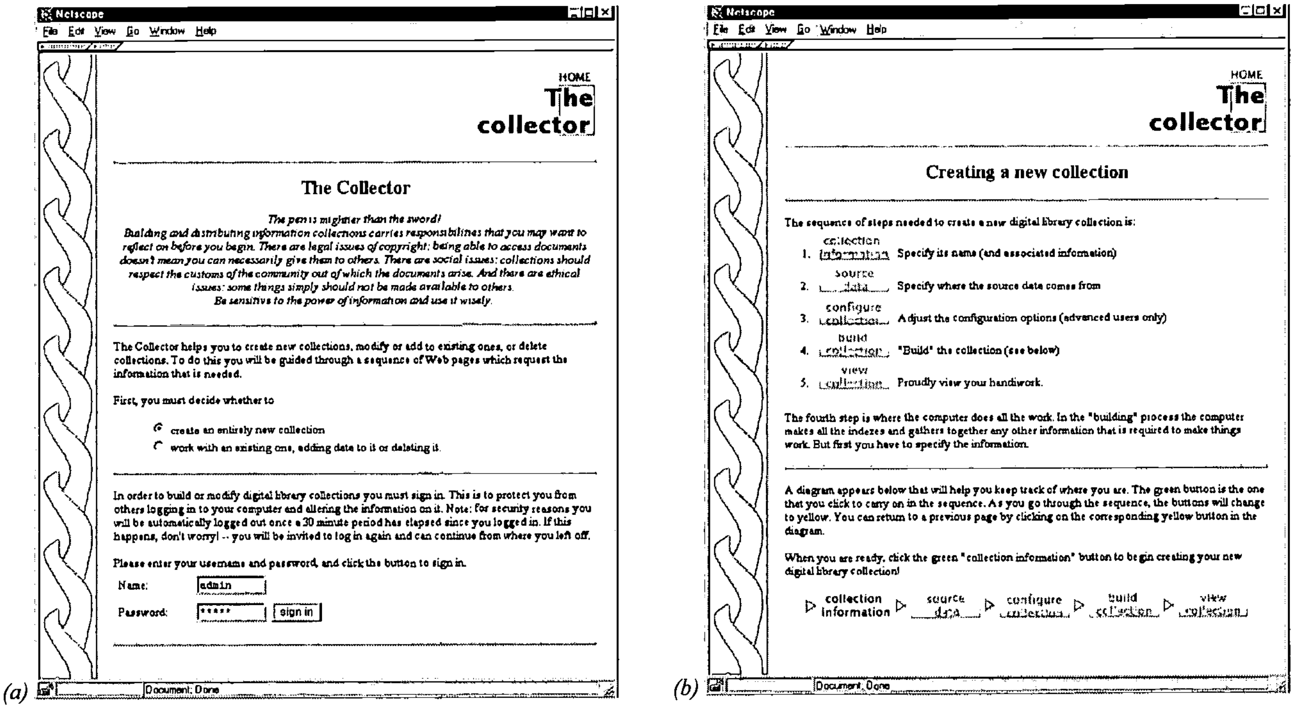
Designed for multi-gigabyte collections. Collections can contain millions of documents, making the Greenstone system suitable for collections up to several gigabytes. Compression is used to reduce the size of the indexes and text [6]. Small indexes have the added bonus of faster retrieval.

New collections appear dynamically. Collections can be updated and new ones brought on-line at any time, without bringing the system down; the process responsible for the user interface will notice (through periodic polling) when new collections appear and add them to the list presented to the user.

Collections can be published on CD-ROM. Greenstone collections can be published, in precisely the same form, on a self-installing CD-ROM. The interaction is identical to accessing the collection on the Web (Netscape is provided on each disk)—except that response times are faster and more predictable. For collections larger than one CD-ROM, a multi CD-ROM solution has been implemented.

Distributed collections are supported. A flexible process structure allows different collections to be served by different computers, yet be presented to the user in the same way, on the same Web page, as part of the same digital library. The Z39.50 protocol is also supported, both for accessing external servers and (under development) for presenting Greenstone collections to external clients.

What you see is what you get. The Greenstone Digital Library is open-source software, available from the New Zealand Digital Library (nzdl.org) under the terms of the GNU General Public License. The software includes everything described above: Web serving, CD-ROM creation, collection building, multi-lingual capability, plugins and classifiers for a variety of different source



The Collector

The pen is mightier than the sword!
Building and distributing information collections carries responsibilities that you may want to reflect on before you begin. There are legal issues of copyright: being able to access documents doesn't mean you can necessarily give them to others. There are social issues: collections should respect the customs of the community out of which the documents arise. And there are ethical issues: some things simply should not be made available to others.
Be sensitive to the power of information and use it wisely.

The Collector helps you to create new collections, modify or add to existing ones, or delete collections. To do this you will be guided through a sequence of Web pages which request the information that is needed.

First, you must decide whether to

● create an entirely new collection

○ work with an existing one, adding data to it or deleting it.

In order to build or modify digital library collections you must sign in. This is to protect you from others logging in to your computer and altering the information on it. Note: for security reasons you will be automatically logged out once a 30 minute period has elapsed since you logged in. If this happens, don't worry! -- you will be invited to log in again and can continue from where you left off.

Please enter your username and password, and click the button to sign in.

Name: admin

Password: ***** [sign in]

Creating a new collection

The sequence of steps needed to create a new digital library collection is:

1. collection information Specify its name (and associated information)
2. source data Specify where the source data comes from
3. configure collection Adjust the configuration options (advanced users only)
4. build collection "Build" the collection (see below)
5. view collection Proudly view your handiwork.

The fourth step is where the computer does all the work. In the "building" process the computer makes all the indexes and gathers together any other information that is required to make things work. But first you have to specify the information.

A diagram appears below that will help you keep track of where you are. The green button is the one that you click to carry on in the sequence. As you go through the sequence, the buttons will change to yellow. You can return to a previous page by clicking on the corresponding yellow button in the diagram.

When you are ready, click the green "collection information" button to begin creating your new digital library collection!

▷ collection information ▷ source data ▷ configure collection ▷ build collection ▷ view collection

Figure 1 Using the Collector to build a new collection (continued on next pages)

document types. It includes an autoinstall feature to allow easy installation on both Windows and Unix.

3. THE COLLECTOR

Our conception of digital libraries is captured by the following brief characterization [1]:

> *A collection of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization and maintenance of the collection.*

It is the last point—selection, organization and maintenance of the collection—that we address in this paper. Our view is that just as new books acquired by physical libraries are integrated with the existing catalog on the basis of their metadata, so one should easily be able to add to a digital library without having to edit its content in any way. Furthermore we strive to do this without manual intervention. Once added, such material should immediately become a first-class component of the library. We accomplish this with a build/rebuild process that imports new documents into a library collection using XML to standardize representation, and use explicitly stated metadata to update searching and browsing structures.

In Greenstone, the structure of a particular collection is determined when the collection is set up. This includes such things as the format, or formats, of the source documents, how they should be displayed on the screen, the source of metadata, what browsing facilities should be provided, what full-text search indexes should be provided, and how the search results should be displayed. Once the collection is in place, it is easy to add new documents to it—so long as they have the same format as the existing documents, and the same metadata is provided, in exactly the same way.

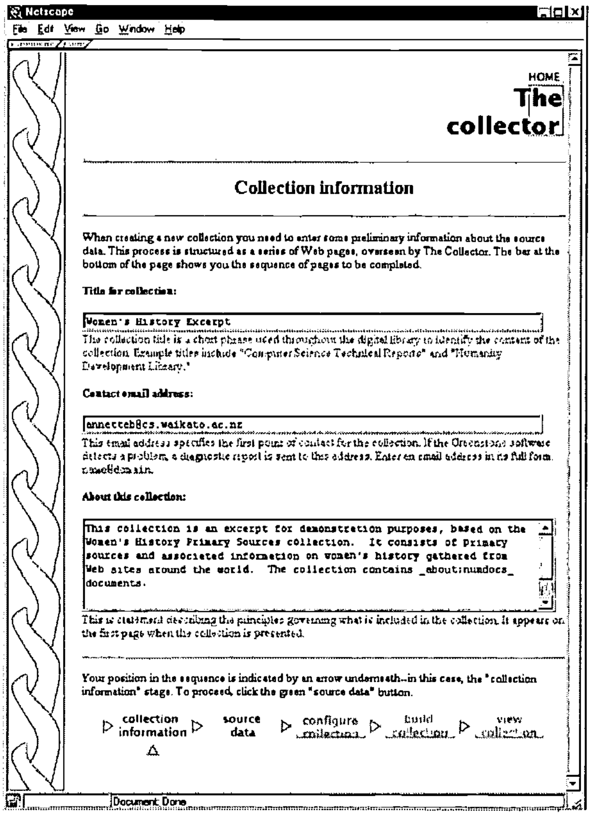
The scheme for collection building has the following basic functions:

- create a new collection with the same structure as an existing one;
- create a new collection with a different structure from existing ones;
- add new material to an existing collection;
- modify the structure of an existing collection;
- delete a collection; and
- write an existing collection to a self-contained, self-installing CD-ROM.

Figure 1 shows the Greenstone Collector being used to create a new collection, in this case from a set of HTML files stored locally. In Figure 1a, the user must first decide whether to work with an existing collection or build a new one. The former case covers the first two options above; the latter covers the remainder. While the example shows a collection being built from existing files, we emphasize that the Collector supports the creation of completely new collections formed from completely new information.

3.1 Logging in

Either way it is necessary to log in before proceeding. Note that in general, users access the collection-building facility remotely, and build the collection on the Greenstone server. Of course, we cannot allow arbitrary people to build collections (for reasons of propriety if nothing else), so Greenstone contains a security system which forces people who want to build collections to log in first. This allows a central system to offer a service to those wishing to build



Collection information

When creating a new collection you need to enter some preliminary information about the source data. This process is structured as a series of Web pages, overseen by The Collector. The bar at the bottom of the page shows you the sequence of pages to be completed.

Title for collection:

Women's History Excerpt

The collection title is a short phrase used throughout the digital library to identify the content of the collection. Example titles include "Computer Science Technical Reports" and "Humanity Development Library."

Contact email address:

annetteb@cs.waikato.ac.nz

This email address specifies the first point of contact for the collection. If the Greenstone software detects a problem, a diagnostic report is sent to this address. Enter an email address in its full form, name@domain.

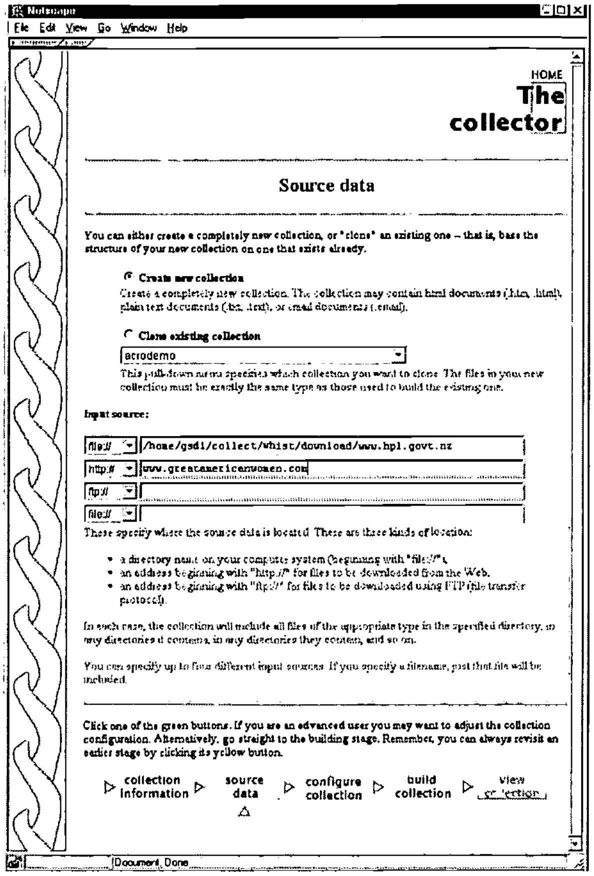
About this collection:

This collection is an excerpt for demonstration purposes, based on the Women's History Primary Sources collection. It consists of primary sources and associated information on women's history gathered from Web sites around the world. The collection contains _about:numdocs_ documents.

This is statement describing the principles governing what is included in the collection. It appears on the first page when the collection is presented.

Your position in the sequence is indicated by an arrow underneath--in this case, the "collection information" stage. To proceed, click the green "source data" button.

(c)



Source data

You can either create a completely new collection, or "clone" an existing one – that is, base the structure of your new collection on one that exists already.

● **Create new collection**

Create a completely new collection. The collection may contain html documents (.htm, .html), plain text documents (.txt, .text), or email documents (.email).

○ **Clone existing collection**

acrodemo

This pull-down menu specifies which collection you want to clone. The files in your new collection must be exactly the same type as those used to build the existing one.

Input source:

file:// /home/gsdl/collect/whist/download/www.hpl.govt.nz

http:// www.greatamericanwomen.com

ftp://

file://

These specify where the source data is located. There are three kinds of location:

- a directory name on your computer system (beginning with "file://"),
- an address beginning with "http://" for files to be downloaded from the Web.
- an address beginning with "ftp://" for files to be downloaded using FTP (file transfer protocol).

In each case, the collection will include all files of the appropriate type in the specified directory, in any directories it contains, in any directories they contain, and so on.

You can specify up to four different input sources. If you specify a filename, just that file will be included.

Click one of the green buttons. If you are an advanced user you may want to adjust the collection configuration. Alternatively, go straight to the building stage. Remember, you can always revisit an earlier stage by clicking its yellow button.

(d)

Figure 1 (continued)

information collections and use that server to make them available to others. Alternatively, a user who is running Greenstone on his or her own computer may build collections locally, but it is still necessary to log in because other people who view these Web pages should not be allowed to build collections.

3.2 Dialog structure

Upon completion of login, the page in Figure 1b appears. This shows the sequence of steps that are involved in collection building. They are:

1. Collection information
2. Source data
3. Configuring the collection
4. Building the collection
5. Viewing the collection.

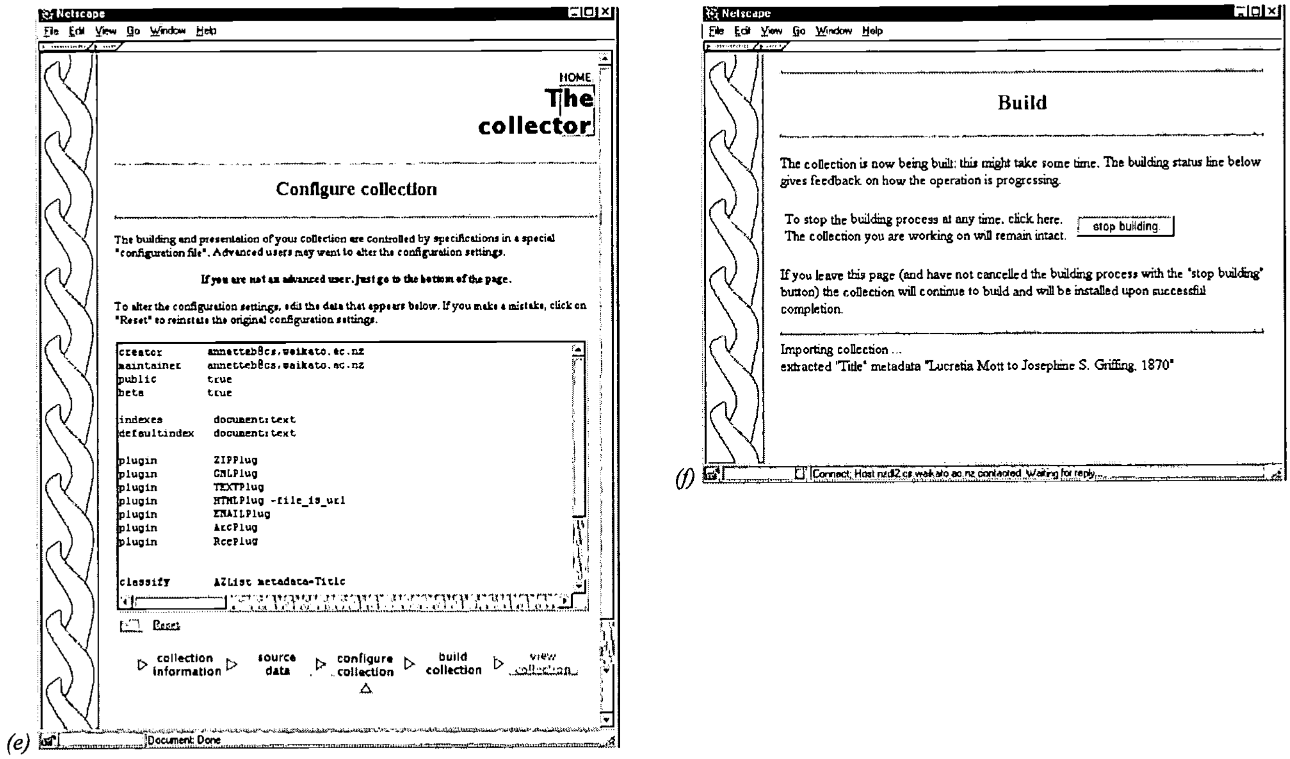
The first step is to specify the collection's name and associated information. The second is to say where the source data is to come from. The third is to adjust the configuration options, which requires considerable understanding of what is going on—it is really for advanced users only. The fourth step is where all the (computer's) work is done. During the "building" process the system makes all the indexes and gathers together any other information that is required to make the collection operate. The fifth step is to check out the collection that has been created.

These five steps are displayed as a linear sequence of gray buttons at the bottom of the screen in Figure 1b, and at the bottom of all other pages generated by the Collector. This display helps users keep track of where they are in the process. The button that should be clicked to continue the sequence is shown in green (*collection information* in Figure 1b). The gray buttons (all the others, in Figure 1b) are inactive. The buttons change to yellow as you proceed through the sequence, and the user can return to an earlier step by clicking the corresponding yellow button in the diagram. This display is modeled after the "wizards" that are widely used in commercial software to guide users through the steps involved in installing new software.

3.3 Collection information

The next step in the sequence, collection information, is shown in Figure 1c. When creating a new collection, it is necessary to enter some information about it:

- title,
- contact email address, and
- brief description.



Configure collection

The building and presentation of your collection are controlled by specifications in a special "configuration file". Advanced users may want to alter the configuration settings.

If you are not an advanced user, just go to the bottom of the page.

To alter the configuration settings, edit the data that appears below. If you make a mistake, click on "Reset" to reinstate the original configuration settings.

```
creator         annetteb@cs.waikato.ac.nz
maintainer      annetteb@cs.waikato.ac.nz
public          true
beta            true

indexes          document:text
defaultindex     document:text

plugin           ZIPPlug
plugin           GMLPlug
plugin           TEXTPlug
plugin           HTMLPlug -file_is_url
plugin           EMAILPlug
plugin           ArcPlug
plugin           RecPlug


classify         AZList metadata=Title
```

Build

The collection is now being built: this might take some time. The building status line below gives feedback on how the operation is progressing.

To stop the building process at any time, click here. The collection you are working on will remain intact.

If you leave this page (and have not cancelled the building process with the 'stop building' button) the collection will continue to build and will be installed upon successful completion.

Importing collection ...
extracted "Title" metadata "Lucretia Mott to Josephine S. Griffing, 1870"

Figure 1 (continued)

The collection title is a short phrase used through the digital library to identify the content of the collection: we have already mentioned such titles as *Food and Nutrition Library*, *World Environmental Library*, and so on. The email address specifies the first point of contact for any problems encountered with the collection. If the Greenstone software detects a problem, a diagnostic report is sent to this address. Finally, the brief description is a statement describing the principles that govern what is included in the collection. It appears under the heading *About this collection* on the first page when the collection is presented.

Lesk [4] recommends that digital libraries articulate both the principles governing what is included and how the collection is organized. *About this collection* is designed to address the first point. The second is taken care of by the help text, which includes a list of access mechanisms that is automatically generated by the system based on what searching and browsing facilities are included in the collection.

The user's current position in the collection-building sequence is indicated by an arrow that appears in the display at the bottom of each screen—in this case, as Figure 1c shows, the *collection information* stage. The user proceeds to Figure 1d by clicking the green *source data* button.

3.4 Source data

Figure 1d is the point where the user specifies the source text that comprises the collection. Either a new collection is created, or an existing one is "cloned." Creating a totally novel collection with a completely different structure from existing ones is a major undertaking, and is not what the interactive Collector interface is designed for. The most effective way to create a new collection is to base its structure on an existing one, that is, to clone it.

When cloning, the choice of current collections is displayed on a pull-down menu. Since there are usually many different collections,[2] there is a good chance that a suitable structure exists. It is preferable that the document file types in the new collection are amongst those catered for by the old one, the same metadata is available, and the metadata is specified in the same way; however, Greenstone is equipped with sensible defaults. For instance, if document files with an unexpected format are encountered, they will simply be omitted from the collection (with a warning message for each one). If the metadata needed for a particular browsing structure is unavailable for a particular document, that document will simply be omitted from the structure.

The alternative to cloning an existing collection is to create a completely new one. A bland collection configuration file is provided that accepts a wide range of different document types and generates a searchable index of the full text and an alphabetic title browser. Title metadata is available for many document types, such as HTML, email, and Microsoft WORD—note, however, that in the latter case it emanates from the system's "Summary" information for the file, which is frequently incorrect because many users ignore this Microsoft feature.

Boxes are provided to indicate where the source documents are located: up to four separate input sources can be specified. There are three kinds of specification:

[2] Collections can be downloaded from nzdl.org.

```
creator                             annetteb@cs.waikato.ac.nz
maintainer                          annetteb@cs.waikato.ac.nz
public                              true
beta                                true

indexes                             document:text
defaultindex                        document:text

plugin                              ZIPPlug
plugin                              GMLPlug
plugin                              TEXTPlug
plugin                              HTMLPlug –file_is_url
plugin                              EMAILPlug
plugin                              ArcPlug
plugin                              RecPlug


classify                            AZList metadata=Title

collectionmeta collectionname       "Women's History Excerpt"
collectionmeta collectionextra      "This collection is an excerpt for demonstration purposes, based on the\
                                    Women's History Primary Sources collection. It consists of primary\
                                    sources and associated information on women's history gathered from\
                                    Web sites around the world. The collection contains _about:numdocs_\
                                    documents"
collectionmeta .document:text       "documents"
```

Figure 2. Configuration file for collection generated in Figure 1

- a directory name on the Greenstone server system (beginning with "file://")
- an address beginning with "http://" for files to be downloaded from the Web
- an address beginning with "ftp://" for files to be downloaded using FTP.

In each case of "file://" or "ftp://" the collection will include all files in the specified directory, any directories it contains, any files and directories *they* contain, and so on. If instead of a directory a filename is specified, that file alone will be included. For "http://" the collection will mirror the specified Web site.

In the given example (Figure 1d) the new collection will contain documents taken from a local file system as well as a remote Web site, which will be mirrored during the building process, thus forming a new resource that is the composite of the two.

3.5 Configuring the collection

Figure 1e shows the next stage. The construction and presentation of all collections is controlled by specifications in a special collection configuration file (see below). Advanced users may use this page to alter the configuration settings. Most, however, will proceed directly to the final stage.

In the given example the user has made a small modification to the default configuration file by including the *file_is_url* flag with the HTML plugin. This flag causes URL metadata to be inserted in each document, based on the filename convention that is adopted by the mirroring package. This metadata is used in the collection to allow readers to refer to the original source material, rather than to a local copy.

3.6 Building the collection

Figure 1f shows the "building" stage. Up until now, the responses to the dialog have merely been recorded in a temporary file. The building stage is where the action takes place.

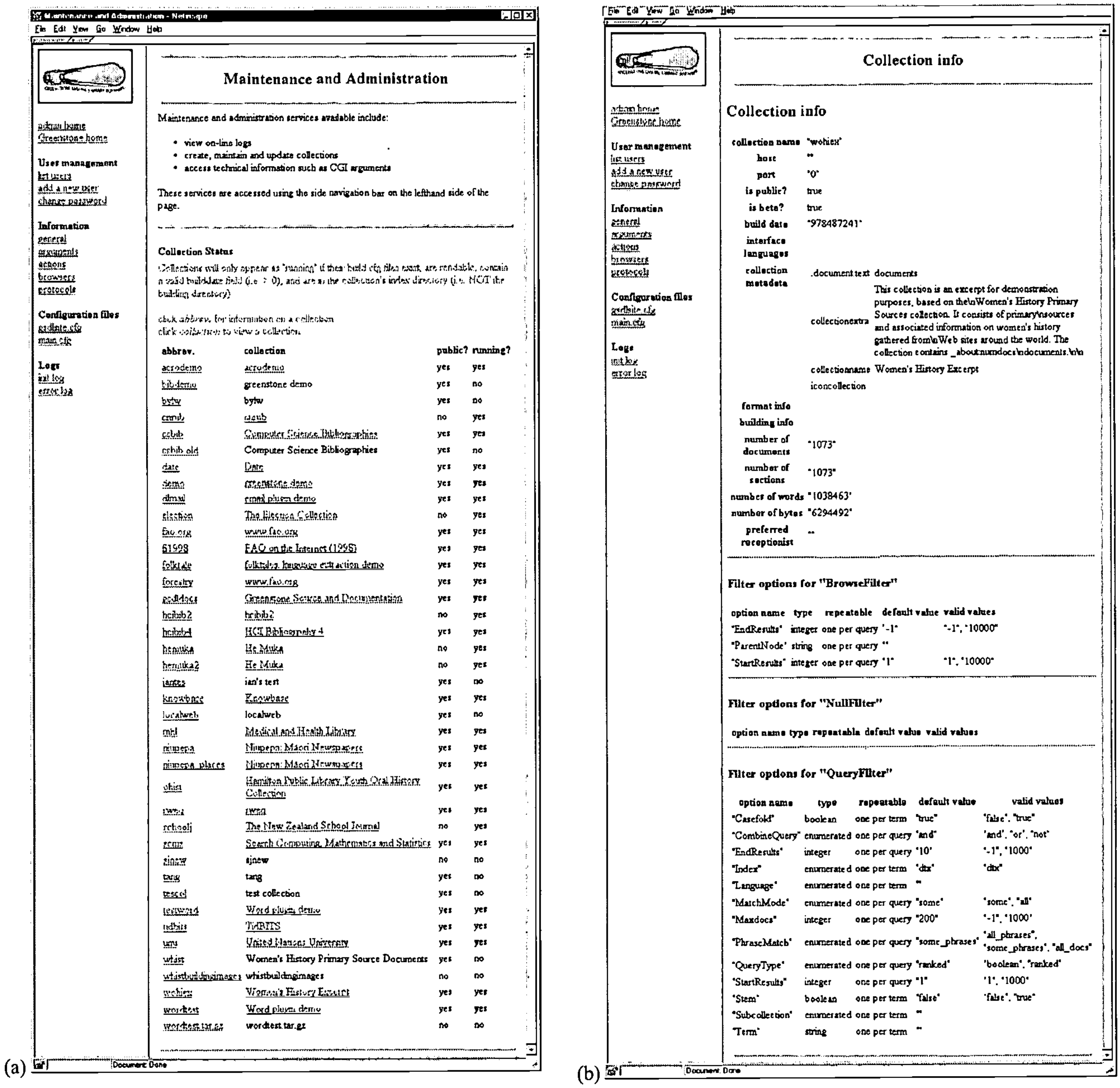
First, an internal name is chosen for the new collection, based on the title that has been supplied (and avoiding name clashes with existing collections). Then a directory structure is created for it that includes the necessary files to retrieve, index and present the source documents. To retrieve source documents already on the file system, a recursive file system copy command is issued; to retrieve offsite files a web mirroring package (we use *wget*[3]) is used to recursively copy the specified site along with any related image files.

Next, the documents are converted into XML. Appropriate plugins to perform this operation must be specified in the collection configuration file. This done, the copied files are deleted: the collection can always be rebuilt, or augmented and rebuilt, from the information stored in the XML files.

Then the full-text searching indexes, and the browsing structures, specified in the collection configuration file are created. Finally, assuming that the operation has been successful, the contents of the building process is moved to the area for active collections. This precaution ensures that if a version of this collection already exists, it continues to be served right up until the new one is ready. Use of global, persistent document identifiers ensures the changeover is almost always invisible to users.

The building stage is potentially very time-consuming. Small collections take a minute or so but large ones can take a day or more. The Web is not a supportive environment for this lengthy kind of activity. While the user can stop the building process

[3] See www.gnu.org



Maintenance and Administration

Maintenance and administration services available include:

- view on-line logs
- create, maintain and update collections
- access technical information such as CGI arguments

These services are accessed using the side navigation bar on the lefthand side of the page.

Collection Status

Collections will only appear as "running" if their build.cfg files exist, are readable, contain a valid builddate field (i.e. > 0), and are in the collection's index directory (i.e. NOT the building directory)

click *abbrev.* for information on a collection
click *collection* to view a collection

abbrev.	collection	public?	running?
acrodemo	acrodemo	yes	yes
bibdemo	greenstone demo	yes	no
bytw	bytw	yes	no
cmub	cmub	no	yes
csbib	Computer Science Bibliographies	yes	yes
csbib.old	Computer Science Bibliographies	yes	no
date	Date	yes	yes
demo	greenstone demo	yes	yes
dlmail	email plugin demo	yes	yes
election	The Election Collection	no	yes
fao.org	www.fao.org	yes	yes
fi1998	FAO on the Internet (1998)	yes	yes
folktale	folktales, language extraction demo	yes	yes
forestry	www.fao.org	yes	yes
gsdldocs	Greenstone Source and Documentation	yes	yes
hcibib2	hcibib2	no	yes
hcibib4	HCI Bibliography 4	yes	yes
hemuka	He Muka	no	yes
hemuka2	He Muka	no	yes
iantest	ian's test	yes	no
knowbase	Knowbase	yes	yes
localweb	localweb	yes	no
mhl	Medical and Health Library	yes	yes
niupepa	Niupepa: Maori Newspapers	yes	yes
niupepa_places	Niupepa: Maori Newspapers	yes	yes
ohist	Hamilton Public Library Youth Oral History Collection	yes	yes
rwsa	rwsa	yes	yes
schoolj	The New Zealand School Journal	no	yes
scms	Search Computing, Mathematics and Statistics	yes	yes
sjnew	sjnew	no	no
tang	tang	yes	no
testcol	test collection	yes	no
testword	Word plugin demo	yes	yes
tidbits	TidBITS	yes	yes
unu	United Nations University	yes	yes
whist	Women's History Primary Source Documents	yes	no
whistbuildingimages	whistbuildingimages	no	no
wohiex	Women's History Excerpt	yes	yes
wordtest	Word plugin demo	yes	yes
wordtest.tar.gz	wordtest.tar.gz	no	no

Collection info

Collection info

collection name	"wohiex"	
host	""	
port	"0"	
is public?	true	
is beta?	true	
build date	"978487241"	
interface languages		
collection metadata	.document text	documents
	collectionextra	This collection is an excerpt for demonstration purposes, based on the\nWomen's History Primary Sources collection. It consists of primary\nsources and associated information on women's history gathered from\nWeb sites around the world. The collection contains _aboutnumdocs\ndocuments.\n\n
	collectionname	Women's History Excerpt
	iconcollection	
format info		
building info		
number of documents	"1073"	
number of sections	"1073"	
number of words	"1038463"	
number of bytes	"6294492"	
preferred receptionist	""	

Filter options for "BrowseFilter"

option name	type	repeatable	default value	valid values
"EndResults"	integer	one per query	"-1"	"-1", "10000"
"ParentNode"	string	one per query	""	
"StartResults"	integer	one per query	"1"	"1", "10000"

Filter options for "NullFilter"

option name	type	repeatable	default value	valid values

Filter options for "QueryFilter"

option name	type	repeatable	default value	valid values
"Casefold"	boolean	one per term	"true"	"false", "true"
"CombineQuery"	enumerated	one per query	"and"	"and", "or", "not"
"EndResults"	integer	one per query	"10"	"-1", "1000"
"Index"	enumerated	one per term	"dtx"	"dtx"
"Language"	enumerated	one per term	""	
"MatchMode"	enumerated	one per query	"some"	"some", "all"
"Maxdocs"	integer	one per query	"200"	"-1", "1000"
"PhraseMatch"	enumerated	one per query	"some_phrases"	"all_phrases", "some_phrases", "all_docs"
"QueryType"	enumerated	one per query	"ranked"	"boolean", "ranked"
"StartResults"	integer	one per query	"1"	"1", "1000"
"Stem"	boolean	one per term	"false"	"false", "true"
"Subcollection"	enumerated	one per term	""	
"Term"	string	one per term	""	

Figure 3 The Greenstone Administration facility

immediately using the button in Figure 1f, there is no reliable way to prevent users from leaving the building page, and no way to detect if they do. In this case the Collector continues building the collection regardless and installs it when building terminates.

Progress is displayed in the status area at the bottom part of Figure 1f, updated every five seconds. The message visible in Figure 1f indicates that when the snapshot was taken, *Title* metadata was being extracted from an input file. Warnings are written if input files or URLs are requested that do not exist, or exist but there is no plugin that can process them, or the plugin cannot find an associated file, such as an image file embedded in a HTML document. The intention is that the user will monitor progress by keeping this window open in their browser. If any errors cause the process to terminate, they are recorded in this status area.

3.7 Viewing the collection

When the collection is built and installed, the sequence of buttons visible at the bottom of Figures 1a–e appears at the bottom of Figure 1f, with the *View collection* button active. This takes the user directly to the newly built collection.

Finally, email is sent to the collection's contact email address, and to the system's administrator, whenever a collection is created (or modified.) This allows those responsible to check when changes occur, and monitor what is happening on the system.

3.8 Working with existing collections

Four further facilities are provided when working with an existing collection: add new material, modify its structure, delete it, and write it to a self-contained, self-installing CD-ROM.

To add new material to an existing collection, the dialog structure illustrated in Figure 1 is used: entry is at the "source data" stage, Figure 1d. The new data that is specified is copied as before and converted to GML, joining any existing imported material.

Revisions of old documents should perhaps replace them rather than being treated as entirely new. However, this is so difficult to determine that all new documents are added to the collection unless they are textually identical to existing ones. While an imperfect process, in practice the browsing structures are sufficiently clear to make it straightforward to ignore near-duplicates. Recall, the aim of the Collector is to support the most common tasks in a straightforward manner—more careful updating is possible through the command line.

To modify the structure of an existing collection essentially means to edit its configuration file. If this option is chosen, the dialog is entered at the "configuring the collection" stage in Figure 1e.

Deleting a collection simply requires a collection to be selected from a list, and its deletion confirmed. This is not as foolhardy as it might seem, for only collections that were built by the Collector can actually be removed—other collections (typically built by advanced users working from the command line) are not included in the selection list. It would be nice to be able to selectively delete material from a collection through the Collector, but this functionality does not yet exist. At present this must be done from the command line by inspecting the file system.

Finally, in order to write an existing collection to a self-contained, self-installing CD-ROM, the collection's name is specified and the necessary files are automatically massaged into a disk image in a standard directory.

4. THE COLLECTION CONFIGURATION FILE

Part of the collection configuration file for the collection built in Figure 1 is shown in Figure 1e; it appears in full in Figure 2. Since we are in the process of updating the collection configuration file format to support a wider variety of services, we will not embark on a detailed explanation of what each line means.

Some of the information in the file (*e.g.* the email address at the top, the collection name and description near the bottom) was gathered from the user during the Collector dialog. In essence this is "collection-level metadata" and we are studying existing standards for expressing such information. The *indexes* line builds a single index comprising the text of all the documents. The *classify* line builds an alphabetic classifier of the title metadata.

The list of plugins is designed to be reasonably permissive. For example, ZIPPlug will uncompress any Zipped files; because plugins operate in a pipeline the output of this decompression will be available to the other plugins. GMLPlug ensures that any documents previously imported into the collection—stored in an XML format—will be processed properly when the collection is rebuilt. TEXTPlug, HTMLPlug and EMAILPlug process documents of the appropriate types, identified by their file extension. RecPlug (for "recursive") expands subdirectories and pours their contents into the pipeline, ensuring that arbitrary directory hierarchies are traversed.

More indicative of Greenstone's power than the generic structure in Figure 2 is the ease with which other facilities can be added. To choose just ten examples:

- A full-text-searchable index of titles could be added with one addition to the *indexes* line.
- If authors' names were encoded in the Web pages using the HTML metaname construct, a corresponding index of authors could also be added by augmenting the *indexes* line.
- With author metadata, an alphabetic author browser would require one additional *classify* line.
- WORD and/or PDF documents could be included by specifying the appropriate plugins
- Language metadata could be inferred by specifying an "extract-language" option to each plugin
- With language metadata present, a separate index could be built for document text in each language
- Acronyms could be extracted from the text automatically [9] and a list of acronyms added
- Keyphrases could be extracted from each document [2] and a keyphrase browser added
- A phrase hierarchy could be extracted from the full text of the documents and made available for browsing [5]
- The format of any of these browsers, or of the documents themselves when they were displayed, or of the search results list, could all be altered by appropriate "format" statements.

Skilled users could add any of these features to the collection by making a small change to the panel in Figure 1e. However, we do not anticipate that casual users will operate at this level, and provision is made in the Collector to by-pass this editing step. More likely, someone who wants to build new collections of a certain type will arrange for an expert to construct a prototype collection with the desired structure, and proceed to clone that into further collections with the same structure but different material.

5. SUPPORT FOR THE SYSTEM ADMINISTRATOR

An "administrative" facility is included with every Greenstone installation. The entry page, shown in Figure 3a, gives information about each of the collections offered by the system. Note that *all* collections are included—for there may be "private" ones that do not appear on the Greenstone home page. With each is given its short name, full name, whether it is publicly displayed, and whether or not it is running. Clicking a particular collection's abbreviation (the first column of links in Figure 3a) brings up information about that collection, gathered from its collection configuration file and from other internal structures created for that collection. If the collection is both public and running, clicking the collection's full name (the second link) takes you to the collection itself.

The collection we have just built has been named *wohiex*, for *Women's History Excerpt*, and is visible near the bottom of Figure 3a. Figure 3b shows the information that is displayed when this link is clicked. The first section gives some information from the configuration file, and the size of the collection (1000 documents, a million words, over 6 Mb). The next sections contain internal information related to the communication protocol through which collections are accessed. For example, the filter options for "QueryFilter" show the options and possible values that can be used when querying the collection.

The administrative facility also presents configuration information about the installation and allows it to be modified. It facilitates examination of the error logs that record internal errors, and the user logs that record usage. It enables a specified user (or users) to authorize others to build collections and add new material to existing ones. All these facilities are accessed interactively from the menu items at the left-hand side of Figure 3a.

5.1 Configuration files

There are two configuration files that control Greenstone's overall operation: the site configuration file *gsdlsite.cfg*, and the main configuration file *main.cfg*. The former is used to configure the Greenstone software for the site where it is installed. It is designed for keeping configuration options that are particular to a given site. Examples include the name of the directory where the Greenstone software is kept, the HTTP address of the Greenstone system, and whether the *fastcgi* facility is being used. The latter contains information that is common to the interface of all collections served by a Greenstone site. It includes the email address of the system maintainer, whether the status and collector pages are enabled, and whether cookies are used to identify users.

5.2 Logs

Three kinds of logs can be examined: usage logs, error logs and initialization logs. The last two are only really of interest to people maintaining the software. All user activity—every page that each user visits—can be recorded by the Greenstone software, though no personal names are included in the logs. Logging, disabled by default, is enabled by including an appropriate instruction in the main system configuration file.

Each line in the user log records a page visited—even the pages generated to inspect the log files! It contains (a) the IP address of the user's computer, (b) a timestamp in square brackets, (c) the CGI arguments in parentheses, and (d) the name of the user's browser (Netscape is called "Mozilla"). Here is a sample line, split and annotated for ease of reading:

```
   /fast-cgi-bin/niupepalibrary
(a) its-www1.massey.ac.nz
(b) [950647983]
(c) (a=p, b=0, bcp=, beu=, c=niupepa, cc=, ccp=0, ccs=0,
      cl=, cm=, cq2=, d=, e=, er=, f=0, fc=1, gc=0,
      gg=text, gt=0, h=, h2=, hl=1, hp=, il=1, j=, j2=,
      k=1, ky=, l=en, m=50, n=, n2=, o=20, p=home, pw=,
      q=, q2=, r=1, s=0, sp=frameset, t=1, ua=, uan=,
      ug=, uma=listusers, umc=, umnpw1=, umnpw2=, umpw=,
      umug=, umun=, umus=, un=, us=invalid, v=0, w=w,
      x=0, z=130.123.128.4-950647871)
(d) "Mozilla/4.08 [en] (Win95; I ;Nav)"
```

The last CGI argument, "z", is an identification code or "cookie" generated by the user's browser: it comprises the user's IP number followed by the timestamp when they first accessed the digital library.

5.3 User authentication

Greenstone incorporates an authentication scheme that can be employed to control access to certain facilities. It is used, for example, to restrict the people who are allowed to enter the Collector and certain administration functions. It also allows documents to be protected on an individual basis so that they can only be accessed by registered users on presentation of a password; however this is currently cumbersome to use and needs to be developed further. Authentication is done by requesting a user name and password as illustrated in Figure 1a.

From the administration page users can be listed, new ones added, and old ones deleted. The ability to do this is of course also protected: only users who have administrative privileges can add new users. It is also possible for each user to belong to different "groups". At present, the only extant groups are "administrator" and "colbuilder". Members of the first group can add and remove users, and change their groups. Members of the second can access the facilities described above to build new collections and alter (and delete) existing ones.

When Greenstone is installed, there is one user called *admin* who belongs to both groups. The password for this user is set during the installation process. This user can create new names and passwords for users who belong just to the *colbuilder* group, which is the recommended way of giving other users the ability to build collections.

User information is recorded in two databases that are placed in the Greenstone file structure. One contains all information relating to users. The other contains temporary keys that are created for each page access, which expire after half an hour. Thus inactive users must reauthenticate themselves.

5.4 Technical information

The links under the *Technical information* heading gives access to more technical information on the installation, including the directories where things are stored.

6. User Evaluation

The Collector and administration pages have been produced and refined through a long period of iterative design and informal testing. The design underwent many revisions before reaching the version presented in this paper. The details were thrashed out over several meetings of our digital library group—some 20 or so individuals from a variety of disciplines including library science, the humanities, and notably within computer science, the field of human computer interaction. It was here, for example, that the idea for the progress bar at the bottom of the page was formulated, and the very name of the tool, the Collector, was conceived. Once satisfied with its development, the tool was added into the public release of the Greenstone software.

Further feedback was obtained through the Greenstone mailing list, a general purpose listserver for Greenstone. In this arena both current issues and future developments are discussed, and users can gain technical assistance for particular problems. Filtering this source specifically for remarks about the Collector and administration pages revealed only technical questions—normally connected, despite extensive prior testing, with scripts performing incorrectly on a given version of a particular operating system. None of the questions fell into the category "how do I do this?" with the Collector. The technical questions indicate that the tools are being used, and the dearth of "how to" questions suggests that it is performing adequately.

Members of our group are presently conducting a usability study designed to establish more clearly how suitable the interface is for performing its intended tasks. Volunteers from a final year computer science class are observed performing set tasks: using both the tools described here and the original set of command line instructions. On completion of the tasks, which take about an hour, the users then fill out a questionnaire. The results from this work are not yet available.

The main difficulty of formal usability testing is that it is inevitably artificial. The tasks are artificial, the work environment is artificial, even the motivation of the users is artificial. Furthermore, the parameters of operation are far more tightly prescribed than in the real world. The question posed is how well does the tool let users perform the tasks it was designed for. But what if the user wants to step outside of the design parameters? On the one hand, an unfair question—the Collector is designed to simplify commonly executed tasks. But as users become more familiar with a tool they naturally expect more from it.

There questions are open ended, but vitally important. For their resolution we look to the open source nature of the Greenstone project. We will continue iteratively developing these tools based on feedback from field trials and user comments. We will also incorporate features added by others who actively develop the Greenstone code.

7. CONCLUSIONS

This paper has described the Collector, a tool integrated into the Greenstone environment that guides an end user through building and maintaining a collection, step by step. It also details the support included for the overall administration and maintenance of a Greenstone site by the system administrator, which is likewise integrated into the runtime system and gives the ability to view logs, create new users and control the access they have to, for example, the collection building tool.

Different users of digital libraries, naturally, have different needs. While access and retrieval is an obvious requirement, and dominates digital library research, we believe that end-user collection creation is another important element that deserves careful attention and further development. Including this capability in digital library systems will help them move away from the large and mostly static entities currently seen, and evolve into more dynamic and responsive environments.

8. REFERENCES

[1] Akscyn, R.M. and Witten, I.H. (1998) "Report on First Summit on International Cooperation on Digital Libraries." ks.com/idla-wp-oct98.

[2] Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C. and Nevill-Manning, C. (1999) "Domain-specific keyphrase extraction." *Proc Int Joint Conference on Artificial Intelligence*, Stockholm, Sweden. San Francisco, CA: Morgan Kaufmann Publishers, pp. 668–673.

[3] Gutwin, C., Paynter, G.W., Witten, I.H., Nevill-Manning, C. and Frank, E. (1999) "Improving browsing in digital libraries with keyphrase indexes." *Decision Support Systems* 27(1/2): 81–104; November.

[4] Lesk, M. (1997) *Practical digital libraries*. San Francisco, CA: Morgan Kaufmann.

[5] Paynter, G.W., Witten, I.H., Cunningham, S.J. and Buchanan, G. (2000) "Scalable browsing for large collections: a case study." *Proc Fifth ACM Conference on Digital Libraries*, San Antonio, TX, pp. 215–223; June.

[6] Witten, I.H., Moffat, A. and Bell, T.C. (1999) *Managing gigabytes: Compressing and indexing documents and images*, second edition. San Francisco, CA: Morgan Kaufmann.

[7] Witten, I.H., Loots, M., Trujillo, M.F. and Bainbridge, D. (2000) "The promise of digital libraries in developing countries."

[8] Witten, I.H., McNab, R.J., Boddie, S.J. and Bainbridge, D. (2000) "Greenstone: A comprehensive open-source digital library software system." *Proc Digital Libraries 2000*, San Antonio, Texas, pp. 113–121.

[9] Yeates, S., Bainbridge, D. and Witten, I.H. (2000) "Using compression to identify acronyms in text." *Proc Data Compression Conference*, edited by J.A. Storer and M. Cohn. IEEE Press Los Alamitos, CA, p. 582. (Full version available as Working Paper 00/1, Department of Computer Science, University of Waikato.)

Web-based Scholarship: Annotating the Digital Library

Bruce Rosenstock
Religious Studies
University of California
Davis, CA 95616, U.S.A.
bbrosenstock@ucdavis.edu

Michael Gertz
Department of Computer Science
University of California
Davis, CA 95616, U.S.A.
gertz@cs.ucdavis.edu

ABSTRACT

The DL offers the possibility of collaborative scholarship, but the appropriate tools must be integrated within the DL to serve this purpose. We propose a Web-based tool to guide controlled data annotations that link items in the DL to a domain-specific ontology and which provide an effective means to query a data collection in an abstract and uniform fashion.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Computer-supported cooperative work

Keywords

Data Annotations, Folk Literature DL

1. THE FOLK LITERATURE DL

The "Folk Literature of the Sephardic Jews"[1] digital library (DL) created at the University of California at Davis consists at present of about 1,000 (half the final total) XML files, conformant to the TEI (Text Encoding Initiative) DTD, and 500 audio files in MPEG3 format, averaging 20 MB each, representing 250 hours of various types of Judeo-Spanish oral literature–ballads, proverbs, wedding songs, memorates, and other genres–collected from dozens of informants over the past forty years. The XML files contain transcriptions of the audio files, with tagging elements dividing and structuring the spoken and sung portions of the material. The informants were Sephardic Jews (descendants of the Jews expelled from the Iberian Peninsula in 1492) who preserved their oral folk traditions over hundreds of years. The transmission of this traditional oral literature has, under the pressure of modern technological civilization, come to an end in this generation, and therefore this digital archive will provide a unique repository of an invaluable cultural treasure within the Jewish and the broader Hispanic world.

In the current project, access to the digital library will be Web-based and multi-tiered, with entry points for the general public and for specialists (medievalists, Romance philologists, folklorists, ethnomusicologists, and Jewish studies scholars). Users will be able to browse the archive for ballad types, informants, provenance and/or date of recording, and they will also be able to retrieve ballads through a search over the entire archive for specified words and phrases. Retrieved XML files are presented to the Web browser in HTML through server-side on-the-fly translation. Although much of the material has been studied and cataloged according to ballad types and motifs, there is a great deal of work remaining to be done. In order to permit not just access to the data set using standard keyword-based searches but the continued cataloging and enrichment of the data, we are designing an annotation tool to facilitate the creation of expert knowledge as metadata integrated within the logical structure of the digital library. This tool will be Web-based, allowing multiple users the selection of an HTML page or any node of its DOM structure as the target of an annotation.

2. CONTROLLED DATA ANNOTATIONS

Web-based annotation systems have been discussed and some implementations have been tested over the past six years since the 1995 W3C Workshop on the World Wide Web and Collaboration [6, 5]. These systems have understood an annotation to consist of unstructured and uncontrolled text which comments on the HTML page. We believe that unstructured textual annotation does not represent the appropriate method for creating new knowledge as an integrated component of the digital library, but that such annotations only add another layer of text on top of the existing data.

Instead, we have adopted an approach to annotation that guides the user to employ an abstract conceptualization of the application domain (ontology, controlled vocabulary). In such a domain conceptualization, concepts specific to the application domain (here folk literature of the Sephardic Jews) are defined, including concept properties and relationships. Concepts can be understood as class definitions and thus provide templates of semantically rich metadata schemes that can be associated with the data in the library. For example, the concept "ballad" has certain properties (ballad name, metre, provenance) and certain well-defined relationships to other concepts, e.g.; "motif", having motif name, actor-role, etc. as properties. Data in Web documents represent instances of concepts (as perceived by the user) and are assigned specific values for the properties.

There has been some previous work on ontology-based data annotation in the AI community (e.g., Ontobroker [3] or WebKB [4]). In these works, the major focus is on us-

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

ing an ontology to reason about the data in order to improve data query schemes, and not on how to efficiently populate the ontology with data. More importantly, data annotation occurs through embedding links to concepts in Web documents. That is, document authors have to include annotations in their documents. Naturally, such an author-centric approach to annotation prevents the enrichment of documents by multiple experts having different views and interpretations of the data.

In contrast, our primary focus is on how an ontology can be used to guide controlled data annotations that link portions of a Web page to concepts in an ontology. A domain specific ontology will be built in a collaborative fashion, providing users with the means to embed their domain knowledge in form of concepts and concept relationships. Our approach is unique in that data annotations are external to Web documents [2]. Multiple users can annotate the same data with different concepts, depending on their specific view and interpretation of the data. Annotations are stored separately from documents and can be queried in combination with the ontology to retrieve data that has been annotated using concepts from the ontology. In what follows, we will outline the architecture and functionality of the annotation tool in context of the folk literature digital library.

3. COMPONENTS OF THE DL

Figure 3 shows the basic components of our digital library.

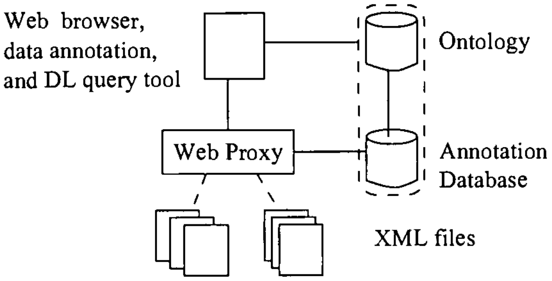


Figure 1: Components of the DL

At the backend, the XML files contain the encoded transcriptions of the informant's performance, as outlined in Section 1. Upon request from a client's Web browser, respective files are translated into HTML files on the fly. A simple keyword-based query interface allows retrieving documents that contain the specified keywords. In that respect, these components resemble standard components of a DL.

The ontology component contains the domain conceptualization. A Web-based interface allows registered users to add and modify concepts and concepts relationships. In the current prototype, we use RDF schema as a means to represent respective information about concepts, concept properties and concept relationships. A simple query interface allows a user to search for specific concepts (based on terms) and to traverse through the ontology.

In order to allow users to annotate data, it is necessary to use a Web proxy to load XML (HTML) documents into a browser. Upon activating a link that is added to a Web page by the proxy, a data annotation tool appears at the client site. The user then can highlight a region (paragraph, sentence, or whole document) of the retrieved document and choose a concept from the ontology, indicating that the portion of the document is an instance of the selected concept. In the annotation tool, the user also specifies the properties of the selected concept instance. The properties that can be specified depend on the respective concept template described in the ontology. Internally, an annotation consists of a pointer (URI) to a concept in the ontology, a pointer (XPath expression) to the selected data in the (XML) document, and concept instances properties. Each annotation is recorded in the annotation database. Keeping annotations external to document allows multiple users to annotate same (portions of) documents using different concepts.

Given a sufficient amount of annotations as semantically rich metadata associated with documents, the ontology together with the annotation database can be used to retrieve documents that have been annotated with user-specified concepts. Respective documents are loaded into a user's Web browser through the Web proxy which embeds "anchors" to the selected concept(s). Users can view annotated document and can traverse from annotations in documents to the ontology. View mechanisms can be specified on annotations, providing users only with annotations that have been made by, e.g., specific users, or that contain certain user-specified features, e.g., creation date or concept instance properties. Users can also browse the ontology and request documents that have been annotated with the selected concept. In this way, annotations provide a semantically rich layer of metadata on top of the original documents and thus facilitate better document retrieval schemes.

A prototypical data annotation tool utilizing a small (in terms of concepts) ontology has already been implemented. We are currently working on extending the tools to allow users to better manage and utilize the ontology. Other extensions include a registry in which users can register their own Web sites and pages and thus will be able to specify links among annotations between our digital library and their scholarly work.

4. REFERENCES

[1] S. G. Armistead, J. H. Silverman, and I. J. Katz. *Folk Literature of the Sephardic Jews, 2: Judeo-Spanish Ballads from Oral Tradition, 1: Epic Ballads.* University of California Press, 1986.

[2] J.M. Bremer and M. Gertz. Web data indexing through external semantic-carrying annotations. In *11th IEEE Int. Workshop on Research Issues in Data Engineering*, IEEE Computer Society, 69–76, 2001.

[3] S. Decker, M. Erdmann, D. Fensel, and R. Studer. Ontobroker: Ontology based access to distributed and semi-structured information. In *Semantic Issues in Multimedia Systems (DS-8)*, 351–369, Kluwer, 1999.

[4] J. Heflin and J. Hendler. Semantic interoperability on the Web. In *Proc. of Extreme Markup Languages 2000*, 111–120. Graphic Communications Association, 2000.

[5] M. Röscheisen, C. Mogensen, and T. Winograd. Beyond browsing: Shared comments, soaps, trails, and on-line communities. Technical Report STAN-CS-TR-97-1582, Stanford, Computer Science Department, 1995.

[6] R. Wilensky and T. A. Phelps. Multivalent documents: A new model for digital documents. Technical Report CSD-98-999, University of California, Computer Science Department, 1998.

A Multi-View Intelligent Editor for Digital Video Libraries

Brad A. Myers, Juan P. Casares, Scott Stevens, Laura Dabbish, Dan Yocum, Albert Corbett

Human Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
bam@cs.cmu.edu
http://www.cs.cmu.edu/~silver

ABSTRACT

Silver is an authoring tool that aims to allow novice users to edit digital video. The goal is to make editing of digital video as easy as text editing. Silver provides multiple coordinated views, including project, source, outline, subject, storyboard, textual transcript and timeline views. Selections and edits in any view are synchronized with all other views. A variety of recognition algorithms are applied to the video and audio content and then are used to aid in the editing tasks. The Informedia Digital Library supplies the recognition algorithms and metadata used to support intelligent editing, and Informedia also provides search and a repository. The metadata includes shot boundaries and a time-synchronized transcript, which are used to support intelligent selection and intelligent cut/copy/paste.

Keywords

Digital video editing, multimedia authoring, video library, Silver, Informedia.

1. INTRODUCTION

Digital video is becoming increasingly ubiquitous. Most camcorders today are digital, and computers are being advertised based on their video editing capabilities. For example, Apple claims that you can "turn your DV iMac into a personal movie studio" [1]. There is an increasing amount of video material available on the World-Wide-Web and in digital libraries. Many exciting research projects are investigating how to search, visualize, and summarize digital video, but there is little work on new ways to support the use of the video beyond just playing it. In fact, editing video is significantly harder than editing textual material. To construct a report or a new composition using video found in a digital library or using newly shot video requires considerably more effort and time than creating a similar report or composition using quoted or newly authored text.

In the Silver project, we are working to address this problem by bringing to video editing many of the capabilities long available in textual editors such as Microsoft Word. We are also trying to alleviate some of the special problems of video editing. In particular, the Silver video editor provides multiple views, it uses the familiar interaction techniques from text editing, and it provides intelligent techniques to make selection and editing easier.

The Silver editor is designed to support all phases of the video post-production process. The storyboard and script views support brainstorming and planning for the video. The project, source, subject and outline views support the collection and organization of the source material. The timeline and script views support the detailed editing, and the preview view can be used to show the result.

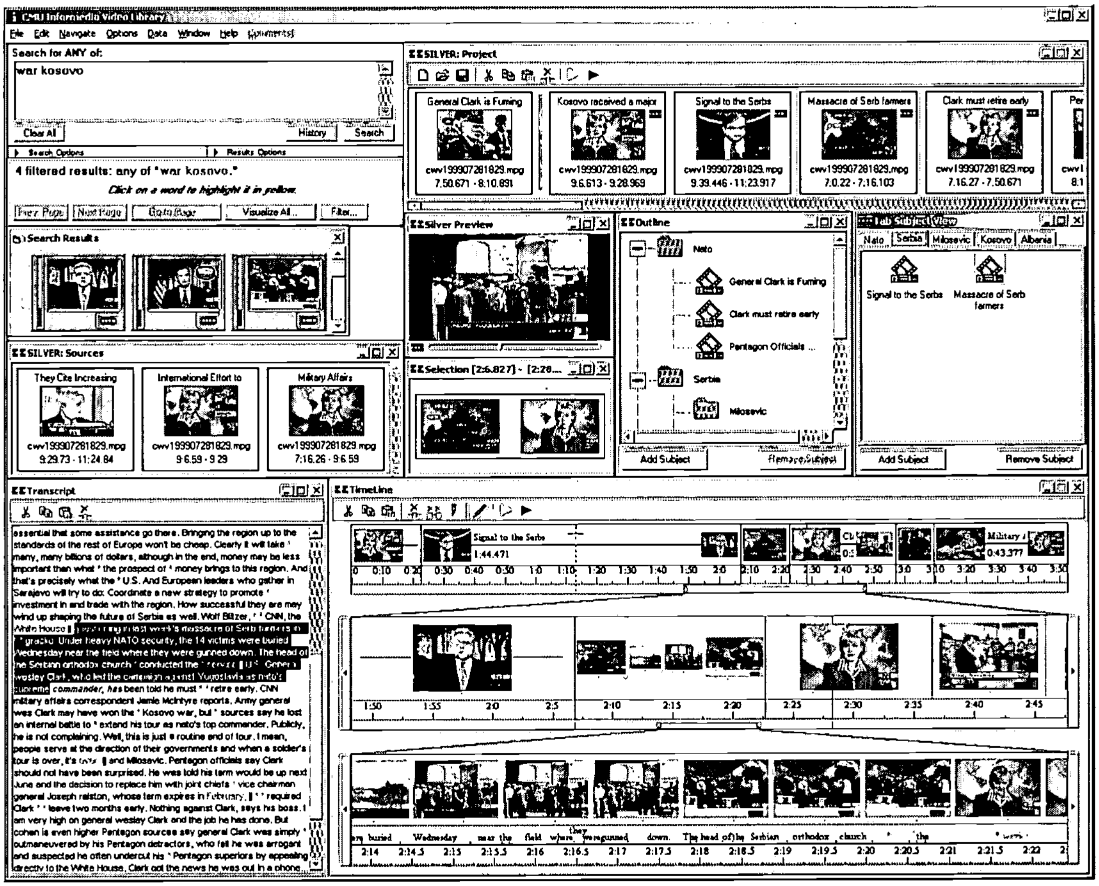
Silver is an acronym and stands for Simplifying Interactive Layout and Video Editing and Reuse. The key innovations in the Silver editor include: providing a transcript view for the actual audio; multiple views with coordinated selections, including the ability to show when one view only contains part of the selection; intelligent context-dependent expansion of the selection for double-clicking; and intelligent cut/copy/paste across the video and audio. These are discussed in this article.

2. STATE OF THE ART

Most tools for editing video still resemble analog professional video editing consoles. Although they support the creation of high quality material, they are not easy for the casual user, especially when compared with applications such as text editors. Current video editing software only operates at a low syntactic level, manipulating video as a sequence of frames and streams of uninterpreted audio. It does not take advantage of the content or structure of the video or audio to assist in the editing. Instead, users are required to pinpoint specific frames, which may involve zooming and numerous repetitions of fast-forward and rewind operations. In text editing, the user can search and select the content by letters, words, lines, sentences, paragraphs, sections, etc. In today's video and audio editing, the only units are low-level frames or fractions of seconds.

As another example, three-point editing is one option in professional editors such as Adobe Premiere. In this kind of editing, the user can locate an "in point" and an "out point"

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.



Search for ANY of:

war kosovo

4 filtered results: any of "war kosovo."

Click on a word to highlight it in yellow.

essential that some assistance go there. Bringing the region up to the standards of the rest of Europe won't be cheap. Clearly it will take ' many, many billions of dollars, although in the end, money may be less important than what ' the prospect of ' money brings to this region. And that's precisely what the ' U.S. And European leaders who gather in Sarajevo will try to do: Coordinate a new strategy to promote ' investment in and trade with the region. How successful they are may wind up shaping the future of Serbia as well. Wolf Blitzer, ' ' CNN, the White House ‖ [illegible] in last week's massacre of Serb farmers in ' ' gracko. Under heavy NATO security, the 14 victims were buried Wednesday near the field where they were gunned down. The head of the Serbian orthodox church ' conducted the ' service ‖ U.S. General wesley Clark, who led the campaign against Yugoslavia as nato's supreme *commander, has* been told he must ' ' retire early. CNN military affairs correspondent Jamie McIntyre reports. Army general wes Clark may have won the ' Kosovo war, but ' sources say he lost an internal battle to ' extend his tour as nato's top commander. Publicly, he is not complaining. Well, this is just a routine end of tour. I mean, people serve at the direction of their governments and when a soldier's tour is over, it's [illegible] ‖ and Milosevic. Pentagon officials say Clark should not have been surprised. He was told his term would be up next June and the decision to replace him with joint chiefs ' vice chairman general Joseph ralston, whose term expires in February, ‖ ' ' required Clark ' ' leave two months early. Nothing against Clark, says his boss. I am very high on general wesley Clark and the job he has done. But cohen is even higher Pentagon sources say general Clark was simply ' outmaneuvered by his Pentagon detractors, who felt he was arrogant and suspected he often undercut his ' Pentagon superiors by appealing directly to the White House. Clark got the news he was out in a phone.

Figure 1. An overview of all of the Silver windows.

in the video source and a third point in the target to perform a copy operation. The fourth point for the edit is computed based on the length of the in-out pair. This method can be traced to the use of physical videotape where the length of the output and input segments must be the same. However, three-point editing is not required for digital video and is very different from the conventional cut/copy/paste/delete technique that is used in other editing tools on computers.

3. RELATED WORK

There is a large body of work on the extraction and visualization of information from digital video (e.g., [29] [12]). However, most of this work has focused on automatic content extraction and summarization during library creation and searching, and on information presentation during library exploration. In the Silver project, we focus on authoring with the content once it is found.

After examining the role of digital video in interactive multimedia applications, Mackay and Davenport realized that video could be an information stream that can be tagged, edited, analyzed and annotated [19]. Davenport et. al. proposed using metadata for home-movie editing assistance [10]. However, they assumed this data would be obtained through manual logging or with a "data camera" during filming, unlike the automatic techniques used in Silver.

The Zodiac system [5] employs a branching edit history to organize and navigate design alternatives. It also uses this abstraction to automatically detect shot and scene boundaries and to support the annotation of moving objects in the video. IMPACT [33] uses automatic cut detection and camera motion classification to create a high level description of the structure of the video, and then visualizes and edits the structure using timeline and tree structure views [32]. IMPACT also detects object boundaries and can recognize identical objects in different shots.

The Hierarchical Video Magnifier [24] allows users to work with a video source at fine levels of detail while maintaining an awareness of the context. It provides a timeline to represent the total duration of the video source, and supplies the user with a series of low-resolution frame samples. There is also a tool that can be used to expand or reduce the effective temporal resolution of any portion of the timelines. Successive applications of the temporal magnifier create an explicit spatial hierarchical structure of the video source. The Swim Hierarchical Browser [35] improves on this idea by using automatically detected shots in the higher level layers. These tools were only used for top-down navigation, and not for editing. We use a similar approach in our Timeline view.

BEST COPY AVAILABLE

The Video Retrieval and Sequencing System [8] semiautomatically detects and annotates shots for later retrieval. Then, a cinematic rule-based editing tool sequences the retrieved shots for presentation within a specified time constraint. For example, the parallel rule alternates two different sets of shots and the rhythm rule selects longer shots for a slow rhythm and shorter shots for a fast one.

Most video segmentation algorithms work bottom-up, from the pixels in individual frames. Hampapur [15] proposes a top-down approach, modeling video-editing techniques mathematically to detect cuts, fades and translation effects between shots.

Video Mosaic [20] is an augmented reality system that allows video producers to use paper storyboards as a means of controlling and editing digital video. Silver's storyboards are more powerful since they can also be used for interactive videos. CVEPS [23] automatically extracts key visual features and uses them for browsing, searching and editing. An important contribution of this system is that it works in the compressed domain (MPEG), which has advantages in terms of storage, speed and noiseless editing.

VideoScheme [22] is a direct manipulation video editing system that provides the user with programming capabilities. This increases the flexibility and expressiveness of the system, for example supporting repetitive or conditional operations. VideoMAP [31] indexes video through a variety of image processing techniques, including histograms and "x-rays" (edge pixel counts). The resulting indices can be used to detect cuts and camera operations and to create visualizations of the video. For example, VideoMAP renders the indices over time, and VideoSpaceIcon represents the temporal and spatial characteristics of a shot as an icon.

The Hitchcock system [14] automatically determines the "suitability" of the different segments in raw video, based on camera motion, brightness and duration. Similar clips are grouped into "piles." To create a custom video, the user drags segments into a storyboard, specifies a total desired duration and Hitchcock automatically selects the start and end points of each clip based on shot quality and total duration. Clips in the storyboard are represented with frames that can be arranged in different layouts, such as a "comic book" style layout [2]. We plan to incorporate similar techniques into Silver.

4. INFORMEDIA

We obtain our source video and metadata through CMU's Informedia Digital Video Library [34]. The Informedia project is building a searchable multimedia library that currently has over 2,000 hours of material, including documentaries and news broadcasts. Informedia adds about two hours of additional news material every day. For all of its video content, Informedia creates a textual transcript of the audio track using closed-captioning information and speech recognition [7]. The transcript is time-aligned with the video using CMU's Sphinx speech recognition system [27]. Informedia also performs image analysis to detect shot boundaries, extracting representative thumbnail images from each shot [6] and detects and identifies faces in the video frame. A video OCR system identifies and recognizes captions in the image [28]. Certain kinds of camera movements such as pans and fades can also be identified. All of this metadata about the video is stored in a database. This metadata is used by Informedia to automatically create titles, representative frames and summaries for video clips, and to provide searching for query terms and visualization of the results. Silver takes advantage of the metadata in the database to enhance its editing capabilities.

5. TYPES OF PRODUCTIONS

The Silver video editor is designed to support different kinds of productions. Our primary goal is to make it easier for middle and high school children (ages 10-18) to create multimedia reports on a particular topic. For example, a social-studies teacher might have students create a report on Kosovo. We want to make it as easy for students to create such a video report using material in the Informedia library, as it would be to use textual and static graphical material from newspaper and magazine articles.

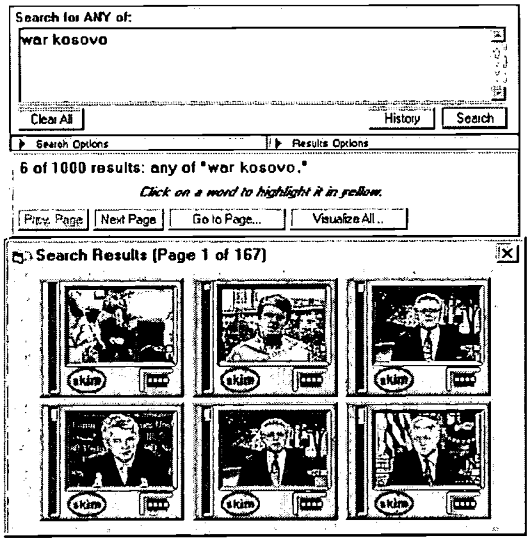
We also want Silver to support original compositions of two types. First, people might just shoot some video with a camcorder, and then later want to edit it into a production. In the second type, there might first be a script and even a set of storyboards, and then video is shot to match the script. In these two cases where new material is shot, we anticipate that the material will be processed by Informedia to supply the metadata that Silver needs. (Unfortunately, Informedia does not yet give users the capability to process their own video, but Silver is being designed so that we are ready when it does.)

Finally, in the future, we plan for Silver to support *interactive* video and other multi-media productions. With an interactive video, the user can determine what happens next, rather than just playing it from beginning to end. For example, clicking on hot spots in "living books" type stories may choose which video is played next. Another type of production is exemplified by the DVD video "What could you do?" [11]. Here, a video segment is played to set up a situation, then the user is asked a question and depending on the user's answer, a different video piece is selected to be played next.

6. MULTIPLE VIEWS

In order to support many different kinds and styles of editing, it is useful to have different views of the material. For example, the Microsoft Word text editor supplies an outline view, a "normal" view that is good for editing, a "print layout" view that more closely shows what the composition will look like, and a "print preview" view that tries to be exactly like the printed output. Similarly, PowerPoint provides outline, normal, notes, slide sorter, and slide show views. However, video editors have many fewer options. Adobe Premiere's principal view is a frame representation, with a thumbnail image representing one or more video frames. Premiere also provides a Project window, a Timeline window, and a Monitor window (to playback video). MGI's VideoWave III editor has a Library window (similar to the project view), a StoryLine window (a simplified timeline), and a View Screen window for playback. Apple's Final Cut Pro provides a Browser window (like a project view that allows clip organization), Timeline view, Canvas (editing palette), and Viewer window for playback.

Silver currently provides nine different views: the Informedia search and search results, project, source, subject, outline, storyboard, transcript, timeline and preview views. These are described below. Each appears in its own window, which the user or system can arrange or hide if not needed (Figure 1 shows an overview of the full screen).



6 of 1000 results: any of 'war kosovo.'

Click on a word to highlight it in yellow.

Figure 2. Search and search results views from Informedia.

6.1 Search Results View

When starting from a search using Informedia, the search results will appear in an Informedia search results window (see Figure 2). Informedia identifies *clips* of video that are relevant to the search term, and shows each clip with a representative frame. Each clip will typically contain many different scenes, and the length of a clip varies from around 40 seconds to four minutes. The user can get longer segments of video around a clip by moving up a level to the full video. If the user makes a new query, then the search results window will be erased and the new results will appear instead. Clicking on the thumbnail image will show the clip using the Windows Media Player window. Clicking on the filmstrip icon displays thumbnail images from each shot in the scene, giving a static, but holistic view of the entire clip.

6.2 Source and Project Views

When the user finds appropriate video by searching in Informedia, the clips can be dragged and dropped into Silver's Project View (actually, a clip can be dragged directly to any other view, and it will be automatically added to the project view). The Project View (see Figure 3) will also allow other video, audio and still pictures from the disk or the World-Wide Web to be loaded and made easily available for use in the production. As in other video editors such as Premiere, the project view is a "staging area" where the video to be used in a production can be kept. However, in the Silver project view, video clips that are in the composition are separated by a line from clips that are not currently in use. Dragging a clip across this line adds or removes it from the composition.

The source view is like a simplified project view, and allows easy access to the original sources of video. Clips in this view represent the video as brought in from Informedia, from a file, or from the web. They cannot be edited, but they may be dragged to other windows to add a copy of the source to the project.

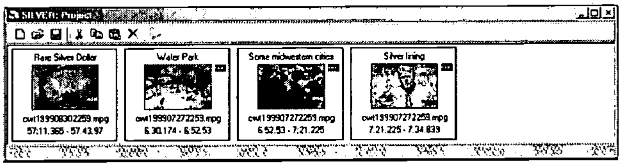


Figure 3. Silver's project view collects the source material to be used in the production. The first clip is shown on the screen outlined in light yellow to indicate it is partially selected.

6.3 Transcript View

An important innovation in the Silver video editor, enabled by Informedia, is the provision of a textual transcript of the video. This is displayed in a conventional text-editor-like window (see Figure 4). Informedia generates the transcript from various sources [16]. If the video provides closed captioning, then this is used. Speech recognition is used to recognize other speech, and also to align the transcript with the place in the audio track where each word appears [7]. Because the recognition can contain mistakes, Silver inserts green "*"s where there appears to be gaps, misalignments, or silence. In the future, we plan to allow the user to correct the transcript by typing the correct words, and then use the speech recognizer to match the timing of the words to the audio track.

The transcript view and the timeline view (section 6.4) are the main ways to specify the actual video segments that go into the composition. In the transcript view, the boundary between segments is shown as a blue double bar ("||"). The transcript and timeline views are used to find the desired portion of each clip. Transcripts will also be useful in supporting an easy way to search the video for specific content words.

We also plan to use the transcript view to support the authoring of new productions from scripts. The user could type or import a new script, and then later the system would automatically match the script to the audio as it is shot.

6.4 Timeline View

In Silver, like many other video editors such as Premiere, the Timeline view is the main view used for detailed editing. In order to make the Timeline view more useful and easier to use, we are investigating some novel formats. As shown in Figure 5, we are currently providing a three-level view.

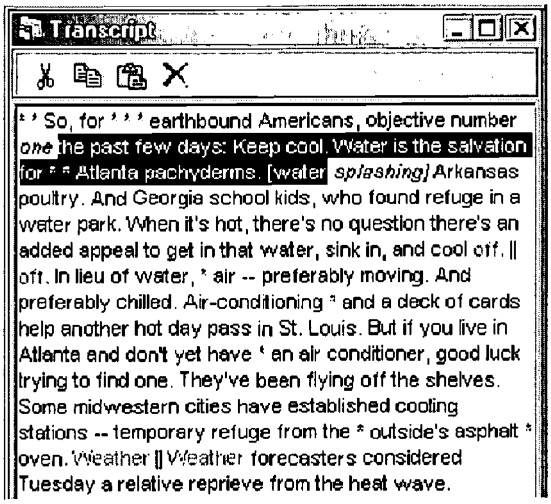
This allows the user to work at a high level of detail without losing the context within the composition. The top level always represents the entire video. The topmost row displays the clips in the composition and their boundaries. For each clip, it shows a representative frame, and if there is enough space, the end frame, title and duration. Below the top row are the time codes. At the bottom of the top level is an indicator showing what portions are being viewed in the bottom two levels. The purple portion is visible in the middle level, and the cyan portion is visible in the bottom level.

The middle level displays the individual shots, as detected by Informedia. Shot boundaries are detected by a change in the video [17]. Each shot is visualized using the representative frame for the shot as chosen by Informedia. The size of the frame is proportional to the duration of the shot.

The bottom level can display the individual frames of the video, so the user can quickly get to particular cut points. The middle row of the bottom level represents the transcript. The bottom level also provides the ability to add annotations or comments to the video (see Figure 6).

A key feature of the Silver timeline is that it allows different representations to be shown together, allowing the user to see the clip boundaries, the time, samples of frames, the cuts, the transcript, annotations, etc. Later, using facilities already provided by Informedia, we can add labels for recognized faces and the waveform for the audio to the timeline. Snapping and double-click selection will continue to be specific to each type of content.

The user can pick what portion is shown in the middle level by dragging the indicator at the bottom of the top level. Alternatively, the scroll buttons at the edges of the middle level cause the viewed section to shift left and right. The scale of the video that is shown in the middle level can be adjusted by changing the size of the indicator at the bottom of the top level, by dragging on the edge of the indicator. This will zoom the middle level in and out. Similarly, the bottom level can be scrolled and zoomed using the indicator at the bottom of the middle level or the bottom's scroll arrows.



* * So, for * * * earthbound Americans, objective number *one* the past few days: Keep cool. Water is the salvation for * * Atlanta pachyderms. [water *splashing]* Arkansas poultry. And Georgia school kids, who found refuge in a water park. When it's hot, there's no question there's an added appeal to get in that water, sink in, and cool off. || off. In lieu of water, * air -- preferably moving. And preferably chilled. Air-conditioning * and a deck of cards help another hot day pass in St. Louis. But if you live in Atlanta and don't yet have * an air conditioner, good luck trying to find one. They've been flying off the shelves. Some midwestern cities have established cooling stations -- temporary refuge from the * outside's asphalt * oven. Weather || Weather forecasters considered Tuesday a relative reprieve from the heat wave.

Figure 4. The Transcript view shows the text of the audio. The "*"s represent unrecognized portions of the audio, and the "||" represent breaks. The reverse video part at the top is selected. Words in italics are partially selected (here "splashing") and words in gray are partially cut out in the current production.

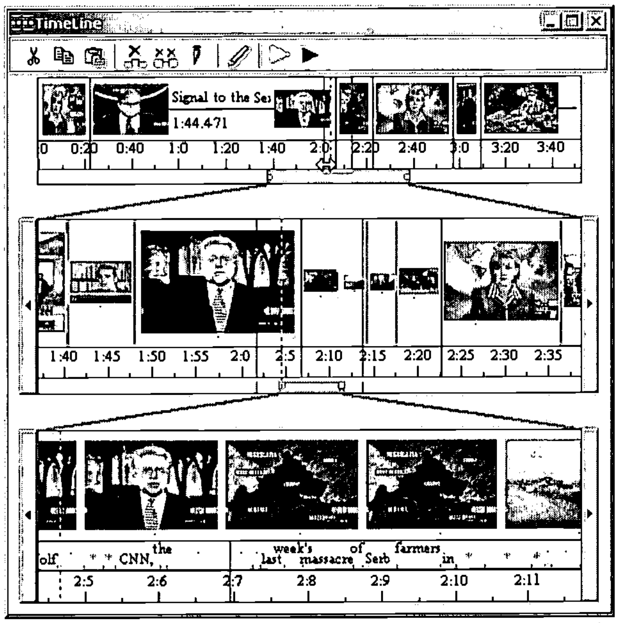


Figure 5. The Timeline view showing the three levels. The portion from about 2:2 to about 2:13 is selected, and appears on the screen in yellow in all three levels (and appears as light gray when printed in black-and-white).

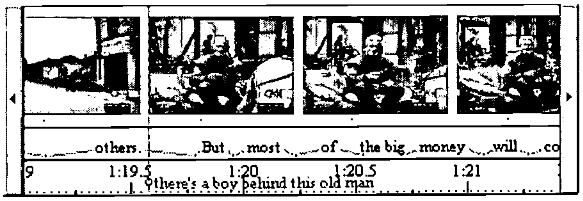


Figure 6. The bottom row of the Timeline view can show the user's annotations (shown in red).

BEST COPY AVAILABLE

The toolbar buttons at the top of the Timeline view window perform editing and playback operations (from left to right in Figure 5): cut, copy, paste, delete, crop (deletes everything but the selection), split (splices the clip at the selection edges), add annotation, play selection, and play the entire video.

Lee, Smeaton, *et. al.* [18] propose a taxonomy of video browsers based on three dimensions: the number of "layers" of abstraction and how they are related, the provision or omission of temporal information (varying from full timestamp information to nothing at all), and the visualization of spatial versus temporal aspects of the video (a slideshow is highly temporal, a timeline highly spatial). They recommend using many linked layers, providing temporal and absolute temporal information, and a spatial visualization. Our timeline follows their recommendations.

The Hierarchical Video Magnifier [24] and Swim [35] also provide multi-level views. These systems are designed to browse video, and navigation is achieved by drilling to higher levels of detail. The goal in Silver is to edit video and the basic interaction for navigating in the timeline is scrolling. Also, Silver is different in using multiple representations of the video within each level.

6.5 Preview View

As the user moves the cursor through the timeline, Silver displays a dotted line (visible to the left of 2.5 in all three levels of Figure 5). The frame at this point is shown in the preview view (Figure 7). If the user moves the cursor too fast, the preview view will catch up when the user stops or slows down sufficiently. The play arrows at the top of the timeline view cause the current video to be played in the preview view. If the black arrow is selected, the video is played in its entirety (but the user can always stop the playback). If the yellow play button is picked, only the selected portion of the video is played. The preview window is implemented using the Windows Media Player control.

6.6 Subject View

When creating a composition, different people have different ways of organizing their material. Some might like to group the material by topic. Silver's Subject View facilitates this type of organization. It provides a tabbed dialog box into which the material can be dragged-and-dropped from the project view. The user is free to label the tabs in any way that is useful, for example by the content of clip, the type of shot, the date, etc. The subject view (Figure 8) will allow the same clip to be entered multiple times, which will help users to more easily find material, since it might be classified in multiple ways.

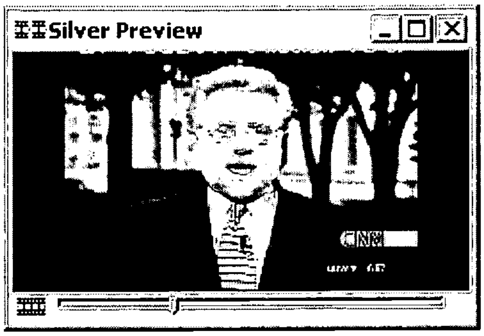


Figure 7. The Preview view shows the frame at the cursor (the dotted lines in Figure 5), and is where the video is played.

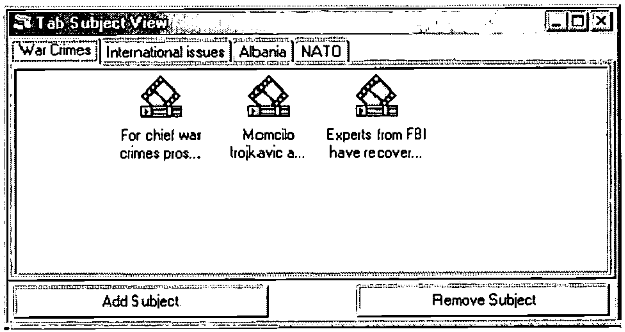


Figure 8. Silver's Subject Views allows users to organize their material by topic, type, date, etc.

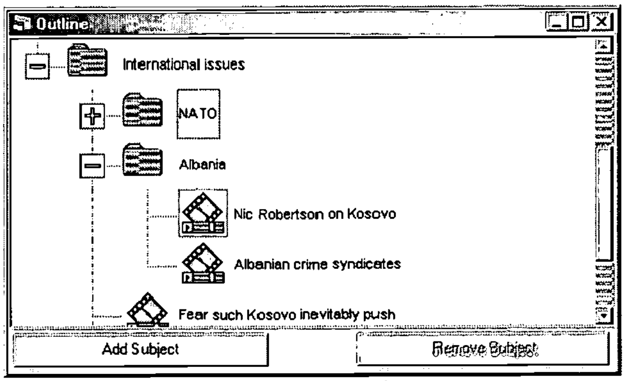


Figure 9. Silver's Outline View organizes the material using a Window's Tree control.

6.7 Outline View

When creating a composition, one good way to organize the material is in an outline. Whereas textual editing programs, such as Microsoft Word, have had outlining capabilities for years, none of the video editors have an outline view. Silver's outline view (shown in Figure 9) uses a conventional Windows tree control, which allows the hierarchy to be easily edited using familiar interaction techniques such as drag-and-drop. Note that for the subject view and the outline view, the subjects (or folders) can be added before there are any clips to put in them, to help guide the process and serve as reminders of what is left to do.

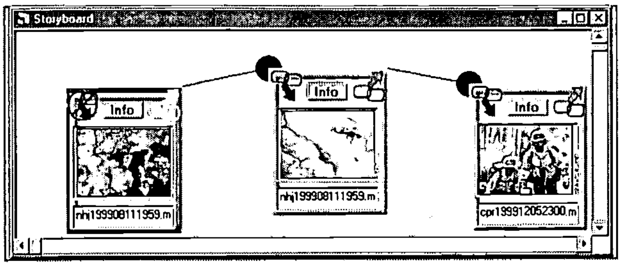


Figure 10. The Storyboard view has segments placed in 2-D.

6.8 Storyboard View

Many video and cinema projects start with a storyboard drawing of the composition, often drawn on paper. Typically, a picture in the storyboard represents each of the major scenes or sequences of the composition. Some video editors, notably MGI's VideoWave III, use a storyboard-like view as the main representation of the composition. Silver's storyboard view (see Figure 10) differs from VideoWave in that it can be used *before* the clips are found, as a representation of the desired video. Stills or even hand-drawn pictures can be used as placeholders in the storyboard for video to be shot or found later. The frames in the storyboard can be hand-placed in two dimensions by the user (and commands will help to visually organize them), which supports organizations that are meaningful to the user. For example, some productions are told using "parallel editing" [3] by cutting between two different stories occurring at the same time (for example, most *Star Wars* movies cut repeatedly between the story on a planet and the story in space). These might be represented in the storyboard by two parallel tracks.

Another important use for storyboards will be *interactive* video compositions (which Silver is planned to support in the future). Some multimedia productions allow the user to interact with the story using various methods to pick which video segment comes next. For example, a question might be asked or the user might click on various hot spots. Our storyboard view allows multiple arrows out of a clip, and we plan to support a "natural" scripting language [26] and demonstrational techniques [25] that will make it easy to specify how to choose which segment to play next based on the end user's input.

6.9 Other Views

In the future, we plan to add support for many other views, all inter-linked. For example, if the transcript window is used to hold an authored script, then it will be important to include "director's notes" and other annotations. These might be linked to views that help manage lists of locations, people, scenery, and to-do items. The ability to add notes, comments, annotations, and WWW links in all other views might also be useful. Other facilities from text documents might also be brought into the Silver editor, such as the ability to compare versions and keep track of changes (as a revision history).

6.10 Selection Across Multiple Views

When the user selects a portion of the video in one view in Silver, the equivalent portion is highlighted in all other views. This brings up a number of interesting user interface design challenges.

The first problem is what is the "equivalent" portion? The different views show different levels of granularity, so it may not be possible to represent the selection accurately in some views. For example, if a few frames are selected in the timeline view what should be shown in the project view since it only shows complete clips? Silver's design is to use dark yellow to highlight the selection, but to use light yellow to highlight an item that is only partially selected. In Figure 3, the first clip has only part of its contents selected, so it is shown in light yellow. If the user selects a clip in the project view, then all video that is derived from that clip is selected in all other views (which may result in discontinuous selections).

A similar problem arises between the timeline and transcript views. A particular word in the audio may span multiple frames. So selecting a word in the transcript will select all the corresponding frames. But selecting only one of those frames in the video may correspond to only part of a word, so the highlight in the transcript shows this by making that word *italic*. This is the case of the words "one" and "splashing" shown in the edges of the selected text in Figure 4. (We would prefer the selection to be yellow and the partially selected words in light yellow to be consistent with other views, but the Visual Basic text component does not support this.) If the selected video is moved, this will cut the word in two pieces. Silver represents this by repeating the word in both places, but showing it in a different font color. The video in Figure 4 was split somewhere during the word "Weather". Thus, this word is shown twice, separated by the clip boundary.

7. INTELLIGENT EDITING

One reason that video editing is so much more tedious than text editing is that in video, the user must select and operate on a frame-by-frame basis. Simply moving a section of video may require many minutes while the beginning and end points are laboriously located. Often a segment in the video does not exactly match with the corresponding audio. For example, the voice can start before the talking head is shown. This gives the video a continuous, seamless feel, but makes extracting pieces much harder because the video and audio portions must often be separately adjusted, with much fiddling to remove extraneous video or audio portions.

We did an informal study to check on the prevalence of these issues, and found it to indeed be significant, at least in recorded news. In looking at 238 transitions in 72 minutes of video clips recorded from CNN by Informedia, 61 (26%) were "L-cuts," where the audio and video come in or stop at different times. Of the 61 L-cuts, 44 (72%) were cases where the audio of the speaker came in before the video but ended at the same time (see Figure 11-Shot A). Most of these were interviews where the voice of the person being interviewed would come in first and then the video of the person after a few seconds. Most of the other ways to overlap the video and audio were also represented. Sometimes, the audio and the video of the speaker came in at the same time, but the audio continued past the video (Figure 11-Shot B) or the video ended after the audio (Shot D). When an interviewee was being introduced while he appeared on screen, the video might come in before the audio, but both end at the same time (Shot C). To copy or delete any of these would require many steps in other editors.

The Silver editor aims to remove much of the tedium associated with editing such video by automatically adjusting the portions of the video and audio used for selection, cut, copy and paste, in the same way that text editors such as Microsoft Word adjust whether the spaces before and after words are selected. These capabilities are discussed in the following sections.

7.1 Intelligent Selection

When the user double-clicks in Microsoft Word and other text editors, the entire word is selected. Triple clicking will get the entire paragraph, sentence or line (depending on the editor). Silver provides a similar feature for video, and the unit selected on multiple clicks depends on which view is active. If the user double-clicks in the text view, the surrounding word or phrase will be selected that Informedia recognized (the minimal unit that can be matched with the audio). In the time-line view, however, the effect of double clicking depends on the specific timeline within the hierarchy. It can mean, for example, a shot (where shot boundaries are located automatically by Informedia) or an entire clip.

7.2 Intelligent Cut, Delete and Copy

When an operation is performed on the selected portion of the video, Silver uses heuristics to try to adjust the start and end points so they are appropriate for both the video and audio.

Using the information provided by Informedia, Silver will try to detect the L-cut situations, as shown in Figure 11. Silver will then try to adjust the selection accordingly. For example, when the user selects a sentence in the transcript and performs a copy operation, Silver will look at the corresponding video. If the shot boundary is not aligned with the selection from the audio, then Silver will try to determine if there is an L-cut as in Figure 11. If so, Silver will try to adjust the selected portion appropriately. However, editing operations using this selection will require special considerations, as discussed next.

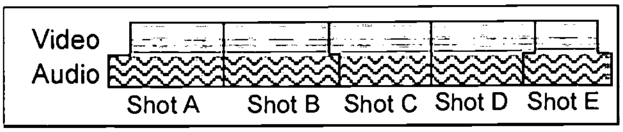


Figure 11. It is very common for the video and audio of a shot not to start and/or end at the same time. For example, Shot B represents a situation where the audio for the shot continues for a little while past when the video has already switched to the next scene. Similarly, for Shot D, the video continues a bit past when the audio has already switched to the next scene.

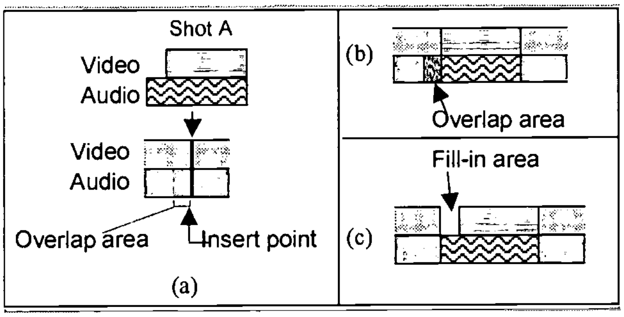


Figure 12. When an L-shot such as Shot A is inserted at a point in the video (a), Silver will check the area that might be overlapped. If the audio is silent in that area, Silver will overlap the audio automatically (b). In some cases (c), the user will have the option to overlay a separate piece of video if the audio cannot be overlapped.

7.3 Intelligent Paste and Reattach

When a segment with an L-cut is deleted or pasted in a new place, then Silver will need to determine how to deal with the uneven end(s). If the audio is shorter than the video (e.g., if Shots C or D from Figure 11 are pasted), then Silver can fill in with silence since this is not generally disruptive. However, it is not acceptable to fill in with black or blank video. For example, if Shot A is pasted, Silver will look at the overlapped area of the audio to see if it is silent or preceeded by an L-cut (Figure 12-a). If so, then the video can abut and the audio can be overlapped automatically (shown in Figure 12-b). If the audio in the overlap area is *not* silent, however, then Silver will suggest options to the user and ask the user which option is preferred. The choices include that the audio in the destination should be replaced, the audio should be mixed (which may be reasonable when one audio track is music or other background sounds), or else some video should be used to fill in the overlap area (as in Figure 12-c). The video to be filled in may come from the source of the copy (e.g., by expanding the video of Shot A) or else may be some other material or a "special effect" like a dissolve.

Although there are clearly some cases where the user will need to get involved in tweaking the edits, we feel that in the majority of cases, the system will be able to handle the edits automatically. It will await field trials, however, to measure how successful our heuristics will be.

8. INTELLIGENT CRITICS – FUTURE WORK

In school, children spend enormous amounts of time learning and practicing how to write. This includes learning the rules for organizing material, constructing sentences and paragraphs, and generally making a logical and understandable composition. However, few people will learn the corresponding rules for creating high-quality video and multimedia compositions. These are generally only taught in specialized, elective courses on film or video production.

Therefore, in order to help people create higher-quality productions, we plan to provide automatic critics that help evaluate and guide the quality of the production. Some of the techniques discussed in section 7 above will actually help improve the quality. We intend to go beyond this to provide many other heuristics that will watch the user's production and provide pop-up suggestions such as "avoid shaky footage" (as in [14]) and "avoid cutting in the middle of a camera pan."

Video editing is a highly subjective part of filmmaking, which can greatly affect the look of the finished product. Therefore, though some of the intelligent editing can be automated to prevent the user from making obvious errors, in some cases it is best to simply inform the user of the rule rather than make the artistic decision for them. For this reason, providing help as an Intelligent Critic is likely to be appropriate in this application.

A century of filmmaking has generated well-grounded theories and rules for film production and editing which can be used by our critic (e.g., [9] [21] [3]). For example, the effect of camera angle on comprehension and emotional impact [4], the effect of shot length and ordering on learning [13], and the effect of lighting on subjects' understanding of scenes [30], are just a small sample of film-making heuristics. As automatically generated metadata improves, it will be possible for Silver to give users more sophisticated assistance. For example, when Informedia vision systems are able to recognize similar scenes through an understanding of their semantic content, a future version of Silver could suggest that the first use of the scene be presented for a longer period than subsequent presentations. This is desirable to keep users' interest, keeping the user from becoming bored with the same visual material. Such capabilities in Silver will still not make Hollywood directors and editors of school children. However, it will provide a level of assistance that should enable naïve users to create much more pleasing productions from video archives.

9. CONCLUSIONS

We are implementing the Silver editor in Visual Basic, and although most functions are implemented, there is much more to be done. One important goal of the Silver project is to distribute our video editor so people can use it. Unfortunately, it is not yet robust enough to be put in front of users. It is also an important part of our plans to do informal and formal user tests, to evaluate and improve our designs.

As more and more video and film work is performed digitally, and as more and more homes and classrooms use computers, there will clearly be an increased demand for digital video editing. Digital libraries contain increasing amounts of multi-media material including video. Furthermore, people will increasingly want easy ways to put their own edited video on their personal web pages so it can be shared. Unfortunately, today's video editing tools make the editing of video significantly harder than the editing of textual material or still images. The Silver project is investigating some exciting ways to address this problem, and hopefully will point the way so the next generation of video editors will be significantly easier to use.

ACKNOWLEDGEMENTS

The Silver Project is funded in part by the National Science Foundation under Grant No. IIS-9817527, as part of the Digital Library Initiative-2. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation. Thanks to Bernita Myers for help with this paper.

REFERENCES

[1] Apple Computer, iMac. 2000. http://www.apple.com/imac/.

[2] Boreczky, J., et al. "An Interactive Comic Book Presentation for Exploring Video," in CHI 2000 Conference Proceedings. 2000. ACM Press. pp. 185-192.

[3] Cantine, J., Howard, S., and Lewis, B. Shot by Shot: A Practical Guide to Filmmaking, Second Edition. 1995, Pittsburgh Filmmakers.

[4] Carroll, J.M. and Bever, T.G. "Segmentation in Cinema Perception." Science, 1976. 191: pp. 1053-1055.

[5] Chiueh, T., et al. "Zodiac: A History-Based Interactive Video Authoring System," in Proceedings of ACM Multimedia '98. 1998. Bristol, England:

[6] Christel, M., et al. "Techniques for the Creation and Exploration of Digital Video Libraries," *Multimedia Tools and Applications*, B. Furht, Editor 1996, Kluwer Academic Publishers. Boston, MA.

[7] Christel, M., Winkler, D., and Taylor, C. "Multimedia Abstractions for a Digital Video Library," in Proceedings of the 2nd ACM International Conference on Digital Libraries. 1997. Philadelphia, PA: pp. 21-29.

[8] Chua, T. and Ruan, L., "A video retrieval and sequencing system." ACM Transactions on Information Systems, 1995. 13(4): pp. 373-407.

[9] Dancyger, K. The Technique of Film and Video Editing: Theory and Practice. Second ed. 1997, Boston, MA: Focal Press.

[10] Davenport, G., Smith, T.A., and Pincever, N. "Cinematic Primitives for Multimedia." IEEE Computer Graphics & Applications, 1991. 11(4): pp. 67-74.

[11] Fischhoff, B., et al. What Could You Do? Carnegie Mellon University, 1998. http://www.cmu.edu/telab/telfair_program/Fischhoff.html. Interactive Video DVD.

[12] Gauch, S., Li, W., and Gauch, J. "The VISION Digital Video Library." Information Processing & Management, 1997. 33(4): pp. 413-426.

[13] Gavriel, S. "Internalization of Filmic Schematic Operations in Interaction with Learners' Aptitudes." Journal of Educational Psychology, 1974. 66(4): pp. 499-511.

[14] Girgensohn, A., et al. "A semi-automatic approach to home video editing," in Proceedings of UIST'2000: The 13th annual ACM symposium on User interface software and technology. 2000. San Diego, CA: ACM. pp. 81-89.

[15] Hampapur, A., Jain, R., and Weymouth, T. "Digital Video Segmentation," in Proceedings of the Second ACM International Conference on Multimedia. 1994. San Francisco: pp. 357-364.

[16] Hauptmann, A. and Smith, M. "Text, Speech, and Vision for Video Segmentation: The Informedia Project," in AAAI Symposium on Computational Models for Integrating Language and Vision. 1995.

[17] Hauptmann, A.G. and Smith, M. "Video Segmentation in the Informedia Project," in IJCAI-95: Workshop on Intelligent Multimedia Information Retrieval. Montreal, 1995. Montreal, Quebec, Canada.

[18] Lee, H., et al. "Implementation and Analysis of Several Keyframe-Based Browsing Interfaces to Digital Video," in Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2000). 2000. Lisbon, Portugal: http://lorca.compapp.dcu.ie/Video/Papers/ECDL2000.pdf.

[19] Mackay, W.E. and Davenport, G. "Virtual Video Editing in Interactive Multimedia Applications." CACM, 1989. 32(7): pp. 832-843.

[20] Mackay, W.E. and Pagani, D. "Video Mosaic: Laying out time in a physical space," in Proceedings of the second ACM international conference on multimedia. 1994. San Francisco, CA: ACM. pp. 165-172.

[21] Mascelli, J. The Five C's of Cinematography: Motion Picture Filming Techiques. 1965, Los Angeles, CA: Silman-James Press.

[22] Matthews, J., Gloor, P., and Makedon, F. "VideoScheme: A Programmable Video Editing System for Automation and Media Recognition," in ACM Multimedia'93 Proceedings. 1993. pp. 419-426.

[23] Meng, J. and Chang, S. "CVEPS: A Compressed Video Editing and Parsing System," in Proceedings of the fourth ACM international conference on multimedia. 1996. Boston, MA: pp. 43-53.

[24] Mills, M., Cohen, J., and Wong, Y. "A Magnifier Tool for Video Data," in SIGCHI '92 Conference Proceedings of Human Factors in Computing Systems. 1992. Monterey, CA: ACM. pp. 93-98.

[25] Myers, B.A. "Demonstrational Interfaces: A Step Beyond Direct Manipulation." IEEE Computer, 1992. 25(8): pp. 61-73.

[26] Myers, B.A. Natural Programming: Project Overview and Proposal. Technical Report, Carnegie Mellon University School of Computer Science, CMU-CS-98-101 and CMU-HCII-98-100, January, 1998. Pittsburgh.

[27] Placeway, P., et al. "The 1996 Hub-4 Sphinx-3 System," in DARPA Spoken Systems Technology Workshop. 1997.

[28] Sato, T. et al. "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption." ACM Multimedia Systems: Special Issue on Video Libraries, 1999. 7(5): pp. 385-395.

[29] Stevens, S.M., Christel, M.G., and Wactlar, H.D. "Informedia: Improving Access to Digital Video." interactions: New Visions of Human-Computer Interaction, 1994. 1(4): pp. 67-71.

[30] Tannenbaum, P.H. and Fosdick, J.A. "The Effect of Lighting Angle on the Judgment of Photographed Subjects." AV Communication Review, 1960. pp. 253-262.

[31] Tonomura, Y., et al. "VideoMAP and VideoSpaceIcon: Tools for Anatomizing Video Content," in Proceedings INTERCHI'93: Human Factors in Computing Systems. 1993. ACM. pp. 131-136.

[32] Ueda, H. and Miyatake, T. "Automatic Scene Separation and Tree Structure GUI for Video Editing," in Proceedings of ACM Multimedia '96. 1996. Boston:

[33] Ueda, H., Tmiyatake, T., and Yoshizawa, S. "Impact: An Interactive Nautral-Motion-Picture Dedicated Multimedia Authoring System," in Proceedings of INTERCHI'93: Conference on human factors in computing systems. 1993. Amsterdam: ACM. pp. 137-141.

[34] Wactlar, H.D., et al. "Lessons learned from building a terabyte digital video library." IEEE Computer, 1999. 32(2): pp. 66 -73.

[35] Zhang, H., et al. "Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution," in ACM Multimedia 95: Proceedings of the third ACM international conference on Multimedia. 1995. San Francisco, CA: pp. 15-24.

VideoGraph: A New Tool for Video Mining and Classification [*]

Jia-Yu Pan, Christos Faloutsos
Computer Science Department
Carnegie Mellon University
Pittsburgh, USA
{jypan, christos}@cs.cmu.edu

ABSTRACT

This paper introduces *VideoGraph*, a new tool for video mining and visualizing the structure of the plot of a video sequence. The main idea is to "stitch" together similar scenes which are apart in time. We give a fast algorithm to do stitching and we show case studies, where our approach (a) gives good features for classification (91% accuracy), and (b) results in *VideoGraphs* which reveal the logical structure of the plot of the video clips.

1. INTRODUCTION

In this paper, we focus on automatically determining and visualizing the 'story-plot' of a video clip. This helps us distinguish between different video types, e.g., news stories versus commercials. The problem is: given a video clip, determine the evolution of its story-plot. For example, do similar scenes alternate, as is the case of dialogue?

2. PROPOSED METHOD

The main idea behind our approach is to spot shots that tend to occur repeatedly. For example, in an interview, shots of the interviewer and the interviewee will occur often, and intermixed. An important concept we introduce is the *shot-group*. Recall that a *shot* is a set of consecutive similar frames.

Definition 1. A *shot-group* is a set of shots, that are similar according to our similarity function.

Notice that shots of a shot-group may or may not be consecutive in time. Our experience showed that we should mainly consider shot-groups that consist of many shots. Specifically, we make the following distinctions:

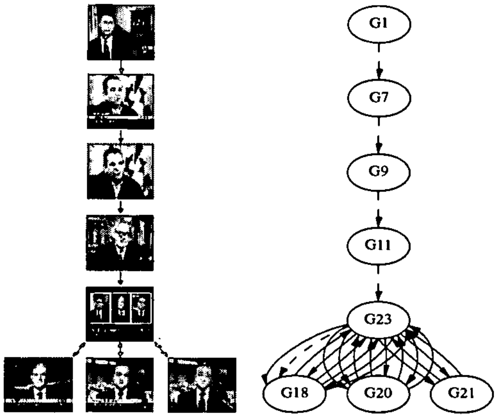


Figure 1: The VideoGraph of a news story. An alternating starlike structure is shown, which corresponds to an interview in the latter part of the story.

A shot-group is called *persistent* if it contains multiple, consecutive shots; while a shot-group that contains multiple, non-consecutive shots, is called *recurrent*. A shot-group that is both persistent and recurrent is called a *basic* shot-group; it is exactly the type of shot-group that we use to construct a VideoGraph. Intuitively, basic shot-groups are the ones that contain shots that occur often, and therefore should be helpful in revealing the plot evolution. This is achieved through the VideoGraph:

Definition 2. The VideoGraph of a video clip is a directed graph, where every node corresponds to a shot-group, and where edges indicate temporal succession.[1]

Figure 1 gives a news clip and its VideoGraph. Notice the pronounced star-like pattern, which, it turns out, corresponds to an interview. The graph contains only the basic shot-groups, namely G1, G7, G9, G11, G23, G18, G20, G21. Notice the heavy traffic between shot-groups G23, G18, G20 and G21, with G23 being the center of the 'star'-shape. In the same Figure 1 we also show the key frames from each shot-group. The key frame of a shot-group was randomly selected among the frames of the participating shots. Notice

[*]This material is based on work supported by the National Science Foundation under Cooperative Agreement No. IRI-9817496.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

[1] Dashed edges indicate that some non-basic shot-groups have been deleted along the way.

again that the 'center' shot-group, G23, indeed is the recurring shot of the interview, which contains shots when the reporter proposed questions to the invited speakers. The shot-groups which interact with G23 heavily, i.e. G18, G20, and G21, contains exactly the shots of the invited speakers replying to the reporter's questions.

Next we describe the algorithm to generate the VideoGraph for a video clip. First, use any off-the-shelf shot detection technique to break the video into shots. Next, use our 'stitching' algorithm (described next) to group similar shots into shot-groups. Each shot-group is assigned a label. Then, collect statistics (i.e. count of persistent shot-groups, count of recurrent ones, e.t.c). Finally, keep only the shot-groups that are both persistent and recurrent; consider the edges among them, indicating temporal succession; and use an off-the-shelf graph-layout tool (e.g., *dot*[2]) to draw the resulting graph. This is exactly the VideoGraph for this video clip.

There are two issues left to discuss: (a) which statistics to keep and (b) how the 'stitching' algorithm works. For the first question, we propose the following statistics: the number of I-frames, the number of shots, and the number of shot-groups in the video clip; the number of persistent, recurrent and basic shot-groups; the count of shots in each of the three classes of shot-groups; the average number of shots per shot-group; and the percentage of shots that are within persistent/recurrent shot-groups.

The second issue to discuss is our *video stitching algorithm ('VS')*, and specifically, how it decides whether two shots should be grouped together or not. Due to lack of space, we just give the main idea: For each I-frame, we used the DCT-coefficients of the macroblocks (MB), and, specifically, the first m=20 attributes, after using a dimensionality reduction method (*FastMap* [1]). Thus, each shot is a 'cloud' of m-dimensional points. We group together two such 'clouds', if the density of the union is comparable to the original, individual densities.

3. EXPERIMENTS - RESULTS

We used MPEG-1 video clips segmented from CNN news reports. Each clip is either a news story or a continuous series of commercials. We chose the technique from [2] to do scene shot detection.

Story-plot visualization Figure 2 shows the VideoGraphs for several news stories (upper part)and commercials (lower part).It is clear that most of the commercial clips have fewer persistent and recurrent shot-groups and much fewer basics, than news stories do. Consequently, VideoGraphs of commercial clips are more likely to be an empty graph or a graph of one or two nodes (Figure 2). This is due to the fact that commercials rarely have a 'plot' being, instead, a collection of shots are not revisited.

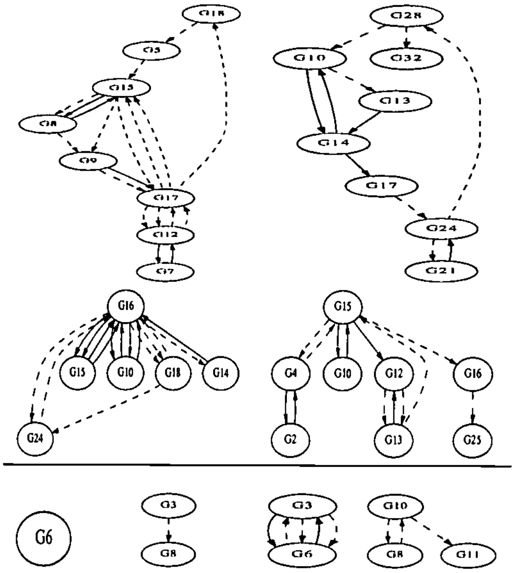


Figure 2: VideoGraphs of news stories (upper part) and commercials (lower part). Note the more complicated structure of news stories.

Classification Here we illustrate that the features we extracted capture the logical structure of a video clip quite well and therefore are promising for video classification. We conducted a classification experiment based on 68 news stories and 33 commercials, whose total size is about 2 gigabytes. Each clip is represented by 20 features (count of recurrent shot-groups, count of persistent shot-groups, etc.) We conducted 5-fold cross-validation using an ID3 classifier and we achieved classification accuracy of 91% (plus or minus 0.16 percentage points, for the 95% confidence interval). We used an off-the-shelf package, $\mathcal{MLC}$++ [3], for the classification.

4. CONCLUSIONS

We presented the *VideoGraph*, a new tool for video mining and story plot visualization. The heart of our approach is to group similar shots together, even if they are not consecutive. We give algorithms that "stitch" similar shots into shot-groups, and automatically derive VideoGraphs. We also show how to derive features (e.g. number of recurrent shot-groups, etc.) for video mining and classification. In a case study with news and commercials, our proposed features achieved 91% classification accuracy, without even using the audio information. VideoGraphs can also be used as video representatives for efficient browsing in digital video libraries such as Informedia [4], and are useful on keyframe selection.

5. REFERENCES

[1] C. Faloutsos and K.-I. Lin. Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *Proceedings of the ACM SIGMOD Conference*, pages 163–174, 1995.

[2] V. Kobla, D. Doermann, and C. Faloutsos. Videotrails: Representing and visualizing structure in video sequences. *Proceeding of the Fifth ACM International Multimedia Conference*, pages 335–346, November 1997.

[3] R. Kohavi, D. Sommerfield, and J. Dougherty. Data mining using MLC++: A machine learning library in C++. *Tools with Artificial Intelligence*, `http://www.sgi.com/Technology/mlc`.

[4] H. Wactlar, M. Christel, Y. Gong, and A. Hauptmann. Lessons learned from the creation and deployment of a terabyte digital video library. *IEEE Computer*, 32(2):66–73, February 1999.

[2] Available at http://www.research.att.com/sw/tools/graphviz/

The Alexandria Digital Earth Prototype System

Terence R. Smith[1], Greg Janée[2], James Frew[3], Anita Coleman[4]

Alexandria Project, University of California at Santa Barbara,
Santa Barbara, CA 93106

1. INTRODUCTION

This note summarizes the system development activities of the Alexandria Digital Earth Prototype (ADEPT) Project.[5] ADEPT and the Alexandria Digital Library (ADL) are, respectively, the research and operational components of the Alexandria Digital Library Project. The goal of ADEPT is to build a distributed digital library (DL) of personalized collections of geospatially referenced information. This DL is characterized by: (1) services for building, searching, and using personalized collections; (2) collections of georeferenced multimedia information, including dynamic simulation models of spatially distributed processes; and (3) user interfaces employing the concept of a "Digital Earth." Important near-term objectives for ADEPT are to build prototype collections that support undergraduate learning in physical, human, and cultural geography and related disciplines, and then to evaluate whether using such resources helps students learn to reason scientifically. Collections and services developed by ADEPT researchers will migrate to ADL as they mature.

2. ARCHITECTURE

The ADEPT architecture (Figure 1) extends the ADL architecture while diverging from the latter's traditional client-server roots. ADEPT is a set of distributed peers, each supporting one or more collections of objects, subsets of which may be "published" (made visible) to other peers. The "library" is the sum of all such collections.

Collections are the primary organizational metaphor in ADEPT. A collection is a set of independent, typed objects (more precisely, a set of references to objects) together with certain contextual and structural collection-level metadata. Objects are largely undefined, save for their ability to provide object-level metadata and to be uniquely identified. While the contents of the library may thus be utterly heterogeneous, three features integrate the library into a uniform whole: (1) common collection-level metadata; (2) the ADEPT "bucket" system (a common model for object-level metadata); and (3) a central collection discovery service.

ADEPT buckets differ notably from other metadata schemes such as detailed content standards (e.g., FGDC [2]) and high-level definitions (e.g., Dublin Core [3]). The ADEPT bucket system is a transparent metadata aggregation system in which semantically similar, quasi-strongly-typed metadata fields may be formally grouped to provide higher-level search and description capabilities. Objects map their relevant native metadata to buckets; significantly, both field names and values are recorded. Collections index, aggregate, and summarize this information. They allow clients to search by entire buckets and to retrieve bucket-level descriptions of objects, and also to "drill down" into buckets (i.e., to search by specific metadata fields that have been mapped to the buckets). Buckets are strongly typed, being either spatial, temporal, hierarchical, textual, qualified textual, or numeric. Associated with each bucket type are a standard representation (e.g., polygons and boxes for spatial buckets) and a set of standard operators (e.g., set algebra). This combination of a formal aggregation system with strong typing yields a metadata model that is flexible enough to accommodate a wide range of metadata, yet expressive enough to support powerful, targeted queries.

ADEPT collection metadata includes contextual information (scope, purpose, etc.) and automatically derived structural information such as the spatiotemporal distribution of objects, the types and counts of objects, and bucket metadata mappings. Metadata mappings are qualified with statistics to give clients an indication of how well represented metadata fields are within buckets.

New ADEPT services layered atop the existing ADL infrastructure support: (1) creating and managing local collections; (2) creating new objects from client-supplied metadata; (3) importing objects from remote collections into local collections; and (4) categorizing objects according to client-supplied thesauri. In creating new objects clients are free to provide any metadata, the only requirements being that the metadata be represented in XML and that the client provide an XSLT script mapping the object metadata to a standard bucket-level description. Remote objects are currently imported by simply copying them, but a future version of ADEPT will also support proxy objects and explicit linking. The ability to create new collections and categorize them with user-supplied thesauri effectively supports building personalized collections.

To facilitate including legacy content in the library, especially content managed with relational database technology, we have developed a generic collection driver for relational databases. The driver assumes only that it is connected to the collection via JDBC. The collection's schema, metadata bucket mappings, thesauri, etc. are all described in configuration files. Metadata reports are described by templates, which are interpreted by custom, query-driven report generation software. Query translation is controlled by a Python-based scripting system that sup-

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

[1] smithtr@cs.ucsb.edu [2] gjanee@alexandria.ucsb.edu
[3] frew@bren.ucsb.edu [4] asc@acm.org
[5] The research reported in this paper was partially supported under NSF IIS-9817432.

ports user-defined functions and customized query formation and optimization.

The ADEPT collection discovery service manages a registry of known collections and, more significantly, supports discovering collections probabilistically according to their relevance to a given query. Using the notion that relevance is determined by the numbers of objects satisfying a query, we are: (1) investigating new indexing techniques that support joint multidimensional indexing of collection contents using sparse, compact data structures; and (2) using the simple histogram and count information contained in collection-level metadata to derive relevance from typed spatiotemporal information density.

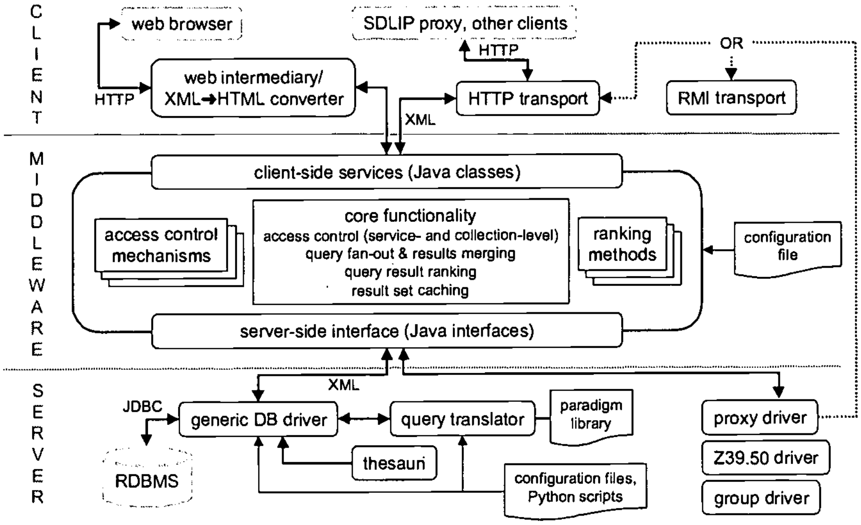


Figure 1. ADEPT architecture and implementation.

Future directions for the ADEPT architecture include supporting a richer information model, facilitating interoperability, and supporting dynamic content such as models of Earth processes and other executable programs. ADEPT's current information model (collections of independent objects) does not capture key relationships of benefit to users. For example, objects that are members of a series share a relationship with each other and to the series as a whole; we plan to make such relationships explicit. With regard to interoperability, we plan to facilitate searching over other types of collections by supporting connections to Z39.50 servers, and by extending the SDLIP protocol. With regard to dynamic content, our focus is on defining an architectural layer that delivers executable content to clients in a platform-neutral fashion, and that allows programs to interact with library contents and to populate the library with new objects. We are designing a suite of additional services that support building personalized collections by, for example, allowing users to: (1) build collections according to personal preferences; (2) build models representing concepts in different ways; and (3) build concept maps.

3. CATALOG AND COLLECTIONS

Three prototypical ADEPT collections are being built to aid in designing future collections and services, and to serve as a basis for evaluating the educational impacts of ADEPT services. The collections focus on processes and forms in (1) physical geography (e.g., landscapes and tectonic phenomena); (2) human geography (e.g., diffusion phenomena); and (3) cultural geography (e.g., religious sites). Initial collections for physical and human geography have already been tested in electronic classroom presentations. The collection building process involves (1) acquisition; (2) cataloging; (3) ingesting; (4) evaluating with respect to user scenarios; and (5) building ISO-standard topic maps. The collections may be searched by concept and reorganized with different levels of granularity into personal collections for various pedagogic purposes (e.g., lectures, self-paced labs, and individual and collaborative learning sessions.) Catalog descriptions of objects employ the ADEPT Learning Objects Metadata Model (ADEPT LOMM) standard, which is based on the IMS [4], DLESE [5], and IEEE LOM [6] metadata standards, and an ADEPT-developed standard for model metadata [7]. The model's key educational and pedagogical elements include (1) type of learning resource; (2) learning context; (3) interactivity level; and (4) description.

4. REFERENCES

[1] *The Alexandria Digital Earth Modeling System (ADEPT).* http://www.alexandria.ucsb.edu/adept/proposal.pdf.

[2] Federal Geographic Data Committee. *Content Standard for Digital Geospatial Metadata.* http://www.fgdc.gov/metadata/csdgm/.

[3] Dublin Core Metadata Initiative. *Dublin Core Metadata Element Set.* http://dublincore.org/documents/dces/.

[4] IMS Global Learning Consortium. *IMS Learning Resource Meta-Data Information Model.* http://www.imsproject.org/metadata/mdinfov1p1.html.

[5] DLESE. http://www.dlese.org/.

[6] IEEE Learning Technology Standards Committee. *LOM: Base Scheme - v3.5.* http://ltsc.ieee.org/doc/wg12/scheme.html.

[7] Crosier, S. *Content Standard for Computer Model Metadata.* http://www.geog.ucsb.edu/~scott/metadata/standard/.

Iscapes: Digital Libraries Environments for the Promotion of Scientific Thinking by Undergraduates in Geography

Anne J. Gilliland-Swetland
Department of Information Studies, University of California, Los Angeles
212 GSE&IS Building
Los Angeles, CA 90095-1520
(+ 001) 310-206-4687
swetland@ucla.edu

Gregory L. Leazer
Department of Information Studies, University of California, Los Angeles
226 GSE&IS Building
Los Angeles, CA 90095-1520
(+001) 310-206-8135
gleazer@ucla.edu

ABSTRACT

This paper reviews considerations associated with implementing the Alexandria Digital Earth Prototype (ADEPT) in undergraduate geography education by means of Iscapes (or Information landscapes). In particular, we are interested in how Iscapes might be used to promote scientific thinking by undergraduate students. Based upon an ongoing educational needs assessment, we present a set of conceptual principles that might selectively be implemented in the design of educational digital library environments.

Keywords

Digital libraries. Geography. Undergraduate education. Scientific thinking.

1. INTRODUCTION

The Alexandria Digital Earth ProtoType (ADEPT) is a 5-year project (1999-2004) funded by the U.S. Digital Libraries Initiative, Phase 2 that is developing a "geolibrary" of georeferenced and geospatial resources. A key aim of ADEPT is to integrate it into undergraduate education in disciplines that might make use of such resources. Iscapes (or Information landscapes) are a construct that specifically addresses the requirements for integrating ADEPT into instructional and learning activities. The Iscape construct, at its simplest, denotes a personalized use environment, both in terms of information resources and functionality. Information resources used in an Iscape might be drawn from many distributed sites. They might include material contributed by the user, derivative materials created by manipulation of ADEPT collections, and ADEPT's own extensive collections. A key aim for Iscapes is to provide instructors and students with the means to discover, model, manipulate, and thus explain dynamic geographical processes and learn the underlying scientific reasoning.

Scientific thinking in geography requires students to master at least five skills: asking geographic questions, acquiring geographic information, organizing geographic information, analyzing geographic information, and answering geographic questions [2]. Digital libraries deployed in education should also allow for "inquiry into authentic questions generated from student experiences" [3]. If a digital library is going to be integrated directly into instructional and learning activities and be instrumental in the inculcation of students' scientific thinking, then it is likely that it will need to make available significantly augmented functionality beyond that of a digital library that is intended primarily as a supporting information resource.

A team of ADEPT researchers at the University of California, Los Angeles and Santa Barbara has been examining the educational implementation and evaluation of ADEPT in undergraduate geography courses on both campuses [1]. As part of an educational user needs assessment, we have been engaged in analyzing instructors' pedagogical goals and identifying the concepts that instructors use in teaching core geographic topics to understand how these might be facilitated by ADEPT. We have been particularly interested in how the use of digital libraries can promote not only the learning of core geographic concepts, but also the thinking processes associated with the relevant disciplinary domains.

2. DESIGNING ISCAPES

2.1 Considerations

In addressing the design of Iscapes to support the acquisition of scientific thinking skills and knowledge by undergraduate students, it is necessary first to identify what needs to be taught and how. In university environments, there are few "objective" frameworks that we can use to understand the scope of learning expectations within a discipline. In the K-12 environment, standardized frameworks provide some guideposts, but in universities, course content and learning objectives can be highly specific to individual instructors, even across instructors teaching different sections of the same course. Commonly used textbooks provide some assistance in identifying core content and perspectives, but there is also wide variability in the texts and editions. As a result, we face several issues, not only with the development of relevant collections, but also in determining how

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

instructors and students can or should be able to retrieve, collate, and manipulate those collections. Some examples of issues that arise include the following: How do we analyze instructors' learning goals for their students? To what extent do instructors' own research interests and the way in which they approach their discipline influence the way they teach core concepts? In what other ways can we identify core collections, core topic and concepts, and best practices for that discipline? How do instructors underscore or reinforce important points? What are the measures and permitted transformations of concepts and information resources that need to be supported? How can a student faced with a variety of ways to solve a problem such as finding an answer to an assigned question know that he or she has taken the "best" approach, i.e., demonstrated the appropriate scientific reasoning techniques? Finally, and key to Iscape development, is it possible to build an educational digital library environment that supports widely varying instructional approaches while supporting and modeling best practices and core concepts?

2.2 Conceptual Requirements

As was mentioned above, one way we have addressed these questions is by determining the core geographical concepts. Based upon analysis of interviews with instructors about their instructional goals and pedagogical approaches, observation of instructors teaching course classes, as well as class syllabi, assignments, and required texts, we have been developing concept maps for major topics within courses.

From our work to date, we have formulated a set of conceptual principles that might selectively be implemented in the design of educational digital library environments such as Iscapes:

Transparency: The technology to support Iscape development and use should not intrude upon developing and learning content, concepts, and processes.

Progressive skill-building: Users who master basic skills and concepts should be able to build upon these as they progress on to more sophisticated Iscapes.

Extensibility: Iscapes should be augmentable through the inclusion of additional content and layers of functionality permitting more sophisticated views of content.

Parameter variation: Iscapes should support interactive visual and data models of dynamic processes and hypothetical scenarios. These models and scenarios will permit users to vary a range of relevant parameters and will display results to users.

Granularity: The level of detail at which Iscape content should be displayed and manipulated can be varied to facilitate different levels of expertise as well as different disciplinary needs.

Computational capabilities: Iscapes should incorporate data manipulation tools to support longitudinal and stochastic analyses.

Linkage: Citation-linking should assist in the identification of additional resources on a topic, and of seminal and influential works. It also aids users in tracing the development of key findings and ideas.

Hybrid collections: Iscapes should support references to related non-digital resources such as maps contained in university library collections.

Multi-disciplinary: Iscapes should be constructed that support a range of domain knowledge, work practices, and discipline-specific reasoning using geo-spatial resources.

Diversity and extensibility of metadata: Creators of Iscapes should be able to integrate new resources and associated metadata, and to create metadata for extant resources specific to their instructional or learning activities.

Authentic science topics: Iscapes should present authentic scientific problems and data.

Expert knowledge and reasoning: Iscapes should reflect best practices and most authoritative current knowledge in a given domain.

Support of disciplinary conventions: Iscapes should adhere to disciplinary practices and conventions for concept representation such as labeling, use of color, and visual perspectives.

Appropriate use of technology: Iscapes should be developed where they will be more effective than existing classroom methods, such as the modeling of dynamic processes, presenting alternate visualizations of data or images, or prediction.

Scaffolding: Iscapes should make available tools to instructors to allow them to collate and annotate specific resources, as well as predetermine how students may move among and manipulate those resources.

Collaboration: Iscapes should facilitate group work by students through the use of collaboration tools.

Annotation: Instructors and students should be able to annotate resources and otherwise embed commentary on their interactions with the Iscapes for others to see. Instructors and students should also be able to embed questions for each other.

Vocabulary: Iscapes should include a spell-checker and a pull-down glossary of technical terms. Instructors should be able to develop customized glossaries for their Iscapes.

Further information about the ADEPT project is available at ADEPT web sites at UCLA (http://is.gseis.ucla.edu/adept/) and UCSB (http://www.alexandria.ucsb.edu/dept/).

3. ACKNOWLEDGMENTS

Our thanks to Christine Borgman for her comments on drafts of this paper. Additional thanks to Richard Mayer. This article is based upon work supported by the National Science Foundation under grant no. IIS-9817432.

4. REFERENCES

[1] Borgman, C.; Gilliland-Swetland, A.; Leazer, G.; Mayer, R.; Gwynn, D.; Gazan, R.; Mautone. P. Evaluating the use of digital libraries in undergraduate education: a case study of the Alexandria Digital Earth Prototype, Library Trends, 49 (2): 228-250.

[2] Geography Education Standards Project. Geography for life: national geography standards. Washington, DC: National Geographic Society (1994).

[3] National Research Council (1996). National science education standards. Washington, D.C.: National Academy Press.

Project ANGEL: An Open Virtual Learning Environment with Sophisticated Access Management

John MacColl
Director of Science & Engineering Library,
Learning & Information Centre (SELLIC) Project/Sub-Librarian, Online Services
University of Edinburgh
Darwin Library
Edinburgh, Scotland, UK
+44 131 650 7275

john.maccoll@ed.ac.uk

ABSTRACT

This paper describes a new project funded in the UK by the Joint Information Systems Committee, to develop a virtual learning environment which combines a new awareness of internet sources such as bibliographic databases and full-text electronic journals with a sophisticated access management component which permits single sign-on authentication.

Categories and Subject Descriptors

[**Authentication; Personalization**]:.

General Terms

Design, Standardization

Keywords

Authentication; Access Management; Virtual Learning Environments.

1. INTRODUCTION

Under its programme for the Development of the DNER [1] (Distributed National Electronic Resource), the UK's Joint Information Systems Committee [2] (JISC) has funded a consortium project known as ANGEL [3] (Authenticated Networked Guided Environment for Learning), which is led by the London School of Economics, with partners at the University of Edinburgh, De Montfort University, South Bank University and associate partners EDINA (Edinburgh Data and Information Access) and Sheffield Hallam University. The ANGEL consortium was formed in the autumn of 2000. System development began early in 2001, with a two-year timescale.

2. A GUIDED ENVIRONMENT FOR LEARNING

ANGEL is developing an environment for learning which is 'guided' in several ways. Student users will be guided in the sense of entering a tailored, customised environment – an individual portal which collocates administrative and academic information which relates to them – from registry, library, finance, school and department. What ANGEL will offer in addition to the features now common to personalised environments is access to relevant databases of content – primary and secondary – as well as to 'pre-coordinated' resources selected by the tutor. The system will not insist on any particular interface (thus it will work with existing customised or institutionally-preferred portals), though it will provide one for institutions which wish.

At the heart of the ANGEL system is the tutor interface, which is the way by which the 'guiding' of the student user is achieved. ANGEL is building tools to make it simple for tutors, as they construct online courses, to select resources which should be added to the environments of all students on their course. They will be prompted to select from a range of resources both local and external, including DNER resources such as the databases and datasets available from the UK's JISC-funded data centres (such as EDINA) as well as useful locally-produced learning resources at various levels. These may range from aggregated resources such as online courses, obviously (all users should normally see several of these), to 'atomic' resources, or course 'granules', which may be located in several different course sites, or within the library system.

3. ACCESS MANAGEMENT

ANGEL aims to extend the compass of a learning environment beyond the normal 'closed' environment of particular resources selected by a tutor, to the more open environment of bibliographic databases, full-text electronic publications and datasets available within the DNER and beyond. Essential to this approach is the development of an authentication and authorisation module which permits, through a 'single sign-on' to ANGEL, access to the range of secured resources to which a user has rights. This requires an approach based upon 'authentication brokering', already developed by the ANGEL lead partner site, the London School of Economics, in an earlier JISC-funded project known as HEADLINE. [4] In this aspect of its development, ANGEL is tracking the work of JISC's Committee on Authentication and Security (JCAS) which is steering the development of a UK-wide access management system known as Sparta. Sparta will replace the Athens system which has been used successfully in the UK for several years, but which is now perceived to be limited in the degree of security it can provide, and in the flexibility which can be offered to institutional management to limit access to specific groups, and draw detailed usage information from the system. Sparta is being developed alongside work going on in the US under the Internet2 'Middleware Initiative' (in particular the work on the 'Shibboleth' system), and in continental Europe, through TERENA (Trans-European Research and Education Network Association) [5]. ANGEL will deploy Sparta technology in prototype.

4. AUTHENTICATION AND AUTHORISATION PRINCIPLES

Development of the authentication and authorisation module is based upon the following principles:

- Responsibility for authentication rests with the institution
- Authentication is based upon centrally-held authoritative institutional user information.
- Authentication and authorisation are functionally separate procedures
- Authorisation involves the identification of roles of authenticated users.
- Authorisation identifies people, not computers.
- Granularity is provided to the level of individual users without any compromise to security
- Migration to the ANGEL system is seamlessly provided to Athens users, and parallel working is available for a reasonable period of time
- Credential subsidiarity operates, whereby credentials are provided only to the level which is appropriate. Thus, in the case of resources which are licensed for use by all members of an institution, with deeper level credentials available also to identified groups or individuals, the credential used for access is determined by the user's role.

This means that institutions are free to use a variety of approaches. The Project will prioritise the deployment of X.509 digital certificates for authentication in the context of an institution acting as a Certificate Authority. Authentication will draw upon existing student databases held in each partner university, and will not require new databases to be established. The Project will assess the difficulties encountered by this requirement.

The directory structure used includes the use of the Lightweight Directory Access Protocol (LDAP) for the storage of user credentials and roles information for authorisation. The roles definitions used by the institution, and that applied by the service provider, will deploy separate namespaces. Work towards recommendations for the definition of a single roles namespace for UK HE will be carried out by the Project, allowing sufficient flexibility to allow individual institutions wide scope for variations. This work will be developed to include the concept of individual user profiling, which is important for the GEL component of the Project.

Throughout the development of the access management module, the Project will monitor Sparta developments alongside other work going on internationally and will adopt solutions where appropriate. The Extensible Markup Language (XML), now a standard for web-based description of a range of data types, including directory information, will be adopted in ANGEL.

Users will thus be guided to an environment in which all networked resources to which they have rights are available via a single sign-on, which means that they will not face multiple ID and password challenges as they take their search from resource to resource, or employ cross-searching of multiple resources. The services to which they do not have rights (perhaps because they are students on a franchised course, for example, or lifelong learners, and resource providers do not recognise them within site licence provisions) will not be visible to them.

5. PROJECT OUTCOMES

The main proposed deliverable is a system that brings together hybrid library technology with learning and teaching resources in a way that:

- Is adaptive to user requirements.
- Includes all of the best features from state-of-the-art systems for resource discovery *and* Instructional Management.
- Employs an authentication and authorisation system which is based upon the principles of single sign-on, granular allocation of resources, a generic account for DNER resources, and integration of subscription-based and local secured services.
- Enables the resources offered to students to be directly selected and maintained by the academic teaching staff responsible for providing, directing and monitoring courses of study.
- Provides monitoring and guidance for the (student) user in a way that shows some 'understanding' of what the user is trying to do when they move from the closed, structured teaching environment to the open environment of resource discovery. The system will learn about the task and the user and be able to remember the user's actions.
- Could be used as an early identifier of potential student study difficulty and hence an early warning of some forms of dropout risk.

The ANGEL system will provide an intermediate layer of 'intelligent' middleware (between users and resources, controlled by metadata about users, and itself learning from the current and past behaviour of each user) so that access to the 'open' resources of the hybrid library can be guided and monitored in similar ways to the guidance and monitoring already available within the best of the 'closed' virtual learning environments.

6. CONCLUSION

We are hopeful that ANGEL will not only provide a useful environment for students, but will also help to achieve coordinated planning between the administrative and academic domains in universities, with the library and information domain important to both. Effective intra-institutional collaboration will be essential to its success.

7. REFERENCES

[1] http://www.dner.ac.uk

[2] http://www.jisc.ac.uk

[3] http://www.angel.ac.uk

[4] http://www.headline.ac.uk

[5] http://www.terena.n

NBDL: A CIS Framework for NSDL

Joe Futrelle
NCSA, University of Illinois
Urbana-Champaign, Illinois
Futrelle@ncsa.uiuc.edu

Su-Shing Chen
University of Missouri
Columbia, Missouri
ChenS@missouri.edu

Kevin C. Chang
University of Illinois
Urbana-Champaign, Illinois
Kcchang@cs.uiuc.edu

ABSTRACT

In this paper, we describe the NBDL (National Biology Digital Library) project, one of the six CIS (Core Integration System) projects of the NSF NSDL (National SMETE Digital Library) Program.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *standards, user issues, dissemination.*

General Terms

Algorithms, Design, Standardization.

Keywords

Digital library, SMET education, Federated search.

1. INTRODUCTION

This NSDL project, NBDL, consists of participating institutions: University of Missouri-Columbia (MU), NCSA, University of Illinois-Urbana/Champaign (UIUC), and Missouri Botanical Garden (MOBOT) [6]. The project focuses on building an interoperable, reliable, and scalable Core Integration System (CIS) framework for coordinating, integrating, and supporting learning environments and resources provided by NSDL collections and services. We emphasize also integration issues of collections and services by building a testbed with a large biological collection, Tropicos, of MOBOT (Missouri Botanical Garden) and its educational services. This testbed can be extended to other disciplines and environments.

2. BIOLOGICAL INFORMATION

Biology has one of the most complex information-structures (concepts, data types and algorithms) among scientific disciplines. Its richness in organisms, species, cells, genes and their pathways provides many challenging issues for biological sciences, computational sciences, and information technology. The advances in biological science and technology urgently need development of very large biological digital libraries for analyzing and managing biological information: sequences, structures, and functions, arising from DNAs, RNAs, genes and proteins, and taxonomies. At present, biological databases are among the best archived, managed, and preserved. The earliest phase of bioinformatics is perhaps the naming and classification of organisms invented by the Swedish biologist, Carolus Linnaeus (1707-1778). His "binomial system" provides a taxonomic hierarchy of types, species, families, orders, classes, and phyla (divisions) for biology, still in use today. Unfortunately, the biology community has never developed an integrated data resource of genomic, morphological, and taxonomic information. That is, users can not search and explore all such information objects serendipitously. The NBDL project will address this important scientific issue.

3. THE EMERGE ARCHITECTURE

The technical infrastructure supporting NBDL is the Emerge distributed IR toolkit [3]. Emerge consists of a set of components to enable federation of distributed, heterogeneous data collection by means of query and data translation. Data sources are proxied with a protocol-translation component called Gazelle and integrated with a broker component called Gazebo which translates client queries into the variety of query formats required by the distributed data sources. Thus, data can be retrieved and integrated independently of its location, access protocol, and query syntax. The following depicts the Emerge architecture:

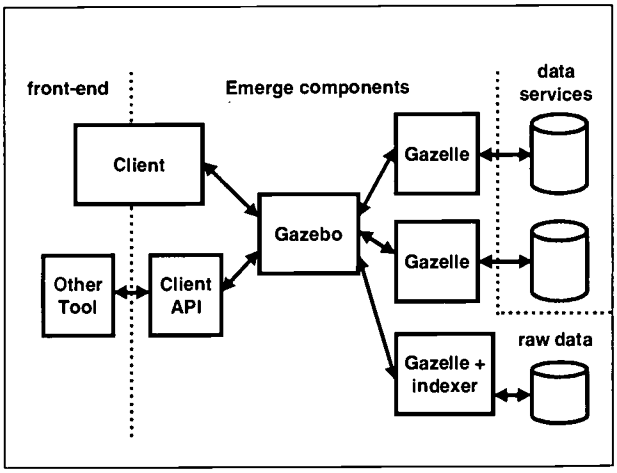


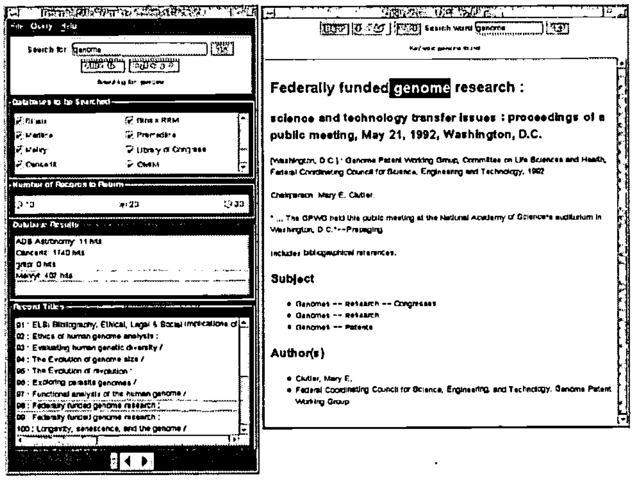
Figure 1. Emerge Architecture

Emerge components access data through a simple interface called the Gazelle Target API, which must be implemented for each data source. In some cases, classes of data sources can be supported by a single implementation. In the case of NBDL, we have developed an implementation of the Gazelle Target API for TROPICOS database, which allows arbitrary queries to be executed against it [7].

Client queries are translated by Gazebo by a general-purpose query translation engine capable of generating queries in a wide variety of

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

syntaxes. We have developed a configuration for the Gazebo query translator, which generates TROPICOS queries. Semantic equivalences between query syntaxes are expressed in Gazebo using meta-attributes, which are similar to the Alexandria Digital Library's search buckets [4]. For TROPICOS, we have developed a set of meta-attributes which can be translated into TROPICOS as well as ZBIG's Darwin Core query syntaxes, allowing a single client query to be targeted at either TROPICOS or any ZBIG service. This will allow tools to be built that can transparently retrieve biological information regardless of whether it originates in TROPICOS or a ZBIG resource. The NBDLuser interface is given the following figure:



Search for: genome

Databases to be Searched:
☑ Bimax ☑ Biomax RRM
☑ Medline ☑ Premedline
☑ Melvyl ☑ Library of Congress
☑ Cancerlit ☑ OMIM

Number of Records to Return: ○ 10 ● 20 ○ 30

Database Results:
ADS Astronomy: 11 hits
Cancerlit: 1740 hits
gmar: 0 hits
Melvyl: 407 hits

Record Titles:
91 : ELSI Bibliography, Ethical, Legal & Social Implications of
92 : Ethics of human genome analysis :
93 : Evaluating human genetic diversity /
94 : The Evolution of genome size /
95 : The Evolution of revolution :
96 : Exploring parasite genomes /
97 : Functional analysis of the human genome /
98 : Federally funded genome research :
99 : Federally funded genome research :
100 : Longevity, senescence, and the genome /

Search word: genome

Federally funded genome research :

science and technology transfer issues : proceedings of a public meeting, May 21, 1992, Washington, D.C.

[Washington, D.C.] : Genome Patent Working Group, Committee on Life Sciences and Health, Federal Coordinating Council for Science, Engineering and Technology, 1992

Chairperson: Mary E. Clutter.

"... The GPWG held this public meeting at the National Academy of Science's auditorium in Washington, D.C."--Prepaging

Includes bibliographical references.

Subject

- Genomes -- Research -- Congresses
- Genomes -- Research
- Genomes -- Patents

Author(s)

- Clutter, Mary E.
- Federal Coordinating Council for Science, Engineering, and Technology. Genome Patent Working Group

Figure 2. User Interface of NBDL

4. SEMANTIC MAPPINGS

A key requirement of the CIS architecture is the development of semantics (or ontology) of information, which are then captured in semantic mappings for the intelligent search engines, such as Emerge. Semantic mapping extends the existing full-text, hypertext, and database indexing and mapping schemes to include semantics or meanings of information content. Furthermore, semantic mappings will reduce the complexity of biological taxonomy and classification, and correlate semantically the genotypes and phenotypes at various levels of biological information in NBDL.

An example of semantic mappings is common/scientific name mapping. A significant barrier for the educational use of resources, such as TROPICOS, is the lack of common-name indexing: species are referred to by scientific names, which are unfamiliar to most educational users. To address this problem, we are extending Gazebo's query translation engine using an approximate query translation algorithm [1]. Common names will be translated into scientific names by querying a name directory service such as ITIS during the query translation process [5]. Although the resulting query will not have the exact semantics of the original query, it will be a close approximation and will allow users to discover relevant information in one integrated step rather than requiring them to use two separate applications. This is an important innovation towards disseminating biological information into the educational community.

Under various current standards (e.g., IEEE LOM and IMS Standards), metadata structures are defined and collected for some specific educational domains. Since metadata structures will always be changing, we are developing "evolving metadata structures," which have adaptive semantic properties [2]. Metadata will evolve and grow through out time and utilization. Semantic mappings are extended to schema mappings of these evolving metadata structures about learning objects and applets in our resources and services, correlating different nomenclatures, syntaxes and semantics. In the LOVE (Learning Object Virtual Exchange) of NBDL [6], we are developing this novel feature so that interoperability, reliability, reusability, and scalability will be maintained even under changes of networked resources and services.

5. RICH NBDL COLLECTIONS

The TROPICOS botanical database at the Missouri Botanical Garden (MOBOT) contains 851,000 name records for plants and associated information on bibliography, types, nomenclature, usage, distribution, and morphology. The data base currently contains over 1.5 million specimen records - mostly new collections gathered over the last 20 years with full locality data, coordinates, and elevation information. A literature file of over 80,000 publications used for vouchering distribution and usage and authority files of people, books and journals, and geographical place names. Most recently images from a variety of sources are included to add a visual impact to the wealth of textural data. At this time thousands of images of plant habitats, structure, type specimens and prologues are available for selected taxa. The web site provides a full range of on-demand and interactive html pages providing a scientific overview of the information accumulated around any of the scientific names in the production database TROPICOS.

6. ACKNOWLEDGMENTS

Our thanks to other NBDL project members: Chip Bruce, Ann Bishop, and Bryan Heidorn. Their contributions to biology and education materials are essential to this project.

7. REFERENCES

[1] C. K. Chang and H. Garcia-Molina, Approximate query translation across heterogeneous information sources, Proc. Of the 26th VLDB Conference, Sept. 2000.

[2] S. Chen, Digital Libraries: The Life Cycle of Information, Better Earth Publisher, 1998, http://www.amazon.com.

[3] "Emerge home page," http://emerge.ncsa.uiuc.edu.

[4] J. Frew, M. Freeston, L. Hill, G. Janée, M. Larsgaard, Q. Zheng. Generic Query Metadata for Geospatial Digital Libraries. IEEE Metadata 99.

[5] "Integrated Taxonomic Information System," http://www.itis.usda.gov/plantproj/itis/.

[6] NBDL. http://cecssrv1.cecs.missouri.edu/NSDLProject/.

[7]W^{3}TROPICOS. http://mobot.mobot.org/Pick/Search/pick.html.

Automatic Identification and Organization of Index Terms for Interactive Browsing

Nina Wacholder
Columbia University
New York, NY
nina@cs.columbia.edu

David K. Evans
Columbia University
New York, NY
devans@cs.columbia.edu

Judith L. Klavans
Columbia University
New York, NY
klavans@cs.columbia.edu

ABSTRACT

The potential of automatically generated indexes for information access has been recognized for several decades (e.g., Bush 1945 [2], Edmundson and Wyllys 1961 [4]), but the quantity of text and the ambiguity of natural language processing have made progress at this task more difficult than was originally foreseen. Recently, a body of work on development of interactive systems to support phrase browsing has begun to emerge (e.g., Anick and Vaithyanathan 1997 [1], Gutwin et al. [10], Nevill-Manning et al. 1997 [17], Godby and Reighart 1998 [9]). In this paper, we consider two issues related to the use of automatically identified phrases as index terms in a dynamic text browser (DTB), a user-centered system for navigating and browsing index terms: 1) What criteria are useful for assessing the usefulness of automatically identified index terms? and 2) Is the quality of the terms identified by automatic indexing such that they provide useful access to document content?

The terms that we focus on have been identified by LinkIT, a software tool for identifying significant topics in text [7]. Over 90% of the terms identified by LinkIT are coherent and therefore merit inclusion in the dynamic text browser. Terms identified by LinkIT are input to Intell-Index, a prototype DTB that supports interactive navigation of index terms. The distinction between phrasal heads (the most important words in a coherent term) and modifiers serves as the basis for a hierarchical organization of terms. This linguistically motivated structure helps users to efficiently browsing and disambiguate terms. We conclude that the approach to information access discussed in this paper is very promising, and also that there is much room for further research. In the meantime, this research is a contribution to the establishment of a solid foundation for assessing the usability of terms in phrase browsing applications.

Keywords

Indexing, phrases, natural language processing, browsing, genre.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

1. OVERVIEW

Indexes are useful for information seekers because they:

- support browsing, a basic mode of human information seeking [17].
- provide information seekers with a valid list of terms, instead of requiring users to invent the terms on their own. Identifying index terms has been shown to be one of the hardest parts of the search process, e.g., [8].
- are organized in ways that bring related information together [16].

But indexes are not generally available for digital libraries. The manual creation of an index is a time consuming task that requires a considerable investment of human intelligence [16]. Individuals and institutions simply do not have the resources to create expert indexes for digital resources.

However, automatically generated indexes have been legitimately criticized by information professionals such as Mulvany 1994 [16]. Indexes created by computer systems are different than those compiled by human beings. A certain number of automatically identified index terms inevitably contain errors that look downright foolish to human eyes. Indexes consisting of automatically identified terms have been criticized on the grounds that they constitute indiscriminate lists, rather than synthesized and structured representation of content. And because computer systems do not understand the terms they extract, they cannot record terms with the consistency expected of indexes created by human beings.

Nevertheless, the research approach that we take in this paper emphasizes fully automatic identification and organization of index terms that actually occur in the text. We have adopted this approach for several reasons:

1. **Human indexers simply cannot keep up with the volume of new text being produced.** This is a particularly pressing problem for publications such as daily newspapers which are under particular pressure to rapidly create useful indexes for large amounts of text.
2. **New names and terms are constantly being invented and/or published.** For example, new companies are formed (e.g., *Verizon Communications Inc.*); people's names appear in the news for the first time (e.g., it is unlikely that *Elian Gonzalez*' name was in a newspaper before November 25, 1999); and new product names are constantly being invented (e.g., *Handspring's Visor PDA*). These terms frequently appear in print some time before they appear in an authoritative reference source.

3. **Manually created external resources are not available for every corpus.** Systems that fundamentally depend on manually created resources such as controlled vocabularies, semantic ontologies, or manually annotated text usually cannot be readily adopted to corpora for which these resources do not exist.
4. **Differing indexing standards across agencies and organizations makes reconciliation of indexes a difficult and time consuming task.** The difficulty of reconciliation is exacerbated when indexes are prepared by different organizations for different user groups, corpora and domains (Hert et al. 2000 [11]). Under some circumstances, it may be preferable to have one large automatically generated index than none at all.
5. **Automatically identified index terms are useful in other digital library applications.** Index term lists are essential for browsing, but can also form data as input to other applications such as information retrieval, summarization and classification [25], [1].

Given these information needs, our goal is develop techniques for identifying and organizing index terms that reduce the number of terms that users need to browse and simultaneously maximize the informativeness of each term.

In this paper, we describe a method for identifying index terms for use in a dynamic text browser (DTB). We have implemented a prototype DTB called Intell-Index which supports interactive navigation of index terms, with hyperlinks to the views of phrases in context and full-text documents. The input to Intell-Index consists of noun phrases identified by LinkIT, a software tool for identifying significant topics in domain independent text. (We describe this software in more detail in Section 2.)

However, little work has been done on the question of what constitutes useful index terms (Milstead 1994 [15], Hert et al. 2000 [11]). In order to move toward our goal, we have therefore found it necessary to identify properties of index terms that affect their usefulness in an electronic browsing environment. To assess the quality of the index terms, we consider three criteria especially pertinent to automatically identified index terms: **coherence, thoroughness of coverage of document content,** and a combined metric of quality and coverage that we call **usefulness.**

- **Coherence:** Because computer systems are unable to identify terms with human reliability or consistency, they inevitably generate some number of junk terms that humans readily recognize as incoherent. We consider a very basic question: are automatically identified terms sufficiently coherent to be useful as access points to document content. To answer this question for the LinkIT output, we randomly selected .025% of the terms identified in a 250MB corpus and evaluated them with respect to their coherence. Our study showed that over 90% of the terms are coherent. Cowie and Lehnert 1996 [3] observe that 90% precision in information extraction is probably satisfactory for every day use of results; this assessment is relevant here because the terms are processed by people, who can fairly readily ignore the junk if they expect to encounter it.
- **Thoroughness of coverage of document content:** Because computer systems are more thorough and less discriminating, they typically identify many more terms than a human indexer would for the same amount of material. For example, LinkIT identifies about 500,000 non-unique terms for 12.27 MB of text. We address the issue of quantity by considering the number of terms that LinkIT identifies, as related to size of the original text from which they were extracted. This provides a basis for future comparison of the number of terms identified in different corpora and by different techniques.
- **Usefulness of index terms:** Techniques for automatically identifying index terms abound. In order to make a preliminary assessment of the usefulness of index terms identified by LinkIT, we performed an experiment to measure user's perceptions of the usefulness of index terms. We presented users with lists of index terms identified by three domain-independent techniques and with the newspaper articles from which the terms had been extracted (Wacholder et al. 2000 [22]). In terms of quality alone, our results show that the technical term extraction method of Justeson and Katz 1995 [14] receives the highest rating. However, in terms of a combined quality and coverage metric, the Head Sorting method, described in Wacholder 1998 [21] used by LinkIT outperforms both other techniques.

Although the terms identified by LinkIT are the primary focus of our analysis, these criteria can be applied systematically to terms identified by other techniques. Techniques for identifying NPs abound and they are difficult to compare (Evans et al. 2000 [7]). With this work, we seek to establish a foundation for further in-depth analysis of the usability of automatically identified terms which will include, *inter alia*, the observation of subjects performing information seeking tasks using index terms generated by different techniques.

We discuss of term coherence, usefulness and thoroughness of coverage of document content in Section 3. But before we turn to this discussion, we describe the technology that we have developed to identify, display, and structure terms.

2. Technology

2.1 Towards a DTB

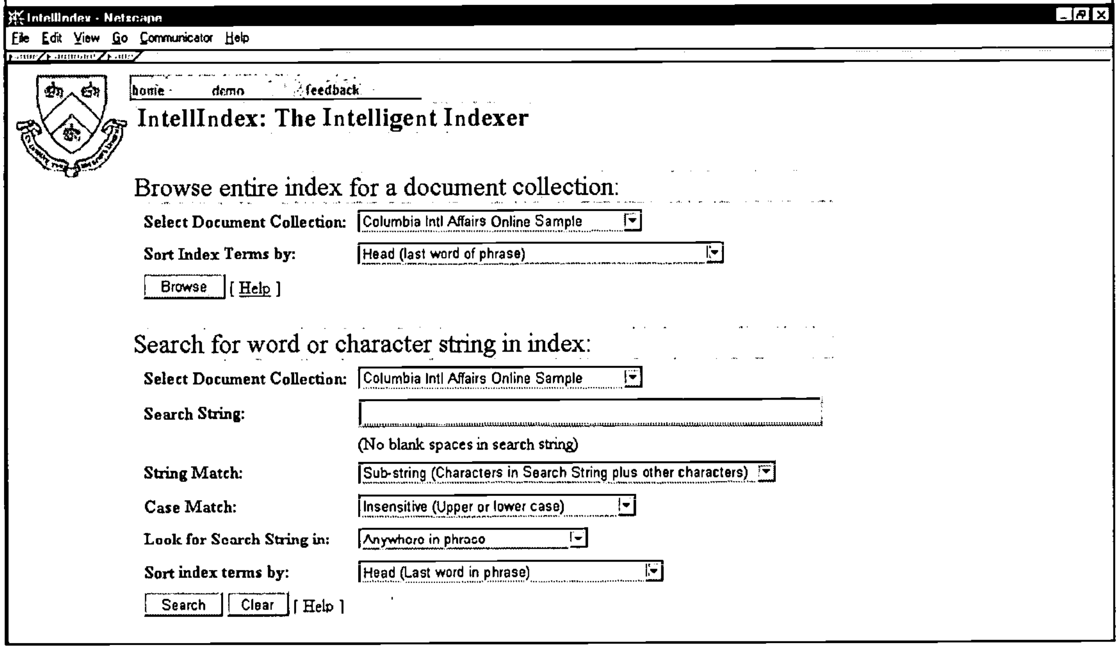
We have designed a domain-independent system called LinkIT, which uses the head sorting method to identify candidate index terms in full-text documents (Wacholder 1998 [21]; Evans 1998[6]). Using a finite state machine compiled from a regular expression grammar, LinkIT parses text which has been automatically tagged with grammatical part-of-speech a finite state machine. LinkIT can process approximately 4.11 MB tagged text per second [7], [6].

The expressions identified by LinkIT are noun phrases (NPs), coherent units whose head (most important word syntactically and semantically) is a noun. For example, *filter* is the head of the NPs *coffee filter*, *oil filter*, *smut filter*, and *water filters at warehouse prices*.

At the present time LinkIT identifies a subset of NPs that occur in a document, simplex NPs. A complex NP *a form of cancer-causing asbestos* actually includes two simplex NPs, *a form* and *cancer-causing asbestos*. A system that lists only complex NPs would list only one term, a system that lists both simplex and complex NPs would list all three phrases, and a system that identifies only simplex NPs would list two. LinkIT identifies Simplex NPs rather than complex ones for a practical reason: Simplex NPs can be identified more reliably because they are structurally more simple. Compared to simplex NPs, the boundaries of complex NPs (e.g., *symptoms that*

Figure 1: Intell-Index opening screen

< http://www.cs.columbia.edu/~nina/IntellIndex/indexer.cgi>



IntellIndex: The Intelligent Indexer

Browse entire index for a document collection:

Select Document Collection: Columbia Intl Affairs Online Sample

Sort Index Terms by: Head (last word of phrase)

Browse [Help]

Search for word or character string in index:

Select Document Collection: Columbia Intl Affairs Online Sample

Search String:

(No blank spaces in search string)

String Match: Sub-string (Characters in Search String plus other characters)

Case Match: Insensitive (Upper or lower case)

Look for Search String in: Anywhere in phrase

Sort index terms by: Head (Last word in phrase)

Search Clear [Help]

crop up decades later) are difficult to identify. The head is also more difficult to identify: for example there are cases, such as *group of children* where the syntactic head (*group*) is distinct from the semantic head (*children*). Complex NPs can be difficult for people to interpret, especially out of context. For example, the expression *information about medicine for babies* is ambiguous: in [[information about medicine] [for infants]], the information is for infants; in [information about [medicine for infants]], the medicine is for infants. The decision to include only simplex NPs in the DTB has important implications for the number of index terms included in the DTB, as discussed below.

Finally, LinkIT sorts the NPs by head, and ranks them in terms of their significance based on head frequency. The intuitive justification for sorting SNPs by head is based on the fundamental linguistic distinction between head and modifier: in general, a head makes a greater contribution to the syntax and semantics of a phrase than does a modifier. This linguistic insight can be extended to the document level. If, as a practical matter, it is necessary to rank the contribution to a whole document made by the sequence of words constituting a documentre, the head should be ranked more highly than other words in the phrase. This distinction is important in linguistic theory; for example, Jackendoff 1977 [13] discusses the relationship of heads and modifiers in phrase structure. It is also important in NLP, where, for example, Strzalkowski 1997 [19] and Evans and Zhai 1996 [5] have used the distinction between heads and modifiers to add query terms to information retrieval systems.

Powerful corpus processing techniques have been developed to measure deviance from an average occurrence or co-occurrence in the corpus. In this paper we chose to evaluate methods that depend only on document-internal data, independent of corpus, domain or genre. We therefore did not use, for example, tf*idf, the purely statistical technique that is the used by most information retrieval systems, or Smadja 1993 [17], a hybrid statistical and symbolic technique for identifying collocations.

We have incorporated LinkIT output into a prototype DTB called Intell-Index.
(http://www.cs.columbia.edu/~nina/IntellIndex/indexer.cgi). Figure 1 above shows the Intell-Index opening screen. The user selects the collection to be browsed and then may browse the entire set of index terms identified by LinkIT. (Note that this "collection" could also arise from the results of a search.) Alternatively, the user may enter a term, and specify criteria to select a subset of terms that will be returned (e.g. heads only, or modifiers and heads). This gives the user better control over the results so that browsing is more effective.

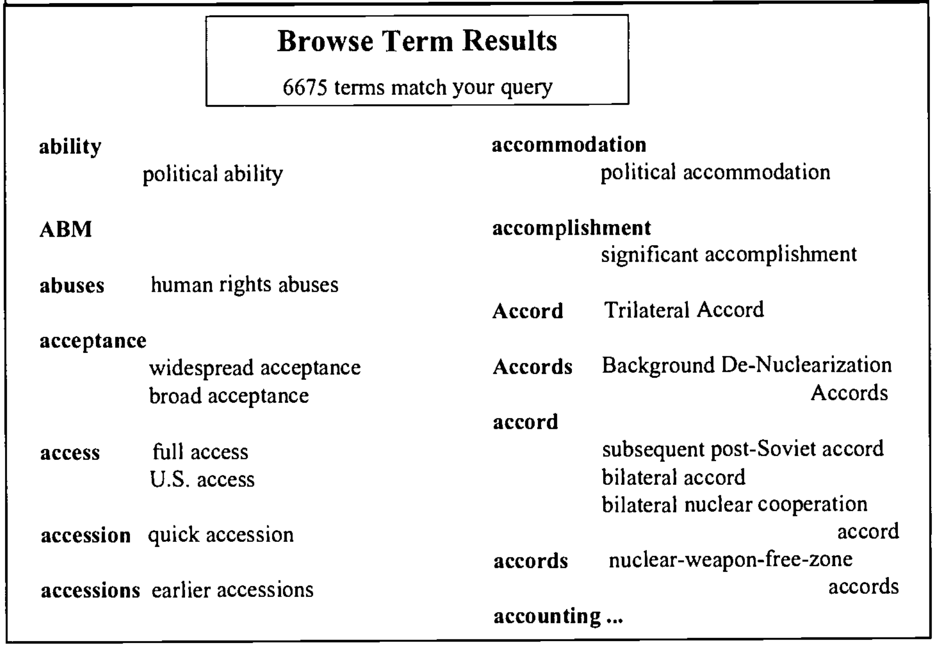
Figure 2 on p.4 shows the beginning of the alphabetized browsing results for the specified corpus. As the user browses the terms returned by Intell-Index, she may choose to view a list of the contexts in which the terms are used; these contexts are sorted by document and ranked by normalized frequency in the document. This view is called index term in context (ITIC) based on its relationship to a simpler version, i.e. keyword in context (KWIC). If the information seeker decides that the list of ITICs is promising, she may view the entire document, or browse another view of the data.

At the present time, the DTB uses only Simplex NPs. However, LinkIT collects information for conflating simplex NPs into complex ones; this will be added to Intell-Index at a later date.

2.2 The head sorting technique

A key advantage of using Simplex NPs rather than Complex ones is that, at least in English, the last word of the NP is reliably the head (Wacholder 1998 [21]). Repetition of heads of phrases in a

Figure 2: Browse term results



Browse Term Results

6675 terms match your query

ability political ability

ABM

abuses human rights abuses

acceptance widespread acceptance
broad acceptance

access full access
U.S. access

accession quick accession

accessions earlier accessions

accommodation political accommodation

accomplishment significant accomplishment

Accord Trilateral Accord

Accords Background De-Nuclearization Accords

accord subsequent post-Soviet accord
bilateral accord
bilateral nuclear cooperation accord

accords nuclear-weapon-free-zone accords

accounting ...

document indicates that the head represents an important concept in the document. As a result, no additional information other than that extracted from the document is required to sort the NPs by head.

Information about frequency with which nouns are used as heads in documents can be used to provide the users with useful views of the content of a single document or a collection of documents. Table 1 shows the topics are identified as most important in a single article using the head sorting technique (*Wall Street Journal 1988* [23]). Heads of terms are italicized.

asbestos *workers*
cancer-causing *asbestos*
cigarette *filters*
researcher(s)
asbestos *fiber*
crocidolite
paper *factory*

Table 1: Most significant terms in document

For example the list of phrases (which includes heads that occur above a frequency cutoff of 3 in this document, with content-bearing modifiers, if any) is a list of important concepts representative of the entire document.

Another view of the phrases enabled by head sorting is obtained by linking NPs in a document with the same head. A single word NP can be quite ambiguous, especially if it is a frequently-occurring noun like *worker*, *state*, or *act*. NPs grouped by head are likely to refer to the same concept, if not always to the same entity (Yarowsky 1993 [24]), and therefore convey the primary sense of the head as used in the text. For example, in the sentence "Those workers got a pay raise but the other workers did not", the same sense of *worker* is used in both NPs even though two different sets of workers are referred to. Table 2 shows how the word *workers* is used as the head of a NP in four different Wall Street Journal articles from the Penn Treebank; determiners such as *a* and *some* have been removed.

workers ... asbestos workers (wsj 0003)
workers ... private sector workers ... private sector hospital workers ... nonunion workers...private sector union workers (wsj 0319)
workers ... private sector workers ... United Steelworkers (wsj 0592)
workers ... United Auto Workers ... hourly production and maintenance workers (wsj0492)

Table 2: Comparison of uses of *worker* as head of NPs across articles

This view distinguishes the type of *worker* referred to in the different articles, thereby providing information that helps rule in certain articles as possibilities and eliminate others. This is because the list of complete uses of the head *worker* provides explicit positive and implicit negative evidence about kinds of workers discussed in the article. For example, since the list for wsj_0003 includes only *workers* and *asbestos workers*, the user can infer that hospital workers or union workers are probably not referred to in this document.

Term context can also be useful if terms are presented in document order. For example, the index terms in Table 3 were extracted automatically by the LinkIT system as part of the process of identification of all NPs in a document (Evans 1998 [6]; Evans et al. 2000[7]).

A form
asbestos
Kent cigarette filters
a high percentage
cancer deaths
a group
workers
30 years
researchers

Table 3: Topics, in document order, extracted from first sentence of *Wall Street Journal* article

For most people, it is not difficult to guess that this list of terms has been extracted from a discussion about deaths from cancer in workers exposed to asbestos. The information seeker is able to apply common sense and general knowledge of the world to interpret the terms and their possible relation to each other. At least for a short document, a complete list of terms extracted from a document in order can relatively easily be browsed in order to get a sense of the topics discussed in a single document.

In the next section we assess, qualitatively and quantitatively, the usability of automatically indexed terms identified by LinkIT. The focus of this discussion is Simplex NPs; in future work, we will discuss techniques for extending the informativeness of Simplex NPs.

3. Assessment of automatically identified index terms

The problem of how to determine what index terms merit inclusion in a DTB is a difficult one. The standard information retrieval metrics of precision and recall do not apply to this task because indexes are designed to satisfy multiple information needs. In information retrieval, precision is calculated by determining how many retrieved documents satisfy a specific information need. But indexes by design include index terms that are relevant to a variety of information needs. To apply the recall metric to index terms, we would calculate the proportion of good index terms correctly identified by a system relative to the list of all possible good index terms. But we do not know what the list of all possible good index terms should look like. Even comparing an automatically generated list to a human generated list is difficult because human indexers add index entries that do not appear in the text; this would bias the evaluation against an index that only includes terms that actually occur in the text. We have therefore identified three basic criteria that affect usability of index terms: coherence of terms, thoroughness of coverage of document content, and usefulness.

3.1 Coherence

For index terms to be useful, they must be coherent. This criterion is particular important because any list of automatically identified index terms inevitably includes some junk. An index with less junk terms is clearly superior to one with more junk.

To assess the coherence of automatically identified index terms, 583 index terms (.025% of the total number of terms identified) were randomly extracted from the 250 MB corpus and alphabetized. Each term was assigned one of three ratings:

- **coherent** -- a term is both coherent and an NP. Coherent terms make sense as a distinct unit, even out of context. Examples of coherent terms identified by LinkIT are *sudden current shifts*, *Governor Dukakis*, *terminal-to-host connectivity* and *researchers*.
- **incoherent** – a term is neither a NP nor coherent. Examples of incoherent terms identified by LinkIT are *uncertainty is*, *x ix limit*, and *heated potato then shot*. Most of these problems result from idiosyncratic or non-standard text formatting. Another source of errors is the part-of-speech tagger; for example, if it erroneously identifies a verb as a noun (as in the example *uncertainty is*), the resulting term is incoherent:
- **intermediate** – any term that does not clearly belong in the coherent or incoherent categories. Typically they consist of one or more good NPs, along with some junk. In general, they are enough like NPs that in some ways they fit patterns of the component NPs. One example is *up Microsoft Windows*, which would be a coherent term if it did not include *up*. We include this term because the term is coherent enough to justify inclusion in a list of references to Windows or Microsoft. Another example is *th newsroom*, where *th* is presumably a typographical error for *the*. There are a higher percentage of intermediate terms among proper names than the other two categories; this is because LinkIT has difficulty of deciding where one proper name ends and the next one begins, as in *General Electric Co. MUNICIPALS Forest Reserve District*.

Table 4 shows the ratings by type of term and overall. The percentage of useless terms is 6.5%. This is well under 10%, which puts our results in the realm of being suitable for everyday use according to the Cowie and Lehnert metric mentioned in Section 1.

	Total	Cohe-rent	Inter-mediate	Inco-herent
Number of words	574	475	62	37
% of total words	100%	82.8%	10.9%	6.5%

Table 4: Coherence of index terms

This study demonstrates that automatically identified terms like those identified by LinkIT are of sufficient quality to be useful in browsing applications.

3.2 Usefulness of index terms

In order to make a preliminary determination of whether the terms identified by LinkIT as likely to be useful index terms, we used a standard qualitative ranking technique to compare the usefulness of terms identified by three domain-independent techniques methods for identifying index terms (Wacholder et al. 2000 [22]):

- **Keywords** are terms identified by counting frequency of stemmed words in a document.
- **Technical terms** are noun phrases (NPs) or subparts of NPs repeated more than twice in a document (Justeson and Katz 1995 [14]);

- **Head sorted NPs** are identified by a method in which simplex noun phrases (as defined below) are sorted by head and then ranked in decreasing order of frequency (Wacholder 1998 [21]).

Table 5 shows examples of the index terms identified by the different techniques. All technical terms are included; a sample of the terms identified by the other two techniques are included.

Keywords	Head sorted NPs	Technical terms
asbestos/asbestosis	workers	cancer deaths
worker/workers /worked	asbestos workers	lung cancer
cancer	160 workers	kent cigarette
death	cancer	dr. talcott
make	lung cancer	cigarette filter
lorillard	asbestos	u.s.
fiber	cancer causing asbestos	
dr.	lung cancer deaths	
...	...	

Table 5: Examples of terms, by technique for one document

To compare the index terms, we presented subjects with an article from the *Wall Street Journal* [23] and a list of terms and asked them to answer the following general question: "Would this term be useful in an electronic index for this article?" Terms were rated on a scale of 1 to 5, where 1 indicates a high quality term that should definitely be included in the index and 5 indicates a junk term that definitely should not be included. For example, the phrase *court-approved affirmative action plans* received an average rating of 1 meaning that it was ranked as definitely useful; the keyword *affirmative* received an average rating of 3.75, meaning that it was less useful; and the keyword *action* received an average ranking of 4.5, meaning that it was not useful. Table 6 shows the results, averaged over three articles.

	Keywords	Head Sorted NPs	Technical Terms
Average ranking	**3.27**	**2.89**	**1.79**

Table 6 : Average rating of types of index terms

Of the three lists of index terms, technical termss received the highest ratings for all three documents—an average of 1.79 on the scale of 1 to 5, with 1 being the best rating. The head sorted NPs came in second, with an average of 2.89, and keywords came in last with an average of 3.27.

However, it should be noted that the average ranking of terms conceals the fact that the number of technical terms is much lower than the other two types of terms. In contrast, Table 7, which shows the total number of terms rated at or below specified rankings, allows us to measure quality and coverage. (1 is the highest rating; 5 is the lowest.)

	Number of terms ranked at or better than			
Method	**2**	**3**	**4**	**5**
Keyword	27	75	124	166
Head sorted NPs	41	96	132	160
Technical Terms	15	21	21	21

Table 7: Running total of terms identified at or below a specified rank

This result is consistent with our observation that the technical term method identifies the highest quality terms, but there are very few of them: an average of 7 per 500 words compared to over 50 for head sorted NPs and for keywords. Therefore there is a need for additional high quality terms. The list of head sorted NPs received a higher average rating than did the list of keywords, as shown in Figure 2. This confirms our expectation that phrases containing more content-bearing modifiers would be perceived as more useful index terms than would single word phrases consisting only of heads.

3.3 Thoroughness of coverage of document content

Thoroughness of coverage of document content is a standard criterion for evaluation of traditional indexes [11]. In order to establish an initial measure of thoroughness, we evaluate number of terms identified relative to the size of the text.

Table 8 shows the relationship between document size in words and number of NPs per document. For example, for the AP corpus, an average document of 476 words typically has about 127 non-unique NPs associated with it. In other words, a user who wanted to view the context in which each NP occurred would have to look at 127 contexts. (To allow for differences across corpora, we report on overall statistics and per corpus statistics as appropriate.)

Corpus	Avg. Doc Size	Avg. number of NPs/doc
AP	2.99K (476 words)	127
FR	7.70K (1175 words)	338
WSJ	3.23K (487 words)	132
ZIFF	2.96K (461 words)	129

Table 8: NPs per document

The numbers in Table 8 are important because they vary depending on the technique used to identify NPs. A human indexer readily chooses whichever type of phrase is appropriate for the content, but natural language processing systems cannot do this reliably. Because of the ambiguity of natural language, it is much easier to identify the boundaries of simplex noun than complex ones [21], as discussed above in Section 2 above.

The option of including both complex and simple forms was adopted by Tolle and Chen 2000 [20]. They identify approximately 140 unique NPs per abstract for 10 medical abstracts. They do not report the average length in words of abstracts, but a reasonable guess is probably about 250 words per abstract. On this calculation, the ratio between the number of NPs and the number of words in the text is .56. In contrast, LinkIT identifies about 130 NPs for documents of approximately 475 words, for a ratio of .27. The index terms represent the content of different units: 140 index terms represents the abstract, which is itself only an abbreviated representation of the document. The 130 terms identified by LinkIT represent the entire text, but our intuition is that it is better to provide coverage of full documents than of abstracts. Experiments to determine which technique is more useful for information seekers are needed.

For four large full-text corpora, we extracted all occurrences of all NPs (duplicates not removed) in each corpus, and then we list the size of the index when duplicate NPs have been removed in Table 9. The numbers in parenthesis are the number of words per document and per corpus for the full-text columns, and the percentage of the full text size for the list of non-unique NPs, and list of unique NPs.

Corpus	Full Text	Non Unique NPs	Unique NPs
AP	12.27 MB (2.0 million words)	7.4 MB (60%)	2.9 MB (23%)
FR	33.88 MB (5.3 million words)	20.7 MB (61%)	5.7 MB (17%)
WSJ	45.59 MB (7.0 million words)	27.3 MB (60%)	10.0 MB (22%)
ZIFF	165.41 MB (26.3 million words)	108.8 MB (66%)	38.7 MB (24%)

Table 9: Corpus Size

The number of NPs reflects the number of occurrences (tokens) of NPs. Interestingly, the percentages are relatively consistent across corpora.

From the point of view of the index, however, the first column in Table 9 represent only a first level reduction in the number of candidate index terms: for browsing and indexing, each term need be listed only once. After duplicates have been removed, approximately 1% of the full text remains for heads, and 22% for NPs. The implications of this are explored in Section 4.

4. Reducing documents to NPs

In this section, we consider how information about NPs and their heads can help facilitate effective browsing by reduce the number of terms that an information seeker needs to look at.

In general, the number of unique NPs increases much faster than the number of unique heads – this can be seen by the fall in the ratio of unique heads to NPs as the corpus size increases.

Corpus	Size in MBs	Unique NPs	Unique Heads	Ratio of Unique Heads to NPs
AP	12	156798	38232	24%
FR	34	281931	56555	20%
WSJ	45	510194	77168	15%
ZIFF	165	1731940	176639	10%
Total	256	2490958	254724	10%

Table 10: Number of unique NPs and heads

Table 10 is interesting for a number of reasons:

1) the variation in ratio of heads to NPs per corpus—this may well reflect the diversity of AP and the FR relative to the WSJ and especially Ziff.

2) the number of heads decreases monotically as the size of the corpus increases. This is because the heads are nouns. No dictionary can list all nouns; this list is constantly growing, but at a slower rate than the possible number of NPs.

In general, the vast majority of heads have two or fewer different possible expansions. There is a small number of heads, however, that contain a large number of expansions. For these heads, we could create a hierarchical index that is only displayed when the user requests further information on the particular head. In the data that we examined, on average the heads had about 6.5 expansions, with a standard deviation of 47.3.

Corp	Max	% <= 2	2 < % < 50	% >= 50	Avg	Std. Dev.
AP	557	72.2%	26.6%	1.2%	4.3	13.63
FR	1303	76.9%	21.3%	1.8%	5.5	26.95
WSJ	5343	69.9%	27.8%	2.3%	7.0	46.65
ZIFF	15877	75.9%	21.6%	2.5%	10.5	102.38

Table 11: Average number of head expansions per corpus

The most frequent head in the Ziff corpus, a computer publication, is *system*.

Additionally, these terms have not been filtered; we may be able to greatly narrow the search space if the user can provide us with further information about the type of terms they are interested in. For example, using simple regular expressions, we are able to roughly categorize the terms that we have found into four categories: NPs, SNPs that look like proper nouns, SNPs that look like acronyms, and SNPs that start with non-alphabetic characters. It is possible to narrow the index to one of these categories, or exclude some of them from the index.

Corpus	# of SNPs	# of Proper Nouns	# of Acronyms	# of non-alphabetic elements
AP	156798	20787 (13.2%)	2526 (1.61%)	12238 (7.8%)
FR	281931	22194 (7.8%)	5082 (1.80%)	44992 (15.95%)
WSJ	510194	44035 (8.6%)	6295 (1.23%)	63686 (12.48%)
ZIFF	1731940	102615 (5.9%)	38460 (2.22%)	193340 (11.16%)
Total	2490958	189631 (7.6%)	45966 (1.84%)	300373 (12.06%)

Table 12: Number of SNPs by category

For example, over all of the corpora, about 10% of the SNPs start with a non-alphabetic character, which we can exclude if the user is searching for a general term. If we know that the user is searching specifically for a person, then we can use the list of proper nouns as index terms, further narrowing the search space to approximately 10% of the possible terms. We regard this technique as an important first step to reducing the number of terms that users must browse in a DTB.

5. CONCLUSION

When we began working on this paper, our goal was simply to assess the quality of the terms automatically identified by LinkIT for use in electronic browsing applications. Through an evaluation of the results of an automatic index term extraction system, we have shown that automatically generated indexes can be useful in a dynamic text-browsing environment such as Intell-Index for enabling access to digital libraries.

We found that natural language processing techniques have reached the point of being able to reliably identify terms that are coherent enough to merit inclusion in a dynamic text browser: over 93% of the index terms extracted for use in the Intell-Index system have been shown to be useful index terms in our study. This number is a baseline; the goal for us and others should be to improve these numbers.

We have also demonstrated how sorting of index terms by head makes it easier to browse index terms. The possibilities for additional sorting and filtering index terms are multiple, and our work suggests that these possibilities are worthy of exploration. Our results have implications for our own work and also for research results with regard to phrase browsers referred to in Section 1.

As we conducted this work, we discovered that there are many unanswered questions about the usability of index terms. In spite of a long history of indexes as an information access tool, there has been relatively little research on indexing usability, an especially important topic vis a vis automatically generated indexes [11][15].

Among them are the following:

1. What properties determine the usability of index terms?
2. What techniques for automatically identifying index terms produce the most useful index terms?
3. How is the usefulness of index terms affected by the browsing environment, the domain of the document, and the user's expertise?
4. From the point of view of representation of document content, what is the optimal relationship between number of index terms and document size?
5. What number of terms can information seekers readily browse? Do these numbers vary depending on the skill and domain knowledge of the user?

Because of the need to develop new methods to improve access to digital libraries, answering questions about index usability is a research priority in the digital library field. This paper makes two contributions: description of a linguistically motivated method for identifying and browsing index terms and establishment of fundamental criteria for measuring the usability of terms in phrase browsing applications.

6. ACKNOWLEDGMENTS

This work has been supported under NSF Grant IRI-97-12069, "Automatic Identification of Significant Topics in Domain Independent Full Text Analysis", PI's: Judith L. Klavans and Nina Wacholder and NSF Grant CDA-97-53054 "Computationally Tractable Methods for Document Analysis", PI: Nina Wacholder.

7. References

[1] Anick, Peter and Shivakumar Vaithyanathan (1997) "Exploiting clustering and phrases for context-based information retrieval", *Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR *'97)*, pp.314-323.

[2] Bush, Vannevar (1945) "As we may think," *Atlantic Monthly.* Available from http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm

[3] Cowie, Jim and Wendy Lehnert (1996) "Information extraction", *Communications of the ACM*, 39(1):80-91.

[4] Edmundson, H.P. and Wyllys, W. (1961) "Automatic abstracting and indexing--survey and recommendations", *Communications of the ACM*, 4:226-234.

[5] Evans, David A. and Chengxiang Zhai (1996) "Noun-phrase analysis in unrestricted text for information retrieval", *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics*, pp.17-24. 24-27 June 1996, University of California, Santa Cruz, California, Morgan Kaufmann Publishers.

[6] Evans, David K. (1998) LinkIT Documentation, Columbia University Department of Computer Science Report. Available at <http://www.cs.columbia.edu/~devans/papers/LinkITTechDoc/>

[7] Evans, David K., Klavans, Judith, and Wacholder, Nina (2000) "Document processing with LinkIT", *Proc. of the RIAO Conference,* Paris, France.

[8] Furnas, George, Thomas K. Landauer, Louis Gomez and Susan Dumais (1987) "The vocabulary problem in human-system communication", *Communications of the ACM* 30:964-971.

[9] Godby, Carol Jean and Ray Reighart (1998) "Using machine-readable text as a source of novel vocabulary to update the Dewey Decimal Classification", presented at the SIG-CR Workshop, ASIS, < http://orc.rsch.oclc.org:5061/papers/sigcr98.html >.

[10] Gutwin, Carl, Gordon Paynter, Ian Witten, Craig Nevill-Manning and Eibe Franke (1999) "Improving browsing in digital libraries with keyphrase indexes", *Decision Support Systems* 27(1-2):81-104.

[11] Hert, Carol A., Elin K. Jacob and Patrick Dawson (2000) "A usability assessment of online indexing structures in the networked environment", *Journal of the American Society for Information Science* 51(11):971-988.

[12] Hodges, Julia, Shiyun Yie, Ray Reighart and Lois Boggess (1996) "An automated system that assists in the generation of document indexes", *Natural Language Engineering* 2(2):137-160.

[13] Jackendoff, Ray, (1977), *X-Bar Syntax: A Study of Phrase Structure,* MIT Press, Cambridge, MA.

[14] Justeson, John S. and Slava M. Katz (1995). "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering* 1(1):9-27.

[15] Milstead, Jessica L. (1994) "Needs for research in indexing", *Journal of the American Society for Information Science.*

[16] Mulvany, Nancy (1993) *Indexing Books*, University of Chicago Press, Chicago, IL.

[17] Nevill-Manning, Craig G., Ian H. Witten and Gordon W. Paynter (1997) "Browsing in digital libraries: a phrase based approach", *Proc. of the DL97*, Association of Computing Machinery Digital Libraries Conference, 230-236.

[18] Smadja, Frank, McKeown, Kathy, and Vasileios Hatzivassiloglou, V. (1996). "Translating collocations for bilingual lexicons: A statistical approach". *Computational Linguistics* 22 (1):1-38.

[19] Strzalkowski, Tomek. 1997. "Building Effective Queries in Natural Language Information Retrieval." *Proc. of the 5th Applied Natural Language Conference* (ANLP-97), Washington, DC. pp. 299-306.

[20] Tolle, Kristin M. and Hsinchun Chen (2000) "Comparing noun phrasing techniques for use with medical digital library tools", *Journal of the American Society of Information Science* 51(4):352-370.

[21] Wacholder, Nina (1998) "Simplex noun phrases clustered by head: a method for identifying significant topics in a document", *Proc. of Workshop on the Computational Treatment of Nominals*, edited by Federica Busa, Inderjeet Mani and Patrick Saint-Dizier, pp.70-79. COLING-ACL, October 16, 1998, Montreal.

[22] Wacholder, Nina, David Kirk Evans, Judith L. Klavans (2000) "Evaluation of automatically identified index terms for browsing electronic documents", *Proc. of the Applied Natural Language Processing and North American Chapter of the Association for Computational Linguistics (ANLP-NAACL) 2000.* Seattle, Washington, pp. 302-307.

[23] Wall Street Journal (1988) Available from Penn Treebank, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

[24] Yarowsky, David (1993) "One sense per collocation", *Proc. of the ARPA Human Language Technology Workshop*, Princeton, NJ, pp.266-271.

[25] Zhou, Joe (1999) "Phrasal terms in real-world applications". In *Natural Language Information Retrieval*, edited by Tomek Strazalowski, Kluwer Academic Publishers, Boston, pp.215-259.

Public Use of Digital Community Information Systems: Findings from a Recent Study with Implications for System Design

Karen E. Pettigrew
Assistant Professor, The Information School
University of Washington
Box 352930
Seattle, Washington, 98195-2930, USA
Voice: 206-543-6238; Fax: 206-616-3152;
kpettigr@u.washington.edu

Joan C. Durrance
Professor, School of Information
University of Michigan
304 West Hall, 550 East University Ave
Ann Arbor, MI 48109-1092
Voice: 734-763-1569; Fax 734-764-2475
durrance@umich.edu

ABSTRACT

The Internet has considerably empowered libraries and changed common perception of what they entail. Public libraries, in particular, are using technological advancements to expand their range of services and enhance their civic roles. Providing community information (CI) in innovative, digital forms via community networks is one way in which public libraries are facilitating everyday information needs. These networks have been lauded for their potential to strengthen physical communities through increasing information flow about local services and events, and through facilitating civic interaction. However, little is known about how the public uses such digital services and what barriers they encounter. This paper presents findings about how digital CI systems benefit physical communities based on extensive case studies in three states. At each site, rich data were collected using online surveys, field observation, in-depth interviews and focus groups with Internet users, human service providers and library staff. Both the online survey and the follow-up interviews with respondents were based on sense-making theory. In our paper we discuss our findings regarding: (1) how the public is using digital CI systems for daily problem solving, and (2) the types of barriers they encounter. Suggestions for improving digital CI systems are provided.

Categories and Subject Descriptors: DL impact, user studies, case studies, evaluation methods, communities of use/practice, social informatics

General Terms: measurement, performance, human factors, theory

Keywords: community information, community networks, information behavior, barriers, sense-making, qualitative methods

1. BACKGROUND

Every day, citizens require equitable and easy access to local resources that can help them deal with the myriad of situations that arise through daily living. Yet, all people--despite their occupation, education, financial status, or social ties--encounter situations where they experience great difficulties in recognizing, expressing and meeting their needs for such community information (Chatman, 1996, in press; Chen and Hernon, 1982; Dervin, et al., 1976; Durrance, 1984a; Harris and Dewdney, 1994; Pettigrew, 2000; Pettigrew et al., 1999). Financial, physical, geographic and cultural barriers also prohibit individuals from successfully seeking information. As a result, many people cannot obtain important information, access needed services, or participate fully in their community's daily life. While information technologies hold significant promise for linking individuals with information and one another, they are foreshadowed by the potential for a deeper digital divide between the information rich and the information poor.

Public libraries have long recognized the importance of community information (CI) for creating and sustaining healthy communities. Comprising three elements: survival or human services information, local information and citizen action information (Durrance, 1984b), CI can be broadly defined as:

> any information that helps citizens with their day-to-day problems and enables them to participate [in their] community. It is all information pertaining to the availability of human services, such as healthcare, financial assistance, housing, transportation, education, and childcare services; as well as information on recreation programs, clubs, community events, and information about all levels of government (Pettigrew, 1996, p. 351).

Since the 1970s public libraries have facilitated citizens' access to CI by providing information and referral (I&R) services, and through organizing and supporting community-wide information initiatives with local service providers (Baker and Ruey, 1988; Childers, 1984). The Internet, along with high-speed personal computers, modems, and graphical interfaces, has suggested new ways for libraries to facilitate citizens' information needs through digital CI systems. One such digital collaboration in which libraries have taken a leading role and is flourishing throughout the world is community networking.

Since the late 1980s libraries have played pivotal roles in developing community networks (community-wide electronic consortia) that provide citizens with equitable access to the Internet for obtaining CI and communicating with others (Cisler, 1996; Durrance, 1993, 1994; Durrance and Pettigrew, 2000; Durrance and Schneider 1996, Gurstein 2000). Often organized and designed by librarians, these digital networks provide citizens with one-stop shopping using community-oriented discussions, question-and-

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

answer forums, access to governmental, social services, and local information, email, and Internet access (Schuler, 1994; 1996). While individuals may interact with other users by posting queries, monitoring discussions, etc., CI is often a central network feature that appears in many forms: libraries, for example, may mount their databases on the Internet, while individual service providers may post information about their programs and services. Thus, the architecture of the Internet makes digital CI possible by linking information files created not only by single organizations such as libraries, but by agencies, organizations, and individuals throughout the community (and, of course, the world). This is a major departure from traditional I&R services where librarians and other CI agency staff work with files about the community that are created on an internal library system. As a result of digital CI systems via community networks people .can access CI through public library terminals while seeking help with related search problems from librarians. In short, digital systems mean that citizens can access CI anytime in any place

Despite the lauding of community networks' potential for strengthening physical communities through increased digital CI flow and civic interaction, findings from recent studies (e.g., Kraut et. al., 1999; Nie and Erbring, 2000) suggest that Internet use has the reverse effect by isolating individuals and decreasing interpersonal interaction, which gain greater importance given Putnam's (1995, 2000) observation regarding the decline of social capital in physical communities. Thus, life in an electronic world poses several fundamental problems for research. Two such questions that are only beginning to be addressed include:

(1) How do individuals use digital CI systems when seeking help for daily situations? and

(2) How do public library-community network initiatives strengthen communities?

To date, little is known about how access to digital CI systems help (or do not help) citizens with daily living, how CI affects their information behavior, and how it may or may not benefit communities. In a recent literature review (Pettigrew, Durrance, and Vakkari, 1999), we observed that research interest in citizens' use of networked CI is increasing. However, the majority of papers were applied and descriptive in nature and were based on questionnaires or analyzed transaction log data that revealed user socio-demographics and system or page use frequency (e.g., Geffert, 1993; Harsh, 1995; Harvey and Horne, 1995; Patrick, 1996, 1997; Patrick and Black, 1996a&b; Patrick et al., 1995; Schalken and Tops, 1994), which confirms Savolainen (1998). Most studies were from the professional literature and reported conflicting user and use statistics, especially regarding user socio-demographics. In this sense, the digital CI system literature has been akin to the general public library literature that Zweizig and Dervin (1977) criticized as providing little insight into the uses that people make of information and information systems. One study of particular note, however, is Bishop, et al., (1999). Through interviews and focus groups in low income neighborhoods with users and potential users of the Prairienet community network, they identified the following categories of digital CI need: community services and activities, resources for children, healthcare, education, employment, crime and safety, and general reference tools. They recommended that libraries might provide more effective digital information services if they focus on ways that complement citizens' lifestyles, constraints and information seeking patterns.

2. CURRENT STUDY

Our research questions addressed the situations that prompt citizens to use/not use digital CI systems for everyday help, the specific types of CI that they are seeking, how they deal with different barriers that they encounter, and how they are helped by the CI that they obtain. Our study also focused on how public libraries and community service providers perceive digital CI systems help their clients, their own organizations, and the community at-large. We were particularly interested in how the public's perceptions of digital CI systems related to those of service providers and librarians.

Since our study was exploratory and aimed at yielding rich data, we used multiple methods over several stages. Stage 1 comprised a national survey with 500 medium and large-sized public libraries regarding their involvement with digital CI systems. For Stage Two, we used a standard design to conduct intensive case studies in three communities (Table 1) that received national recognition for their respective community network and in which the local public library system played a leading role.

Data collection methods at each site included (a) an online survey and follow-up telephone interviews with adult community network users who access "tagged" CI web pages, along with (b) in-depth interviews, field observation and focus groups with public library-community network staff, local human service providers, and members of the public. The survey was posted (during different time periods) on the main CI page of each network. The steps we took to address methodological considerations when conducting online surveys (as discussed by Witte et al., (2000) and Zhang (2000)) are discussed in an earlier paper (Pettigrew and Durrance, 2000). The number of days each survey ran and the total number of responses for each network are summarized in Table 2.

Table 1. Overview of Data Collection Sites

Site	Counties/ Areas Served	Public Library System	Community Network	
			Name (URL)	Est.
Northeastern Illinois	Cook, DuPage, Kane, Lake, McHenry & Will	Suburban Library System	NorthStarNet (nsn.nslsilus.org)	1995
Pittsburgh, Pennsylvania	Southwestern Pennsylvania	Carnegie Library of Pittsburgh	Three Rivers Free-Net (trfn.pgh.pa)	1995
Portland, Oregon	Multnomah County	Multnomah County Library	CascadeLink (www.cascadelink.org)	1996

Table 2. Overview of Online Survey Responses

Community Network / Area Served	# Days Survey Posted	# Responses	Gender			Age Range						
			M	F	NA	18-25	25-35	36-45	46-55	56-65	66+	NA
NorthStarNet Northeastern Illinois	60 days	34	10	20	4	6	9	9	5	2	1	2
Three Rivers Free-Net Pittsburgh, Pennsylvania	90 days	123	57	61	5	10	30	22	30	15	9	7
CascadeLink Multnomah County, Portland	70 days	40	17	20	3	5	7	9	11	3	2	3
Total:	220 days	197	84	101	12	21	46	40	46	20	12	12

Both the user survey and follow-up interviews were based on Dervin's sense-making theory (c.f., Dervin, 1992; Savolainen, 1993), which comprises a set of user-centered assumptions and methods for studying the uses individuals make of information systems. It asserts that throughout daily life, people encounter gaps in their knowledge that they can only fill or bridge (in Dervin's terms) by making new sense of their situations through seeking information. Thus they use varied strategies to seek and construct information from different resources or ideas as they cope with different barriers. Sense-making facilitates the study of different aspects of information behavior. Our research included two aspects: (1) users' assessments of the helpfulness of digital CI, and (2) users' and service providers' constructions or images of these systems. Both were investigated using the micro-moment time-line technique where respondents were asked "to reconstruct a situation in terms of what happened (time-line steps) [and then] to describe each step in detail" (p. 70), which enabled us to gather and comparing the perceptions of different players regarding how CI is constructed and used through electronic communication. The framework's social constructionist orientation suggested it would be viable for studying citizens' online information behavior. In addition to the sense-making propositions, we examined our qualitative data for such themes such as indicators of social capital, and analyzed our quantitative data for such patterns as the relationship between users' perceptions of how they were helped by the digital CI and their willingness to access it again for help in similar situations.

In the remainder of this paper, we share our findings regarding: (1) how the public is using digital CI systems (i.e., their information needs), and (2) the barriers they encounter in the process. Suggestions for improving the design of digital CI systems are also discussed. In future publications we are addressing how users are helped by digital CI systems and how these systems contribute to building social capital at the individual and community levels.

3. How the Public is Using Digital CI Systems

The respondents' age groups followed a normal distribution with most respondents (71.4%) falling between the ages of 25 and 55, while slightly more women (54.6%) responded than men. Thus our findings suggest that a typical user is non-existent, socio-demographically speaking: users equally represent both genders, a distributed range of age groups, and a diverse range of occupations: from students to blue-collar workers to white-collar professionals. Moreover, our respondents comprised both first-time or novice users as well as very experienced searchers.

Our respondents reported that they use digital CI systems for many different types of situations, including those of a personal nature and those regarding the workplace. This confirms a tenet of information behavior, namely that all individuals require community information at one point or another and that it is the individual's situation that reveals most insight into information seeking and use (Harris and Dewdney, 1994). We found that users seek the following types of digital CI (in alphabetical order):

- Business
- Computer and Technical Information
- Education
- Employment Opportunities
- Financial Support
- Governmental and Civic
- Health
- Housing
- Library Operations and Services
- Local Events
- Local History and Geneaology
- Local Information (local accommodations, community features)
- Local News (weather, traffic, school closures)
- Organizations and Groups
- Other People (both local and beyond the community)
- Parenting
- Recreation and Hobbies
- Sale, Exchange, or Donation of Goods
- Social Services
- Volunteerism

These categories are markedly different from those traditionally used to classify CI needs. Moreover, they also broaden findings reported by Bishop, et al. (1999). Notable differences between our categories and those reported in CI studies conducted prior to the Internet are: (1) a strong emphasis on employment opportunities, volunteerism, and social service availability; and, (2) the inclusion of such new categories as: sale, exchange and donation of goods, local history and genealogy, local news, computer and technical information, and other people (residing both within and beyond the community).

What is the reason for this emphasis on employment information, etc., and the emergence of novel categories? Our analysis indicates that the Internet is responsible. Increased computer capabilities and online connectivity have enabled many different types of service providers to make information available about themselves that was previously unavailable or only in limited amounts via a public library's CI record. In other words, service providers are now able to share information about themselves first-hand. Prior to the Internet, such information was largely only available on paper and had to be searched manually and often through intermediaries (although many public libraries maintained electronic, in-house databases, these databases were seldom available to the public for direct end-user searching). The breadth of CI available, along with new search engine and software capabilities, has contributed to extending the notion of what CI comprises. Just as the Internet is broadening our concept of community, so it is changing the scope of community information. Due to digital CI systems, people can search for other people online, sell and trade goods, research their family history, exchange neighborhood information—and all at a faster, more immediate pace. Increased access to the Internet, and hence community network, especially via public libraries has led to an increased public awareness of what's available, what's going on and what might be found in a community. This enhanced access is undoubtedly facilitating CI flow. Whereas people once relied on conversations over backyard fences, postings on notice boards at supermarkets, and local newspapers, they are now drawing upon the capabilities of the Internet, as fueled by public library efforts, to seek and share information about their communities.

While the above categories are useful for understanding the types of digital CI that users seek, further insights are gained when one considers the actual situations. The following are just a few examples:

- Teenagers used the network to find summer employment because it has all the local job information in one place and is trusted it as a reliable, current source;
- A senior used the network to find out about an important upcoming town council meeting;
- A man looking for a local directory of gay and lesbian organizations searched the Web, but only came across national resources. The network directed him to the exact local organization he needed.
- A homebound person used the network to research his family's genealogy because its comprehensive organization of local resources, including public library, county agency and local historical association materials;
- A former resident organized a family reunion from across the country using the network to arrange everything from activities to hotels;
- A woman used the network to learn about local government information, such as current ordinances pertaining to matters ranging from trash pick up to flood damage prevention, and to identify sources of funding for a community service project intended to help a nearby low-income community;
- A man, who sometimes uses the network to find miscellaneous information, said he uses it "mostly for help with lung cancer and possible cures or ways of living longer whether it be conventional or alternative medicine."

According to sense-making theory, information needs cannot be considered in isolation of the situations that create them since any situation is likely to yield multiple information needs, i.e., information found for one aspect of a query frequently opens another, related information need. As we found, the situations for which users sought digital CI were complex and usually required multiple pieces of information. In this sense, our users described how their searches were ongoing and how they anticipated having to pose several different queries or consult multiple sources. This notion of the ongoing search is similar to Bates' (1989) "berrypicking" concept where users search for information "a bit at a time" and alter their search strategies according to what they find and what barriers they encounter.

Beyond analyzing the CI that users sought by need category and situation type, we also focused on the information's enabling aspects, i.e., the attributes of the information that would aid users in whatever it was they were trying to accomplish. This approach builds on Dervin's notion of "verbings." We derived the following "information enabling" categories for classifying the types of CI requested:

- o ***Comparing*** (similar to verifying but may come earlier in the cognitive process)
- o ***Connecting*** (how to find people with related interests, etc.)
- o ***Describing*** (services offered, cost, eligibility, etc)
- o ***Directing*** (information about where something is located or how to get somewhere)
- o ***Explaining*** (in-depth, content-oriented information that explains how something works)
- o ***Problem solving*** (information that will help bridge a gap or solve a problem)
- o ***Promoting*** (want others to know about them, e.g., that they're available for employment, that they've started a new club, etc)
- o ***Relating*** (information that is relevant to the individual's needs and situational constructs as perceived by the indiviudual).
- o ***Trusting*** (information that individuals perceive as coming from a trusted source. This is similar to high-quality CI, i.e., CI that is accurate and current, which people said they wanted)
- o ***Verifying*** (a form of corporate intelligence, people want to keep up with what their competition is doing, be aware of new trends, etc.

These "enabling" attributes provide a novel way of viewing information needs because they focus on what users are trying to accomplish for a particular situation. When considered in conjunction with (a) the user's initial need (as presented to the digital CI system by either point and click or by typing a search phrase), (b) the situation that prompted that need, and (c) what is known about the barriers that users encounter—as discussed later—these enabling categories reveal several implications for the design of digital CI systems.

4. Other Findings Regarding the Public's Online Information Behavior

Several other novel themes emerged regarding citizens' online information behavior that contribute to the literature and may aid in digital CI system design. For example, respondents indicated that they often tried other sources (e.g., friends, newspapers, telephone directories, etc.,) for help with their questions before turning to the system. Such was the case of a user from Pittsburgh, who accessed the Three Rivers Free-Net after friends and co-workers told him that it contained job listings and other sources such as local newspapers had proven unsuccessful Since the 1960s, information science research has indicated that social ties and face-to-face communication are primary sources of information, regardless of the setting (home, workplace, school, etc.). Our findings suggest that this remains the case: the Internet has not replaced the role of social ties in citizens' information behavior. During out interviews, several respondents described how they spoke about their information need or situation with a social tie before searching online. Thus, we found that the Internet is supplementing other information-seeking behaviors in addition to creating new pathways for obtaining information: the public is using digital CI systems as an additional source. Moreover, we learned that people *want* their community networks to promote social interaction by bringing people together. This notion was expressed by a user who said: "a bulletin board or someway to facilitate people meeting each other and getting around would be very helpful. I've recently moved to town and am looking for ways to meet people. Maybe a place where people could find others who are interested in a super club or playing cards, or informal sporting groups, etc."

Users also tended to be highly confident that they could find what they needed through the community network. Despite the difficulties with using the Internet noted in previous studies, such as lack of content, low retrieval rates with search engines, inaccurate information, etc., our respondents tended to perceive their community network as an ubiquitous source and gateway to all knowledge. In this sense we identified a mismatch between what users think they can obtain via the Internet and the likelihood that that information exists and can be easily located. This finding expands on a principle of everyday information behavior: that a mismatch exists between what users believe service providers offer and what they actually do (Harris and Dewdney, 1994). Another plausible explanation is that users are transferring their mental model of what public libraries contain and how they function to the Internet in general. In other words, of community networks and the Internet, users hold the same "information" expectations that they associate with public libraries. The difficulty here, of course, is that public libraries and the Internet are not the same thing: they provide different sorts of information in vastly different ways, with the roles played by professional librarians making a critical difference. An interesting and representative example of users' perceptions of the Internet and community networks came from a young man who asserted that the Internet and community network provided non-biased information—something he associated with public libraries. Later, acknowledging that sometimes information is "sensationalized," he added that he tries to balance information retrieved from the Internet with that gleaned from other sources before making a final decision.

On a different theme, it was interesting how some respondents revealed that they were searching for CI on behalf of another person (e.g., relative, friend), and not always at that person's behest. This notion of proxy searching, of gathering requested and unrequested CI for others, supports recent findings regarding the Web by Erdelez and Rioux (in press), which they describe as information encountering, and by Gross (in press), who describes how users present "imposed queries" at reference desks in public and school libraries. On many levels, it seems that the Internet has made it easier for researchers to label and identify a particular social type, one that might be best described as "information gatherers" or "monitors" to borrow from Baker and Pettigrew (1999). In our study, these active CI seekers, who may be considered somewhat akin to information gatekeepers, appeared to relish time spent browsing and poking about the community network and the Internet. But the greatest satisfaction they described was when they found something that they believed might of interest to someone else, which they would quickly pass on, either by email or in-person. Hence, a distinguishing feature of these CI gatherers is that they are socially connected or active, and, perhaps more importantly, are aware of the potential CI needs or interests of the people with whom they interact. These CI gatherers do not wait for someone to say "I need to know about X."; instead, they take mental notes of what's going on in the lives of the people around them, their interests and situations, and then keep an eye out for CI that might be of interest or helpful—not by initiating an actual, purposive search. In this sense, CI monitors are able to recognize the potential CI needs of the people around them. Another defining element of this social type is that they do not really care if the CI they pass on is actually used, and they exhibit an understanding that sometimes information is used and proven helpful at a later point in time. For systems design, this information gatherer social type has important implications. In communities, for example, that are considered information poor, individuals who represent this social type could be identified and given advance training in Internet searching as well as in how to identify information needs and how to provide information in ways that best facilitate those needs.

We also found support for Wellman's (in press; Hampton and Wellman, 2000) notion that the Internet has created "glocalization" where it is being used by individuals for both local and long-distance interaction. In our study, respondents used the community network as a personal gateway to websites located throughout the world, while people far beyond the network's physical home were using it to obtain local information. A woman in Florida, for example, used the Three Rivers Free-Net to locate information about seniors' housing for her elderly father who was moving to the Pittsburgh area. A different user, who was accessing the network from another region, remarked on how it helps her connect with her family: "although I haven't lived there in years, I can keep up with the events and what is going on." Respondents also expressed interest in having a strong regional and neighborhood emphasis in their networks' content.

5. Barriers to Using Digital CI Systems

The notion of barriers, which is central to the sense-making framework, represents the ways in which people are prevented or blocked from seeking information successfully. By identifying barriers, one can devise ways of improving the design of digital CI systems that facilitate users' information behavior. Our respondents were asked several open-ended questions that address types of barriers. Specifically, we asked them to explain what, if anything, would make it easier for them to find what they're looking for, and to describe any past actions they might have taken regarding their search topic.

Our analysis revealed that users encounter several types of barriers when using community networks and the Internet, in general. We labeled the main barrier as ***"Information-Related."*** Barriers that fell under this broad category included:

- *Low Retrieval Rates: Due to poor search engines and site indexing, users frequently complained that they retrieved too much CI, that search engines did not provide enough specificity (e.g., for retrieving information at the neighborhood level), and that they were challenged with discerning what was relevant to their search. Regarding the Internet, in general, one user said he didn't "like it a lot" because most sites and search engines gave him 10 billion leads that get him sidetracked until he's forgotten what it was he was looking for;*
- ***Information Overload:*** Users were often daunted by a site's layout (e.g., it appeared too busy, too many bells and whistles,

poor font and color choice, especially for those who are color-blind) and the amount of text displayed on a single screen;

- ***Poorly organized (classified):*** Users complained that they often did not find CI where they expected to find it, and that there was little cross-referencing. As one female user explained: "I have a difficult time finding this information. None of the [system] categories apply to this even though I know the entity exists. Search engines didn't help either."
- ***Out-of-date and inaccurate information:*** Users found CI that was either out-of-date or there was no way of discerning when a page was created or last updated. Inaccuracies in content were also noted;
- ***Authority:*** Without proper identifiers and author credentials or association endorsements, users said it was difficult to gauge the "quality" of the CI source, i.e., whether they should trust the CI (and its source) or not;
- ***Missing:*** Users sometimes commented that information was missing although it was described as existing at the beginning of a page or document;
- ***Dead links:*** Users were frustrated when finding a link to a page or site that they believe will be highly relevant to their information need, only to find that the link is inactive or otherwise unavailable;
- ***Language used:*** Beyond most information appearing in English only, users also commented on how some sites contained information that was written using jargon or at a level that was too high for many to understand;
- ***Security:*** Users want strong evidence that the information they submit and retrieve is confidential—"reasurred security"—as one user phrased it;
- ***Specificity:*** Users want to be able to search for information at the neighborhood level. As one user explained, "what's the use of providing information concerning neighborhoods if you then don't make it easy for someone to determine exactly which neighborhood they're in or belong to?"
- ***Non-anticipatory systems:*** Although users were unable to articulate this barrier themselves, their responses in the surveys and interviews indicated that users' information behavior would be greatly facilitated if digital CI systems were "smart enough" either to anticipate their next information need (based on the need posed to the system by typed query or by point and click) or a related information need. All too often users described how the site they found was not quite what they were looking for but they did not know where to go to next.

These information-related barriers point to problems as well as potential solutions for improving the usability and helpfulness of digital CI systems. However, ***other barriers*** that users encounter also emerged from our analysis. Such barriers included:

- ***Technological barriers:*** computer connection speeds were very slow, software worked slowly or unavailable or incompatible with connecting systems, etc.;
- ***Economic barriers:*** users who could not afford their own computing equipment or online access felt they felt were at a disadvantage unless they were able to access equipment at a public library or other public computing site, which even at the best of times, was still not as convenient as having a home system;
- ***Geographic barriers:*** People were hindered in accessing computers because they lived far away from a public library or other public access site, or because high-speed connectivity was unavailable in their area;
- ***Search skill barriers:*** Community network users did not know how to search the system (or Internet in general) or how to use advanced methods. This was reflected by several respondents, one of whom commented "I have a hard time finding information even though I think I'm a pretty savvy web surfer;"
- ***Cognitive barriers:*** Users did not understand how the Internet works in terms of how it is indexed and how search engines work, how links are created, who creates and manages the information, how sites are updated, etc. As one user explained "I am not Internet savvy enough to know what would make it easier—I just muddle through," while another remarked: "there is probably more to the website that I know about;"
- ***Psychological barriers:*** Users frequently expressed a lack of confidence in their own ability to find needed information. In other words, they internalized their search failures: instead of attributing them to the Internet or just a plain lack of availability, they believed the reason they could not find something was because they were unable to carry out the search successfully.

These barriers are highly significant because they represent the impediments that users encounter when seeking information. People who are job seeking, for example, feel that they cannot get ahead unless they have access to a computer, not only so they can become more computer literate, but also because that's how they perceive people learn about job opportunities these days. For any one situation or information need, a user might be confronted by several barriers, which, collectively, can overwhelm the user and prevent him or her from locating needed information.

6. Discussion

Our analysis of users' online information behavior reveals a rich portrait of how individuals are getting faster access to more detailed information in ways that were never possible, even a decade ago due to digital CI system initiatives. These systems are valued and used by all segments of the adult population, and enable individuals, from near and far, to find information about local services and events, and facilitate different types of information seeking. Our analysis of the situations that create users' needs for CI revealed a plethora of rich findings that expand on previous reports, and, more importantly, signify several novel ways in which people are seeking CI at the turn of the century by drawing upon new technologies supported by public libraries. However, our results also indicated that users' mental models of what information exists, is retrievable, and is accurate on the Internet are overly optimistic. Although many barriers are associated with digital CI system access, these same barriers can reveal optimal solutions that will assist in creating even stronger and more information literate communities. Our findings suggest the following ways in which digital CI systems might be improved:

1. Provide users with greater specificity in their searches by improving the capability of search engines and searchable fields. Users, for example, want to be able to search for CI by neighborhood and zipcode, which reflects their notions of community.
2. Incorporate anticipatory search features that offer users suggestions or "next steps" on other types of information that are related to the information currently retrieved. For example, if a user is searching for genealogical information, then the system could suggest other sources of genealogical

information as well as genealogical software for family tree building, etc. This heuristic approach might be developed by querying the user about the context of his/her search, and by linking categories of CI based on users' perceptions of CI and how these categories are used or connected in real-life situations.

3. Query the user automatically regarding the enabling aspect(s) of the information that they are seeking and then use this data to provide information holistically. For example, if a user is seeking "directing" information, then the system might also bring up local bus schedules and routes, directions, etc., through a geographic information system.
4. Use a community information taxonomy, such as Sales (1994), for organizing and indexing CI records and make the taxonomy available online as part of the digital CI system.
5. Follow established interface design principles, such as those proposed by Head (1999) and Raskin (2000), to reduce incidents of information overload. Incorporating easy-to-use search engines that have different levels of search sophistication and following solid design standards can contribute greatly to reducing users' frustrations with pages that appear "too busy" and list too much text.
6. Indicate when the CI displayed on a page was last updated.
7. Indicate the CI source and that person's credentials.
8. Ensure that pages contain the information indicated as therein on higher level screens, i.e., ensure that pages actually contain the contents as described on introductory screens.
9. Remove dead links regularly by implementing periodic checking and updating practices.
10. Use appropriate language when providing CI that is understandable to users.
11. Provide help mechanisms that explain the very basics, i.e., how the digital CI system and Internet are organized and function, how search engines work, etc., and explain that sometimes information is unavailable at no fault of the user.
12. Provide users with contact information (email and phone number) for someone who can assist with matching their information needs to the system and with general system use.
13. Incorporate mid-way features that allow the systems to be used by people with slower machines, etc.
14. Incorporate more ways of linking people together to facilitate social interaction via bulletin boards, etc.

By carefully considering the information needs and seeking behavior of users when designing digital information systems, many of the barriers noted earlier can be avoided or greatly reduced. Systems that anticipate related information needs and the actual activities or functions that users are trying to accomplish can go even further in facilitating users' online information behavior.

7. References

[1] Baker, L. M., and Pettigrew, K. E. Theories for practitioners: Two frameworks for studying consumer health information-seeking behavior. Bulletin of the Medical Library Association, 87, 4 (1999), 444-50.

[2] Baker, S. L., and Ruey, E. D. Information and referral services — attitudes and barriers: A survey of North Carolina public libraries. Reference Quarterly, 28, 3 (1988), 243-52.

[3] Bates, M. The design of browsing and berrypicking techniques for the online search interface. Online Review. 13, 5 (1989), 407-424.

[4] Bishop, A. P., Tidline, T., Shoemaker, S., and Salela, P. Public libraries and networked information services in low-income communities. Libraries and Information Science Research, 21, 3 (1999), 361-90.

[5] Chatman, E. A. Framing social life in theory and research. In L. Höglund (Ed.), ISIC 2000: The Third International Conference on Information Seeking in Context. Gothenburg, Sweden, 16-18 August 2000. London, England: Taylor Graham.

[6] Chatman, E. A. The impoverished life-world of outsiders. Journal of the American Society for Information Science, 47, (1996), 193-206.

[7] Chen, C., and Hernon, P. Information Seeking: Assessing and Anticipating User Needs. Neal-Schuman, New York, 1982.

[8] Childers, T. Information and Referral: Public Libraries. Ablex, Norwood, NJ, 1984.

[9] Cisler, S. Weatherproofing a great, good place. American Libraries, 27, 9 (1996), 42-46.

[10] Dervin, B. From the mind's eye of the user: The sense-making qualitative-quantitative methodology. In J. D. Glazier and R. R. Powell (Eds.), Qualitative Research in Information Management (pp. 61-84). Libraries Unlimited, Englewood, CO, 1992.

[11] Dervin, B., et al. The development of strategies for dealing with the information needs of urban residents: Phase I: The citizen study. Final report of Project L0035J to the U.S. Office of Education. Seattle, WA: University of Washington, School of Communications. ERIC: ED 125640, 1976.

[12] Durrance, J. C. Meeting Community Needs through Job and Career Centers. Neal-Schuman, New York, 1994.

[13] Durrance, J. C. Armed for Action: Library Response to Citizen Information Needs. Neal-Schuman, New York, 1984a.

[14] Durrance, J. C. Community information services — An innovation at the beginning of its second decade. Advances in Librarianship 13, (1984b) 99-128.

[15] Durrance, J. C., et al. Serving Job Seekers and Career Changers: A Planning Manual for Public Libraries. American Library Association, Chicago, IL, 1993.

[16] Durrance, J. C., and Pettigrew, K. E. Community information: The technological touch. Library Journal, 125, 2 (2000), 44-46.

[17] Durrance, J. C., and Schneider, K. G. Public library community information activities: Precursors of community networking partnerships. Ann Arbor, MI: School of Information, University of Michigan. 1996: http://www.si.umich.edu/Community/taospaper.html.

[18] Erdelez, S., and Rioux, K. Sharing information encountered for others on the web. In L. Höglund (Ed.), ISIC 2000: The Third International Conference on Information Seeking in Context. Gothenburg, Sweden, 16-18 August 2000. London, England: Taylor Graham.

[19] Geffert, B. Community networks in libraries: A case study of the Freenet P.A.T.H. Public Libraries, 32, (1993), 91-99.

[20] Gross, M. Imposed information seeking in school library media centers and public libraries: A common behavior? In Höglund, L., (Ed.), ISIC 2000: The Third International Conference on Information Seeking in Context. Gothenburg, Sweden, 16-18 August 2000. London, England: Taylor Graham.

[21] Gurstein, M. Community Information: Enabling Communities with Information and Communications Technologies. Idea Group Publishing, London, 2000.

[22] Hampton, K., and Wellman, B. Examining community in the digital neighbourhood: Early results from Canada's wired suburb. In T. Ishida and K. Isbister (Eds.), Digital Cities: Technologies, Experiences, and Future Perspectives (pp. 475-492). Springer-Verlag, Berlin, 2000.

[23] Harris, R. M., and Dewdney, P. Barriers to Information: How Formal Help Systems Fail Battered Women. Greenwood Press, Westport, CN, 1994.

[24] Harsh, S. An analysis of Boulder Community Network usage. 1995: http://bcn.boulder.co.us/community/resources/harsh/harshproject.html.

[25] Harvey, K., and Horne, T. Surfing in Seattle: What cyber-patrons want. American Libraries, 26, (1995), 1028-1030.

[26] Head, A. J. Design Wise: A Guide for Evaluating the Interface Design of Information Resources. CyberAge Books, Medford, NJ, 1999.

[27] Kraut, R., Lundmark, V., Patterson, M., Kiesler, S., Mukopadhyay, T., and Scherlis, W. Internet paradox: A social technology that reduces social involvement and psychological well-being? American Psychologist, 53, 9 (1999), 1017–1031.

[28] Nie, N. H., and Erbring, L. Internet and Society: A Preliminary Report. Stanford Institute for the Quantitative Study of Society, Stanford, CA, 2000.

[29] Patrick, A. S. Media lessons from the National Capital FreeNet. Communications of the ACM, 40, 7 (1997), 74-80.

[30] Patrick, A. Services on the information highway: Subjective measures of use and importance from the National Capital FreeNet. 1996: http://debra.dgbt.doc.ca/services-research/survey/services/.

[31] Patrick, A. S., and Black, A. Implications of access methods and frequency of use for the National Capital Freenet. 1996a: http://debra.dgbt.doc.ca/services-research/survey/connections/.

[32] Patrick, A. S., and Black, A. Losing sleep and watching less TV but socializing more: Personal and social impacts of using the NCF. 1996b: http://debra.dgbt.doc.ca/services-research.

[33] Patrick, A. S., Black, A., and Whalen, T. E. Rich, young, male, dissatisfied computer geeks? Demographics and satisfaction with the NCF. In D. Godfrey and M. Levy (Eds.), Proceedings of Telecommunities *95:* The international community networking conference (pp. 83-107). Victoria, British Columbia: Telecommunities Canada, 1995. http://debra.dgbt.doc.ca/services-research/survey/demographics/vic.html.

[34] Pettigrew, K. E. Lay information provision in community settings: How community health nurses disseminate human services information to the elderly. The Library Quarterly, 70, 1 (2000), 47-85.

[35] Pettigrew, K.E. Nurses' perceptions of their needs for community information: Results of an exploratory study in southwestern Ontario. Journal of Education for Library and Information Science, 37, (1996), 351-60.

[36] Pettigrew, K. E., and Durrance, J. C. Community building using the 'Net: Perceptions of organizers, information providers and Internet users. Internet Research 1.0: The State of the Interdiscipline, First Annual Conference of the Association of Internet Researchers, University of Kansas, Lawrence, Kansas, September 14-17, 2000. [url: www.cddc.vt.edu/aoir/]

[37] Pettigrew, K. E., Durrance, J. C., and Vakkari, P. Approaches to studying public library Internet initiatives: A review of the literature and overview of a current study. Library and Information Science Research, 21, 3 (1999), 327-60.

[38] Putnam, R. D. Bowling Alone: The Collapse and Revival of American Community. Simon and Schuster, New York, 2000.

[39] Putnam, R. D. Bowling alone: America's declining social capital. Journal of Democracy, 6 (1995), 65-78.

[40] Raskin, J. The Humane Interface: New Directions for Designing Interactive Systems. Addison Wesley, Reading, MA, 2000.

[41] Sales, G. A Taxonomy of Human Services: A Conceptual Framework with Standardized Terminology and Definitions for the Field. Alliance of Information and Referral System and INFO LINE of Los Angeles, 1994.

[42] Savolainen, R. Use studies of electronic networks: A review of the literature of empirical research approaches and challenges for their development. Journal of Documentation, 54 (1998), 332-351.

[43] Savolainen, R. The sense-making theory: Reviewing the interests of a user-centered approach to information seeking and use. Information Processing and Management, 29 (1993), 13-28.

[44] Schalken, K., and Tops, P. The digital city: A study into the backgrounds and opinions of its residents. Paper presented at the Canadian Community Networks Conference, August 15-17, 1994. Carleton University, Ottawa. http://cwis.kub.nl/~frw/people/schalken/schalken.htm.

[45] Schuler, D. New Community Networks: Wired for Change. Addison-Wesley, New York, 1996.

[46] Schuler, D. Community networks: Building a new participatory medium. Communications of the ACM, 37 (1994), 39-51.

[47] Wellman, B. Physical place and CyberPlace: The rise of networked individualism. Journal of Urban and Regional Research, 25 (in press).

[48] Witte, J.C., Amoroso, L.M., and Pen, H. Research methodology - Method and representation in Internet-based survey tools - Mobility, community, and cultural identity in Survey 2000. Social Science Computer Review, 18, 2 (2000), 179-195.

[49] Zhang, Y. Using the Internet for survey research: A case study. Journal of the American Society for Information Science, 51, 1 (2000), 57-68.

[50] Zweizig, D., and Dervin, B. Public library use, users, uses: Advances in knowledge of the characteristics and needs of the adult clientele of American public libraries. Advances in Librarianship, 7 (1977), 231-55.

Evaluating the Distributed National Electronic Resource

Peter Brophy
Director
Centre for Research in Library & Information Management (CERLIM)
The Manchester Metropolitan University
Manchester M15 6LL
United Kingdom
+44-161-247-6153
p.brophy@mmu.ac.uk

Shelagh Fisher
Reader
Department of Information & Communications
The Manchester Metropolitan University
Manchester M15 6LL
United Kingdom
+44-161-247-6718
s.m.fisher@mmu.ac.uk

ABSTRACT

The UK's development of a Distributed National Electronic Resource (DNER) is being subjected to intensive formative evaluation by a multi-disciplinary team. In this paper the Project Director reports on initial actions designed to characterise the DNER from multi-stakeholder perspectives.

Categories and Subject Descriptors

C.2.1 Network Architecture and Design; C.2.4 Distributed Systems.

General Terms

Management, Measurement, Performance, Design, Economics, Reliability, Human Factors, Verification.

Keywords

Distributed collections; information environments; evaluation.

1. INTRODUCTION

The United Kingdom has, over the last decade, invested heavily in both infrastructure and information content to support higher education's research and teaching functions. The JANET and SuperJANET networks provide very high bandwidth (gigabit) connections. Content is supplied through a series of national-level 'deals' with public and private sector suppliers, with services being delivered from three datacentres (in Bath, Edinburgh & Manchester): a guiding principle is that such content should be 'free at the point of use'. This has helped ensure massive take-up.

Considerable experimentation has taken place with service delivery, not least through the Electronic Libraries Programme (eLib) [1] which as well as investigating specific service developments (such as e-journals, subject gateways, electronic document delivery) has explored generic issues, both through its final (3rd) phase concentration on 'hybrid libraries' and 'clumps' and through a series of supporting studies, of which the UKOLN-led MODELS workshops are probably the best–known and most influential [2].

In 1999, agreement was reached on the need to engineer a step-change in these services by moving towards an integrated information environment, in which access to any desired resources could be managed coherently. Issues such as resource description, location, request and delivery, alongside authentication and authorisation, need to be considered within a national and international framework if interoperability, coherence, sustainability and scalability are to be secured. The new environment was to be known as the Distributed National Electronic Resource (DNER).

In late 1999, a Call for Proposals (known as JISC 5/99 [3]) was issued to the higher education community in the UK. Subsequently over 40 development projects were selected for funding. In addition, a major formative evaluation project was funded with a primary aim of learning as much as possible about the impacts of the developing DNER on end-users, and feeding this learning back into the development process. The formative evaluation, known as EDNER, is led by the Centre for Research in Library & Information Management (CERLIM) at the Manchester Metropolitan University; the Centre for Studies in Advanced Learning Technology (CSALT) at Lancaster University is a partner. The project will extend over 3 years, and is funded at approx. $1,000,000 over that period [4]. This paper reports on work in progress and initial observations of the evaluation team.

2. DEFINITIONS

The first issue has been to determine exactly what the DNER is and what it is intended to become. To arrive at a clear understanding of what is intended and of its appropriateness, the evaluation team has used a dual approach, analysing the DNER's objectives, content and use and then comparing it with a variety of models of real-world services and environments. In essence the first approach is inductive, the second deductive.

An initial project workpackage, which involved garnering views from programme participants, potential users and other stakeholders, suggested that there was a wide range of perspectives, from those who regard the DNER as an e-university to those who see it as a large library or even museum. This variety of view was confirmed by documentary analysis of stated DNER project objectives.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

The JISC itself has used a variety of definitional statements about the DNER. The 'official' definition, as given in the Call for Proposals for the DNER's major development programme, is 'a managed environment for accessing quality assured information resources on the Internet' [3]. In other documents, however, JISC refers to the DNER as 'a comprehensive collection of electronic resources' [5] and 'the main academic apparatus required for research and teaching in the full range of main subject areas' [6].

3. MODELS

An alternative approach to understanding the DNER, used in parallel with that reported above, is to make deductions by mapping it to models of other services and environments. This suggests that the DNER has features which show significant commonality with ten other identifiable models, as descried very briefly below.

3.1 Publisher

As with any other service in the scholarly communication chain, the DNER has features which suggest parallels with both traditional and emerging models of publishing. For example, it needs to address the quality assurance of content and to provide facilities to enable that content to be distributed, often with payment mechanisms (*pace* the Open Archives initiative!).

3.2 Traditional Library

There are many models of traditional libraries available, but the DNER appears to be replicating or replacing some of this functionality, such as the organisation of content and its archiving (preservation), the provision of enquiry services and the provision of (in this case, virtual) study 'spaces'.

3.3 Museum

Some DNER development projects are explicitly designed to digitise museum content. Also, the traditional museum function of organising materials coherently for themed display – as well as, again, the preservation function – find direct parallels.

3.4 Digital Library

Perhaps most obviously, although definitionally it is complex, the DNER has features of a digital library: 'an organized collection of multimedia data with information management methods that represent the data as information and knowledge' [7].

3.5 Hybrid Library

The idea of the 'hybrid library' emerged during the eLib programme [8]. A model of this 'library in the twenty-first century' suggests that its key role is as an 'invisible intermediary', dynamically linking each user with exactly the information they need. To achieve this, it is highly sophisticated both in the intelligence it has about its users and in its knowledge of potential information sources.

3.6 Gateway

In DNER terms, gateways are effectively ordered lists of Internet resources; the eLib gateways, now part of the Resource Discovery Network (RDN), form a key component of the DNER. A variety of other gateways are on view.

3.7 Portal

Using definitions suggested within the DNER, a portal differs from a gateway in that the user is not directed to another site in response to a query (as, for example, when a URL displayed by a gateway is clicked). Rather the portal accepts the query, itself interrogates a series of resources, intelligently interprets the results (e.g. deduplicating) and then presents a result to the user. To date portals, on this definition, remain experimental.

3.8 Managed Learning Environment

MLEs should provide not just the communications structures, materials and online tuition needed for virtual learning, but all the support infrastructure that goes with them. It is difficult not to conclude that, unless the DNER can demonstrate interoperability with institutionally-based MLEs, it will not succeed.

3.9 e-university

Going beyond the MLE, an e-university needs a variety of services which enable it to offer and deliver its products across electronic networks. These would include, at a minimum, a brand, quality assurance, awards and financial systems. It could be argued, again, that the DNER will need to interoperate with, and possible to be integrated with, many of these functions.

3.10 dot.com

Again, the dot.com sector provides some lessons for characterisation of the DNER. In brief: dot.coms need <u>both</u> a high profile brand <u>and</u> a high quality product; excellent marketing; robust yet simple payment mechanisms; reliable and rapid delivery mechanisms. The DNER needs all of these.

4. Interim Conclusions

As indicated earlier the EDNER project is in its early stages, and what is reported here is very much work in progress. Nevertheless, it is already apparent that the task of building and evaluating national-level services is complicated by very different perspectives among key stakeholders, and by the lack of any single, clear model on which to base development and evaluative judgements. It is likely that consensus will emerge gradually both through debate on competing models and through assessment of the impact of achievements in service delivery.

5. REFERENCES

[1] http://www.ukoln.ac.uk/services/elib/

[2] http://www.ukoln.ac.uk/dlis/models/

[3] JISC 5/99 http://www.jisc.ac.uk:8080/pub99/c05_99.html

[4] http://www.cerlim.ac.uk/projects/edner.htm

[5] Joint Information Systems Committee. JISC Strategy 2001-05 (October 2000 draft). http://www.jisc.ac.uk/curriss/general/strat_01_05/draft_strat.html

[6] http://www.jisc.ac.uk:8080/dner/background/dner_intro1.pdf

[7] http://www.ccic.gov/pubs/iita-dlw/

[8] Brophy, P. & Fisher, S. The hybrid library. The New Review of Information & Library Research, 4, 1998, pp. 3-15

Collaborative Design with Use Case Scenarios

Lynne Davis
DLESE Program Center
University Corporation for Atmospheric Research
P.O. Box 3000
Boulder, CO. 80307-3000
303 497-8313
lynne@ucar.edu

Melissa Dawe
Center for LifeLong Learning and Design
Dept. Of Computer Science
University of Colorado at Boulder
303 492-4932
meliss@colorado.edu

1. ABSTRACT

Digital libraries, particularly those with a community-based governance structure, are best designed in a collaborative setting. In this paper, we compare our experience using two design methods: a Task-centered method that draws upon a group's strength for eliciting and formulating tasks, and a Use Case method that tends to require a focus on defining an explicit process for tasks. We discuss how these methods did and did not work well in a collaborative setting.

Categories and Subject Descriptors

H.3.7 **[Information Storage and Retrieval]**: Digital Libraries – *user issues.*

General Terms

Design, Experimentation, Human Factors.

Keywords

Use case, task-centered, design, collaboration, methodology.

2. INTRODUCTION

Designing a digital library system is a challenging effort; doing so collaboratively through community participation is even more so. Nevertheless, for the last twelve months we have actively engaged the community in the design of the Digital Library for Earth System Education (DLESE). DLESE serves a broad base of educators and students interested in any scientific aspect of the Earth and as such requires a systematic design approach to define requirements. Because DLESE is envisioned, designed and sustained by its community of users, such an approach for defining requirements needs likewise to be a highly collaborative effort. As we approached this problem, two design models emerged as likely candidates to provide this system: Task-centered design [3] and Use Case development [1]. Because both models use natural language to define requirements, users can read and understand them more easily than traditional software requirements documents. Also, both assume an iterative development model for rapid prototyping and flexibility.

But, while both profess to consider the needs of the users to be paramount in a successful design, they differ considerably in their methods. In July of 2000, at the DLESE Leadership Workshop at Montana State University in Bozeman, MT, our informed efforts to use the Use Case development method failed to result in the expected outcomes. We explain those methods and why we conclude that the Use Case methods were not appropriate for use in a highly collaborative design setting.

3. METHODOLOGIES

3.1 Task-centered Design

Task-centered design methodology [3], well known among the human-computer interaction (HCI) community and interface designers, provides techniques to elicit tasks from users that capture the user's goals and work setting. It centers on developing a small suite of tasks that represent, in essence, who will use the system and what they want to do. We used it prior to the workshop to elicit user tasks. Here is an excerpt from one:

> *Alan teaches an introductory undergraduate geology class. He wants to find a few photographs taken by the Sojourner of the surface of Mars and compare them to similar geographical areas on earth for his class.*

Tasks are then used to guide design and testing of the system. The process of creating a user interface based on these tasks occurs through an iterative process of design and user testing. The precise steps a user would take to accomplish a task are determined while a tangible user interface develops. This iterative design method encourages consideration of alternatives.

3.2 Use Case Design

The Use Case design methodology as defined by Alistair Cockburn [1] is popular among software engineers to define requirements in terms of a specific sequence of steps. It begins as a narrative that is much like the tasks in the Task-centered design methodology. A major difference is that the narratives are then broken down into a sequence of execution steps, called a scenario. For each narrative, many scenarios describe the different outcomes a user might experience when executing a task. Thus, Use Cases try to provide more detail about the system. However, the model does not include contextual or environmental factors that can influence use, and that can infer commitment to a design prematurely [2], before design alternatives can be considered.

In preparing for the DLESE Leadership Workshop, the Task-centered design methodology appealed to the user interface designers, while the Use Case methodology appealed to software engineers, who wanted the system requirements to be in a form

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

more readily usable by them. So, our design process involved elements of both.

4. SUITABILITY FOR COLLABORATION

Where these approaches really differ becomes apparent in our collaborative design setting. The Use Case approach assumes that the narrative for a goal is already written and is ready to be broken down into its various scenarios. The process of creating Use Cases that conform to Cockburn's model requires skill to break down a Use Case into a formatted sequence of steps and assumes a readiness on the part of the collaborators to commit to those steps. At the Leadership Workshop, we started with representative tasks that were expressed by the user community previously. The participants were asked to form groups of 4-6 and over a one-hour period, begin to articulate step-by-step scenarios to accomplish these tasks.

This approach was problematic in our collaborative setting. Few participants understood the process. To list a single sequence of steps the group had to exclude other interrelated steps, at least temporarily. The group had to accept assumptions about pre- and post-conditions to carry out the goal. Also, by starting with a given narrative, it disregarded a key need for the user community to contribute to the understanding of their tasks and work setting, and hence, for the designers to situate the system most appropriately. Finally, the collaborators were not ready to commit to a defined sequence of steps which made the group feel uneasy and hampered successful collaboration.

The Task-centered design process, on the other hand, begins by trying to understand the user and the user's situation. In the very early stages of the DLESE design, we asked our community to give us stories about how they envision the library might support them in their work. What we received were highly visionary, detailed descriptions. This approach facilitated group participation that was useful to the designers for describing user settings and system context.

The user interface developers transformed task descriptions into tangible mockup designs. Participants then used the mockups to test whether they could accomplish a task. They had something tangible about which they might suggest improvements.

5. RESULTS AND DISCUSSION

Accepting the Use Case narratives as written and having to develop the scenarios proved to be very incongruous with the needs, expectations, skills and motivation of our workshop participants. While it did elicit valuable information, participants did not develop scenarios. Instead, they reformulated user tasks. We feel there are two main reasons why this exercise didn't work.

5.1 Preparation of Participants

Participants did not understand the processes and constraints of developing Use Case scenarios. The semi-formal process was foreign to them, and despite minimal training and guidance from the facilitators, they essentially ignored it.

There was a mismatch of purpose. The cognitive characteristics that are well suited to developing Use Case scenarios were incongruous with the motivation and expectations of the group. Participants wanted to describe ways that the system can support them in getting their job done and they were not ready to exclude possibilities. They expected to discuss issues in an open forum. They were motivated to contribute at a high level and had neither the skills, nor the inclination, to develop structured scenarios.

5.2 Readiness to Commit

Participants were asked to accept the Use Case narratives in order to proceed with developing the scenarios. The narratives originated from the community but participants had no direct ownership, or "buy-in" to them. They felt they were too narrow, and did not represent all possibilities. They wanted to discuss the narratives in terms of their experience. While this dialog was useful and valuable for expanding the existing Use Cases, it indicated a lack of readiness to commit to specific scenarios.

6. CONCLUSIONS

Defining requirements for a digital library can be done collaboratively in conjunction with community members. In our experiment, the community was better suited to participate successfully in task creation rather than scenario development. Task-centered design leads quickly to a tangible, user-testable object that can serve as a picture to see how Use Case scenarios might develop. Task-centered design appears to be better suited for design collaboration in the case where the participants are not skilled in the semi-formal process of developing Use Cases, and where they are neither motivated nor committed to do so.

Cockburn argues that user interface design should come after the development of Use Cases. Our experience in collaborative design suggests that user interface design should come first through the Task-centered design process, where design alternatives can be included rather than excluded. Then, when the mockups and low-level prototypes begin to pass user testing and the users are ready to commit to a design, the scenarios could be formulated by software engineers or by those who are trained in this process to carry the design further.

7. ACKNOWLEDGMENTS

The work described here is funded under Cooperative Agreement #ATM-9732665 between the National Science Foundation (NSF) and the University Corporation for Atmospheric Research (UCAR), and NSF Grants #9978338 and #0085600, and by supplemental funding from the National Aeronautics and Space Administration (NASA). The views expressed herein are those of the author(s) and do not necessarily reflect the views of NSF, its subagencies, or UCAR.

8. REFERENCES

[1] Cockburn, A., 1997. Goals and use cases. *Journal of object-oriented programming,* (Sep 1997), pp 35-40, and (Nov-Dec 1997), pp 56-62.

[2] Green, T. R. G. and Petre, M., 1996. Usability analysis of visual programming environments. *J. Visual Languages and Computing*, 7, pp 131-174.

[3] Lewis, C. & Reiman, J., 1993. Task-centered user interface design. Available via ftp.cs.colorado.edu /pub/cs/distribs/clewis/HCI-Design-Books.

Human Evaluation of Kea, an Automatic Keyphrasing System

Steve Jones Gordon W. Paynter
Department of Computer Science
University of Waikato
Private Bag 3105, Hamilton, New Zealand
+64 7 838 4021
{stevej, paynter}@cs.waikato.ac.nz

ABSTRACT

This paper describes an evaluation of the Kea automatic keyphrase extraction algorithm. Tools that automatically identify keyphrases are desirable because document keyphrases have numerous applications in digital library systems, but are costly and time consuming to manually assign. Keyphrase extraction algorithms are usually evaluated by comparison to author-specified keywords, but this methodology has several well-known shortcomings. The results presented in this paper are based on subjective evaluations of the quality and appropriateness of keyphrases by human assessors, and make a number of contributions. First, they validate previous evaluations of Kea that rely on author keywords. Second, they show Kea's performance is comparable to that of similar systems that have been evaluated by human assessors. Finally, they justify the use of author keyphrases as a performance metric by showing that authors generally choose good keywords.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *user issues*. I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis*.

General Terms

Algorithms, Performance, Experimentation.

Keywords

keyphrase extraction, author keyphrases, digital libraries, subjective evaluation, user interface

1. INTRODUCTION

Some types of document (such as this one) contain a list of key words specified by the author. These keywords and keyphrases—we use the latter term to subsume the former—are a particularly useful type of summary information. They condense documents, offering a brief and precise description of their content. They have many further applications, including the classification or clustering of documents [12, 28], search and browsing interfaces [10, 11, 13], retrieval engines [3, 6, 14] and thesaurus construction [15, 18].

Keyphrases are often chosen manually, usually by the author of a document, and sometimes by professional indexers. Unfortunately not all documents contain author- or indexer-assigned keyphrases. Even in collections of scientific papers those with keyphrases are in the minority [13]. Manual keyphrase identification is tedious and time-consuming, requires expertise, and can give inconsistent results, so automatic methods benefit both the developers and the users of large document collections.

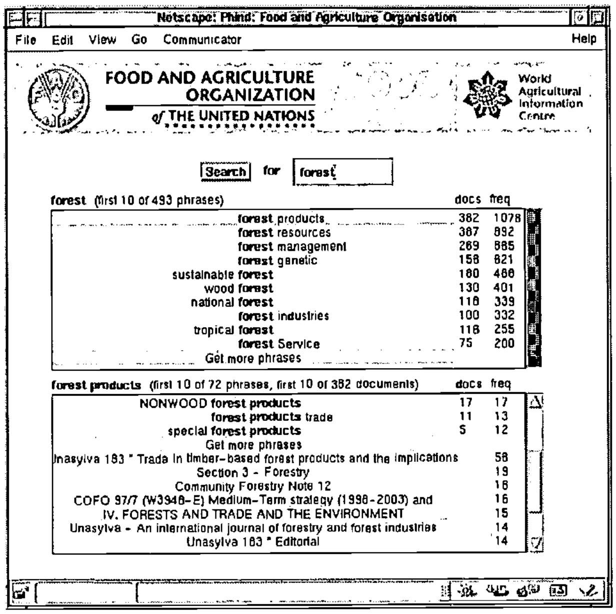
In this paper we describe a human evaluation of Kea [9, 27], an automatic keyphrase extraction algorithm developed by members of the New Zealand Digital Library Project [26]. Kea uses machine learning techniques to build a model that characterises document keyphrases, and later uses the model to identify likely keyphrases in new documents.

Previous evaluations show Kea's performance is state-of-the-art, but are weakened by their assumption that a document's author-specified keyphrases are its best possible set of keywords. In practice, the author keyphrases may not be exhaustive, and may not even be particularly appropriate—they can be chosen for purposes other than summarisation: to associate a document with a particular discipline, for example.

Our evaluation makes a number of contributions in respect of automated keyphrase extraction. First, it augments and tests the validity of the previous evaluations of Kea using a different evaluation technique—a subjective evaluation involving human assessment of the quality and appropriateness of keyphrases. Second, it compares Kea's performance as determined by human assessors against the results of similar evaluations of other systems. Finally, it investigates whether comparison against author keyphrases is a good measure of the results of keyphrase extraction systems.

In the next section of this paper we present a range of keyphrase-based interfaces developed by ourselves and others. We then describe two approaches to associating keyphrases with documents, along with techniques for keyphrase extraction, and the Kea algorithm. We discuss issues and techniques in evaluating keyphrases, providing a summary of previous research results, before proceeding to describe an experiment in which human assessors judged the quality of keyphrases generated by Kea and gathered by other means. Finally, we discuss our findings and the conclusions that we draw from the experimental results.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.


Netscape: Phind: Food and Agriculture Organisation
File Edit View Go Communicator Help
FOOD AND AGRICULTURE ORGANIZATION of THE UNITED NATIONS
World Agricultural Information Centre
Search for forest

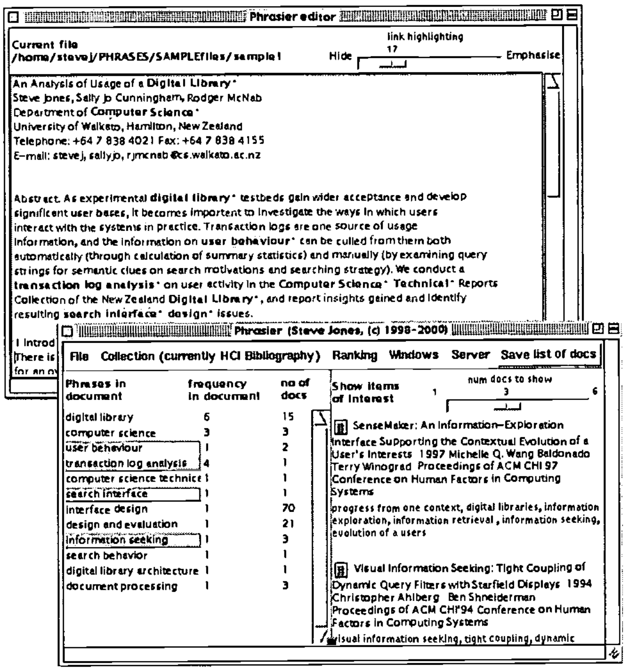
forest (first 10 of 493 phrases)

	docs	freq
forest products	382	1078
forest resources	387	892
forest management	269	885
forest genetic	158	821
sustainable **forest**	180	486
wood **forest**	130	401
national **forest**	118	339
forest industries	100	332
tropical **forest**	118	255
forest Service	75	200
Get more phrases		

forest products (first 10 of 72 phrases, first 10 of 382 documents)

	docs	freq
NONWOOD **forest products**	17	17
forest products trade	11	13
special **forest products**	5	12
Get more phrases		
Jnasylva 183 " Trade in timber-based forest products and the implications		58
Section 3 - Forestry		19
Community Forestry Note 12		16
COFO 97/7 (W3948-E) Medium-Term strategy (1998-2003) and		16
IV. FORESTS AND TRADE AND THE ENVIRONMENT		15
Unasylva - An international journal of forestry and forest industries		14
Unasylva 183 " Editorial		14

Figure 1: The Phind user interface.



Current file
/home/stevej/PHRASES/SAMPLEfiles/sample1

An Analysis of Usage of a **Digital Library**˚
Steve Jones, Sally Jo Cunningham, Rodger McNab
Department of **Computer Science**˚
University of Waikato, Hamilton, New Zealand
Telephone: +64 7 838 4021 Fax: +64 7 838 4155
E-mail: stevej, sallyjo, rjmcnab@cs.waikato.ac.nz

Abstract. As experimental **digital library**˚ testbeds gain wider acceptance and develop significant user bases, it becomes important to investigate the ways in which users interact with the systems in practice. Transaction logs are one source of usage information, and the information on **user behaviour**˚ can be culled from them both automatically (through calculation of summary statistics) and manually (by examining query strings for semantic clues on search motivations and searching strategy). We conduct a **transaction log analysis**˚ on user activity in the **Computer Science**˚ **Technical**˚ Reports Collection of the New Zealand **Digital Library**˚, and report insights gained and identify resulting **search interface**˚ **design**˚ issues.

1 Introd
There is
for an ov

Phrases in document	frequency in document	no of docs
digital library	6	15
computer science	3	3
user behaviour	1	2
transaction log analysis	4	1
computer science technic	1	1
search interface	1	1
interface design	1	70
design and evaluation	1	21
information seeking	1	3
search behavior	1	1
digital library architecture	1	1
document processing	1	3

SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests 1997 Michelle Q. Wang Baldonado Terry Winograd Proceedings of ACM CHI 97 Conference on Human Factors in Computing Systems
progress from one context, digital libraries, information exploration, information retrieval, information seeking, evolution of a users

Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays 1994 Christopher Ahlberg Ben Shneiderman Proceedings of ACM CHI'94 Conference on Human Factors in Computing Systems
visual information seeking, tight coupling, dynamic

Figure 2: The Phrasier user interface.

2. KEYPHRASE-BASED INTERFACES

Our evaluation is motivated by our use of keyphrases in user interfaces for searching and browsing. We have built a number of novel systems that use keyphrases to support new styles of interaction with digital libraries.

Phind [19] adds a browsable topic-oriented structure to collections of documents where no structure existed before—a structure that cannot be uncovered through conventional keyword queries. Users interact with a phrase hierarchy that has been automatically extracted from the documents. The phrase hierarchy resembles a paper-based subject index or thesaurus, and is presented to the user via a World Wide Web page.

The user begins by entering an initial query term, and a list of phrases that contain the term is displayed (Figure 1, top pane). When the user clicks on a phrase of interest, a further panel appears, listing longer phrases that contain the phrase, and the documents where it occurs. The user can continue to descend through the phrase hierarchy, viewing increasingly specific phrases. At each stage documents containing the phrase can be selected for display.

In Phind, users must move back and forth between result lists and document content. Another system, called Kniles, eliminates this extraneous navigation by embedding the browsing interface directly into documents as they are viewed [13].

Kniles uses keyphrases to automatically construct browsable hypertexts from plain text documents that are displayed in a conventional Web browser. Link anchors are inserted into the text wherever a phrase occurs that is a keyphrase in another document or documents. A second frame of the Web page provides a summary of the keyphrase anchors that have been inserted into the document. When a user clicks on a phrase a new web page is generated that lists the documents for which the phrase is a keyphrase. Selecting a document from the list loads it, with hyperlinks inserted, into the web browser.

Kniles is a simplified, Web-based version of Phrasier, a program that supports authors and readers who work *within* a digital library [11, 13]. In Phrasier, browsing and querying activities are seamlessly integrated with document authoring and reading tasks (see Figure 2).

Keyphrases are used to dynamically insert hypertext link anchors into the text of a retrieved document. Each anchor has two levels of *gloss* (preview information about the link destination), allowing users to navigate directly to desirable documents. Phrasier uses variable highlighting of the phrases to help users to skim the document and find sections of interest. Keyphrases are displayed more prominently than the rest of the text. Multiple keyphrases can be selected, retrieving a ranked list of documents related to the combination of selected topics.

Because links are introduced dynamically when the document is viewed users can load a document from their own filestore into Phrasier, and it will behave in the same way as documents from an established collection. In fact, the user can create a document by typing it directly into Phrasier. As the user enters text, keyphrases are identified in real time, highlighted and turned into link anchors with associated destination documents, providing immediate access to related material.

A number of other systems exploit phrases to enhance user interaction. The Journal of Artificial Intelligence Research (http://extractor.iit.nrc.ca/jair/keyphrases/) can be accessed through an interface based on phrases produced by Extractor [25]. Larkey [17] describes a system for searching a database of patent information. Within the system phrases are used to suggest query expansions to users based on the search terms that have been specified. Similarly, Pedersen *et al* [20] use phrases to support query reformulation in their Snippet Search system. Krulwich and Burkey [16] exploit heuristically extracted phrases to inform InfoFinder, an 'intelligent' agent that learns user interests during access to on-line documents.

The utility of each of these systems depends upon the availability of accurate, reliable keyphrases. The remainder of this paper shows that Kea can supply appropriate candidates.

3. ASSOCIATING KEYPHRASES WITH DOCUMENTS

There are two dominant approaches to associating keyphrases with documents: keyphrase assignment and keyphrase extraction. In keyphrase assignment (also known as text categorization) an analysis of a document leads to selection of keyphrases for that document from a controlled vocabulary [8]. It has two main advantages: the controlled vocabulary ensures that similar documents are classified consistently, and documents can be associated with concepts that are not explicitly mentioned in their text. However, there are also disadvantages: potentially useful keyphrases are ignored if they are not in the vocabulary; and controlled vocabularies require expertise and time to build and maintain, so are not always available.

In the second approach, keyphrase extraction, the text of a document is analysed and the most appropriate words and phrases that it contains are identified and associated with the document. Every phrase that occurs in the document is a potential keyphrase of the document. This approach does not require a predefined vocabulary, and is not restricted to the concepts in such a vocabulary. However, the keyphrases assigned to each document are less consistent, and it is not easy to identify the "most appropriate" words and phrases.

A wide range of techniques has been applied to the problem of phrase extraction. Turney [24, 25] uses a set of heuristics that are fine tuned using a genetic algorithm. Chen [5] uses statistical measures exploiting importance, frequency, co-occurrence and distance attributes of word pairs. Larkey [17] builds a phrase dictionary by tagging word sequences as parts of speech and retaining noun phrases. Krulwich and Burkey [16] exploit markup, such as capitalisation, emphasis, and section headings to select possibly significant phrases from documents. Anick and Vaithyanathan [2] carry out part of speech tagging and identify noun compounds—word sequences of two or more adjectives and nouns terminating in a head noun. Smeaton and Kelledy [22] identify 2 or 3 word candidate phrases from text by using stopword delimiters, and then consider phrases to be meaningful if they occur in the document collection more than some fixed number of times. Barker and Cornacchia [4] identify noun phrases using dictionary lookup, and then consider the frequency of a given noun as a phrase head within a document, discarding those that fall below a given threshold. Tolle and Chen [23] use pattern matching rules to select phrases from texts that have been tokenized and tagged by a part-of-speech tagger.

Of these approaches, Turney and Barker and Cornacchia explicitly attempt to simulate the author's choice of keywords and evaluate their methods by comparing the algorithm's choices against the author's.

4. KEA

Kea is a keyphrase extraction algorithm developed by members of the New Zealand Digital Library Project. The algorithm is substantially simpler, and therefore less computationally intensive, than many previous approaches.

Kea has been described in detail elsewhere [9, 27], and its operation is summarised below. Kea uses a *model* to identify the phrases in a document that are most likely to be good keyphrases. This model must be learned from a set of training documents with exemplar keyphrases. The exemplar phrases are usually supplied by authors, though it is also acceptable to manually provide exemplar keyphrases.

To learn a model, Kea extracts every phrase from each of the training documents in turn. Many phrases are discarded at this stage, including duplicates, those that begin or end with a stopword, those which consist only of a proper noun, those that do not match predefined phrase length constraints, and those that occur only once within a document. Three attributes of each remaining phrase are calculated: whether or not it is an author-specified keyphrase of the document, the distance into a document that it first occurs, and how specific it is to the document (its TF•IDF value). The attribute values of every phrase in every training document are used to construct a Naive Bayes classifier [7] that predicts whether or not a phrase is an author keyphrase based on its other attributes.

A range of options allows control over the model building process, and consequently the characteristics of the keyphrases that will eventually be extracted. These include maximum and minimum acceptable phrase length (in words), and an extension to the model that incorporates the number of times that phrase occurs as an author-specified keyphrase in a corpora of related documents.

Once a model for identifying keyphrases is learned from the training documents, it can be used to extract keyphrases from other documents. Each document is converted to text form and all its candidate phrases are extracted and converted to their canonical form. Many are immediately discarded, using the same criteria as described for the training process. The distance and TF•IDF attributes are computed for the remaining phrases. The Naïve Bayes model uses these attributes to calculate the probability that each candidate phrase is a keyphrase. The most probable candidates are output in ranked order; these are the keyphrases that Kea associates with the document.

The number of phrases extracted from each document can be controlled, and is typically around 10. The length of the phrases, expressed as the minimum and maximum number of words it contains, can also be controlled.

Several predefined models are distributed with Kea, including models based on generic World Wide Web pages and computer science technical reports. Previous work shows that models built for specific collections are more likely to account for the idiosyncrasies of that collection's keyphrases [9].

5. EVALUATING KEYPHRASES

There are two basic approaches to evaluating automatically generated keyphrases. The first adopts the standard Information Retrieval metrics of precision and recall to reflect how well generated phrases match phrases which are considered to be 'relevant.' Author phrases are usually used as the set of relevant phrases, or the 'Gold Standard.' This approach was adopted in previous evaluations of Kea [9, 27].

Table 1: Profile of phrases associated with each document

	Number of keyphrases			
Paper	Author	Merged Kea	Food	Combined List
1	5	49	6	58
2	6	47	6	58
3	10	51	6	66
4	8	54	6	68
5	5	51	6	57
6	7	55	6	67

There are several problems with evaluations based purely on author-chosen keyphrases. Barker and Cornacchia identify four [4]. First, author keyphrases do not always appear in the text of the document to which they belong. Second, authors choose keyphrases for purposes other than document description—to increase the likelihood of publication, for example. Third, authors rarely provide more than a few keyphrases—far fewer than may be extracted automatically. Fourth, author keyphrases are available for a limited number and type of documents.

A second approach is to gather subjective keyphrase assessments from human readers. Previous studies involving human phrase assessment [4, 5, 23, 25] follow essentially the same methodology. Subjects are provided with a document and a phrase list and asked to assess in some way the relevance of the individual phrases (or of sets of phrases) to the given document.

The study reported here adopts the second approach, and represents the first direct human evaluation of the keyphrases generated by Kea. It incorporates a human evaluation of author keyphrases, to better inform the first type of evaluation.

The evaluation had three aims. First, we wished to evaluate the keyphrases produced by Kea with a variety of models and settings. Second, we wished to compare a subjective evaluation of Kea to the results of evaluations based on the author keyphrases. Finally, we wished to determine if the author's keyphrases are a good standard against which to measure performance—do readers think the author keywords are good keyphrases?

5.1 Experimental Texts

A set of six English language papers from the Proceedings of ACM Conference on Human Factors 1997 (CHI 97; [1]) was used for the test documents. They were suitable for our purposes because they contain author-specified keywords and phrases, and provide a good fit with the background and experience of our subjects. Each paper was eight pages long.

The author's keyphrases were removed from each paper so that they would not influence extraction and assessment, and so that the papers would better represent the bulk of technical reports that do not have author keyphrases.

5.2 Subjects

Subjects were recruited from a final year course on Human Computer Interaction taken as part of an undergraduate degree programme in Computer Science. 28 subjects were recruited, of which 23 were male and five female. All had completed at least three years of undergraduate education in computer science or a related discipline and were nearing completion of a fifteen week course on human-computer interaction. The first language of 15 of the subjects was English. The youngest subject was 21, the oldest 38, and the mean age was 25.

5.3 Allocation

Two papers were allocated to each of the subjects. Papers were allocated randomly to the subjects, though presentation order, number of viewings of each paper, and subjects' first language were controlled. Two subjects chose to read only one paper during the experimental session. All other subjects were able to complete both tasks, and did so within two hours.

5.4 Instructions

The subjects were instructed to first read the paper fully. They were then told to reveal a list of phrases for the paper and asked: "How well does each of the following phrases represent what the document is either wholly or partly about?" The list of phrases was presented in the following form:

hypertext

Not at all Perfectly

0 1 2 3 4 5 6 7 8 9 10

co-citation analysis

Not at all Perfectly

0 1 2 3 4 5 6 7 8 9 10

Subjects indicated their rating by drawing a circle around the appropriate value. Subjects could refer back to the paper and reread it as often as required.

5.5 Candidate Phrase Lists

Each phrase list contained phrases from a variety of sources: Kea keyphrases extracted from the paper, author keyphrases specified in the paper, and unrelated control phrases.

Three Kea models were used to extract keyphrases. The first, *aliweb*, was trained on a set of typical web pages found by Turney [24, 25]. The second, *cstr*, is derived from a collection of computer science technical reports as described by Frank *et al.*[9]. The third, *cstr-kf*, was trained on the same documents as *cstr*, but uses a further attribute which reflects how frequently a phrase occurs as a specified keyphrase in a set of training documents. Experiments using information retrieval measures show that, averaged over hundreds of computer science documents, the *cstr* model extracts better phrases than the *aliweb* model, and that the *cstr-kf* model extracts better phrases than either [9].

The minimum phrase length was varied for each model. Two phrase sets were produced with each model, corresponding to phrases of 1–3 words and 2–3 words. The first variation reflects the way that Kea is typically used to approximate author keyphrases, and 15 phrases were extracted. The latter reflects Kea's use in Phind and Phrasier, which ignore phrases that consist of a single word.

Table 2: An example of sets of 15 keyphrases associated with paper 5

		Keywords extracted by Kea (length 1-3)			
	Author keywords	aliweb model	cstr model	cstr-kf model	Food
1	History mechanisms	revisit	revisit	**navigation**	onion
2	WWW	URL	**web**	browsers	garlic
3	web	history	**navigation**	World Wide Web	milk
4	hypertext	user	URL	browsing	ham and eggs
5	navigation	**history mechanisms**	history	patterns	pumpkin pie
6		**navigation**	**history mechanisms**	web browsers	vegetable soup
7		pages	pages	predict	
8		patterns	web pages	**WWW**	
9		**web**	browsers	empirical	
10		web pages	user	**hypertext**	
11		stack	Tauscher	accessed	
12		visited	World Wide Web	methods	
13		recency	visited	list	
14		predict	browsing	recurrence	
15		frequency	stack	actions	

Six unrelated phrases were introduced into each phrase list to enable coarse measurement of how carefully the subjects considered the task. This set consists of the names of food products.

In total there were 8 phrase sets for each paper: two phrase length variations for each of three Kea models, the author keyphrases, and the food set. The 8 sets describing each document were merged into a single master list for each paper and exact duplicates were removed. The number of phrases from each source and the total number of phrases in the list for each paper are shown in Table 1.

For every paper, there is overlap between the Kea phrase lists, and between the Kea lists and the author keyphrases. In only one paper—paper 5—was the full set of author keyphrases extracted by Kea. Table 2 shows show some of the phrase sets extracted from this paper. Phrases in bold are those that Kea extracted that are equivalent to author keyphrases (after case-folding and stemming). The shaded areas indicate the keyphrases that would be extracted using the default settings of each model. No single model found all five author phrases in the first fifteen extracted phrases.

6. RESULTS

6.1 Inter-Subject Agreement

We have measured the level of inter-subject agreement using two statistical techniques: the Kappa Statistic K and the Kendall Coefficient of Concordance W [21]. If we find significant agreement between the subjects we can rule out the hypothesis that any effects we observe occur merely by chance.

The Kappa Statistic is based on the assumption that the scores given by the assessors are (unordered) categories to which phrases are assigned. Agreement is represented by the Kappa score (K), a number that ranges from 0, which means there is no more agreement than might be expected by chance, to 1, which means the assessors are in complete agreement.

Table 3 illustrates the agreement between the subjects using the Kappa score. Three different levels of granularity are considered. First, the categories are the scores marked by the user on the 11 point scale. Second we translate subjects' 11 point responses to three categories, simulating responses of bad, average and good. The three categories are formed from the ranges 0-3, 4-6 and 7-10. Third, the 11 point responses are translated into two categories, effectively a bad/good judgement. The two points are formed by the ranges 0-5 and 6-10. Two statistics are shown for each paper: the Kappa score K, and the z score, a test of the significance of K. The number of phrases considered by each assessor for each paper is large. Therefore, across all subjects and phrases for a given paper, the Kappa values are low. We have looked beyond the absolute values and tested the significance of K (as described by Siegel and Castellan [21]), producing z scores..

As expected, the level of agreement increases as the number of categories decreases from 11 to 3 to 2. Although the values of K are small, a test of their significance shows that the inter-assessor agreement was significant at the 0.01 level in all cases.

We found substantially greater agreement between subjects than Barker and Cornacchia [4] observed in a study of keyphrase produced by *Extractor* and their system *B&C*. They reported that "on average, the judges agree only about as much as can be expected by chance". In all cases, our subjects agreed more than one would expect by chance.

A drawback of the Kappa statistic is that it considers agreement on unordered categories. As we are interested in whether one phrase is better or worse than another, not in the specific scores for each phrase, it is useful to consider agreement between the subjects' relative ranking of the phrases.

The Kendall Coefficient of Concordance (W) is a measure of agreement between rankings. As with K, it has a value between 0 (agreement as expected by chance) and 1 (complete agreement). Table 4 shows the result using the full 11 point scale. In each case, W is non-zero, indicating that there is inter-subject agreement. The X^2 score and degrees of freedom (df) can be used to determine the level of significance of the W value. The level of agreement is significant to at least the 0.01 level for all papers.

Table 3: Inter-assessor agreement measured by Kappa

	Number of points in scale					
	11		3		2	
Paper	K	z	K	z	K	z
1	0.13	14.99	0.26	16.06	0.32	14.14
2	0.14	15.74	0.32	18.14	0.39	18.37
3	0.15	17.44	0.28	15.22	0.29	10.58
4	0.08	12.48	0.14	9.85	0.16	8.86
5	0.13	15.29	0.22	13.11	0.27	11.92
6	0.16	5.88	0.29	6.50	0.37	6.69

Table 4: Inter-assessor agreement measured by the Kendall Coefficient of Concordance

Paper	*W*	X^2	*df*
1	0.63	321.03	58
2	0.70	400.67	58
3	0.63	368.90	66
4	0.32	236.11	68
5	0.38	215.65	57
6	0.72	237.81	67

The Kendall Coefficient demonstrates that there are significant (and sometimes strong) levels of agreement between the subjects when they assess the keyphrases. We conclude that subjects agree sufficiently to justify further investigation into the relative quality of the different keyphrase extraction methods.

6.2 Human Assessments

Our objective is to compare the quality of the Kea and author-specified phrases based on assessments by subjects. We do this by averaging the scores that subjects assigned to individual keyphrases derived from each source.

Figures 3 and 4 show the scores allocated by the subjects to the authors' keyphrases, various sets of Kea phrases, and the unrelated *food* phrases. The Y axes are the average score (across all subjects and all documents) assigned to phrases in the set from each source. The X axes are the number of phrases considered from each set. The leftmost point of each curve is the average score when we consider only the first phrase in a set. The rightmost point is the average score when we consider all of the phrases in a set. Intermediate points represent the average score when the first N phrases are considered The curves for the author and food sets are shorter because those sets contain fewer keyphrases that those produced by Kea.

Figure 3 shows the scores for Kea sets containing keyphrases of 1–3 words. Figure 4 shows the scores for Kea sets in which the length of keyphrases is 2–3 words. The experiment also considered phrases of length 1–4 and 2–4, but these results are not reported as they are very similar to their counterparts of length 1-3 and 2-3 respectively.

Several interesting results are revealed in the graphs. First, the phrases which are unrelated food products are rated very lowly. Overall, only nine *food* phrases received a non-zero score, and no *food* phrase was assigned a non-zero score by a subject whose first language is English.

Second, the curves for Kea sets are downward sloping for all models. The author keyphrases also follow this trend.

Third, the author keyphrases initially receive higher scores than the automatically extracted phrases. However, the scores of author keyphrases decrease more sharply than those of the Kea keyphrases, and it is only over the first two phrases that the disparity between author phrases and the best Kea phrase set is strongly apparent.

Fourth, *cstr-kf* phrases were not rated as highly as those produced by the other models. The score curves for *cstr* and *aliweb* are very similar, and are almost identical when single word keyphrases are allowed (Figure 3).

Finally we note that almost all the curves are above the mid-point (5) of the 0-10 scale used by the subjects. Subjects consistently rated the phrases positively. The exceptions to this are the *food* set, and the end of the curve produced by the *cstr-kf* model when single word keyphrases are allowed (Figure 3).

7. DISCUSSION

7.1 Integrity of Subjects' Assessments

A potential risk with such subjective and repetitive tasks is that the assessors fail to maintain a high level of discrimination throughout the process. For this reason we randomly included 'noise' phrases (the *food* set) into the phrase lists. The fact that almost all of these phrases received zero ratings, in conjunction with the agreement measures, leads us to believe that subjects gave appropriate consideration to their responses throughout the tasks. Clearly, these noise phrases are highly distinct from the topic domain of the documents. We wished to ensure that, at a coarse level, subjects had not allocated random or identical ratings to the large number of phrases they were asked to consider. The *food* phrases were unambiguously 'noisy' and served this purpose.

A second risk in this type of evaluation is that assessor agreement is so low that little can be determined from the data. For example, both Chen [5] ("Inter-indexer inconsistency is obvious in our experiment") and Barker and Cornacchia [4] ("Kappa values are spectacularly low") experienced this difficulty. However, we have established that the subjects in our experiment achieved a significant level of agreement. We attribute this to differences between experimental methodologies. Our study used documents from a restricted topic domain, with a sample population of human assessors who had a degree of knowledge about the topic, similar educational backgrounds and comparable baseline skills in the language of the evaluation documents. Studies that report lower inter-subject agreement are characterised by more diverse documents and sample populations.

7.2 Author Keyphrases as a 'Gold-Standard'

One of the aims of the evaluation was to determine whether or not comparison against author keyphrases is a good measure of automatically produced keyphrases. The results indicate that author keyphrases are consistently viewed as good representations of the subject of a document. Consequently we believe the precision and recall measures described in previous work can serve as useful

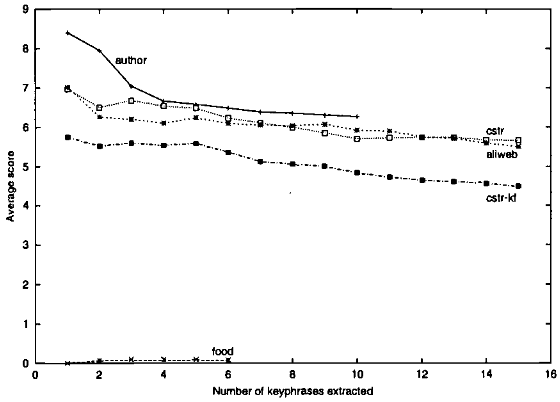


Figure 3: Average scores for phrase sets, with Kea phrases of length 1–3

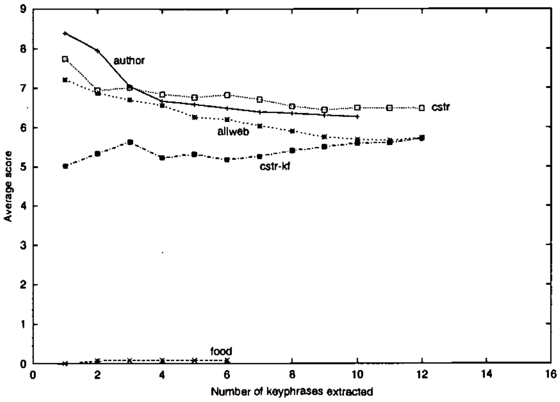


Figure 4: Average scores for phrase sets, with Kea phrases of length 2–3

indicators of the quality of automatically produced keyphrases. Of course, these methods should be adopted with an awareness of the potential problems described earlier.

Each set of author keyphrases was sorted in the order they appeared in the paper, and the resulting curves were downward sloping. We can infer that authors attempt to put the most important keyphrases first (as we might expect), and that their judgement generally matched that of the subjects in the evaluation.

The author's apparent ranking of keyphrases suggests that the basic notion of relevance in information retrieval-based evaluations—an extracted keyphrase matches an author keyphrase—may be too simplistic in some cases. Such measures might take into account the fact that there is an implicit ranking within author keyphrase lists, and consider not only *how many* author keyphrases are identified, but also the *rank* of those keyphrases.

7.3 Quality of Kea Keyphrases

Kea outputs keyphrases for a document in ranked order, and the human assessments provide some insight into the efficacy of that ordering. The downward sloping curves of the mean scores for Kea keyphrases are encouraging. The mean keyphrase score decreases as lower ranked Kea phrases are added, indicating that that Kea phrase lists are ranked effectively, and that the phrases Kea chooses first are usually the best candidates in the phrase lists.

An aim of the evaluation was to determine the effect of various Kea settings on the quality of extracted keyphrases, including the phrase length, the model employed and the use of keyphrase frequency data. One significant effect is that the use of keyphrase frequency data adversely affects keyphrase quality. The poor result is clear regardless of the phrase length and the characteristics of subjects (such as their first language). This contradicts previous studies that found that data regarding the number of times a phrase occurs as an author-specified keyphrase improves the performance of Kea [9]. These results rely on the observation that phrases that are commonly used as author keyphrases in a topic area form a pseudo-controlled vocabulary, and consequently are more likely to be selected by authors writing new papers in the same domain.

One possible explanation for the poor performance of *cstr-kf* in our study is that the domain of the model differs from the domain of the target documents. *cstr-kf* was trained on general Computer Science documents (pre 1996), and consequently favours common author-keyphrases from the training corpus. These may be inappropriate for the experimental documents, which focus on the topic of Computer-Human Interaction. The *cstr* model does not suffer from this problem, supporting other evidence that adding author-keyphrase information makes the model strongly domain-specific [9].

A second possible cause is the quality of the input texts. The *cstr-kf* model was learned on training documents that had been crudely converted from PostScript format without human intervention, resulting in texts with mistakes and poor formatting. The six documents used in the evaluation were converted from PDF to text manually with an interactive tool, resulting in substantially cleaner texts. Further, the length of the training documents varied more widely than the test set. These dissimilarities between the training and test documents may contribute to *cstr-kf's* poor performance.

7.4 Cross-language Suitability of Keyphrases

A secondary aspect of our study allows us to compare the perceived quality of keyphrases for users who do and do not have English as their first language. In fact, when we split the data based on a subject's first language we observe little difference. This is most likely a characteristic of the subject population—final year students undertaking university study in the English language—where adequate English language skills are a necessity.

7.5 Related Human Evaluations

Turney carried out a simple Web-based subjective evaluation [25] of the keyphrases produced by Extractor. Self-selecting subjects were requested to gauge keyphrases as good or bad with respect to a document that they themselves submitted, and provide feedback via a Web page form. Subjects could choose between a 'Good' or 'Bad' rating for a keyphrase. Turney reports that 82% of all keyphrases were acceptable to the subjects when seven phrases were extracted for each document. This result counts phrases with no response to be acceptable; 62% of the total phrases were judged 'Good'.

We can simulate a similar measure of acceptable phrases in our study by calculating the proportion of ratings greater than 5 that were assigned to the top seven phrases for each model. For phrase sets based on *cstr* and *aliweb* models, between 71% and 80% of the phrases were acceptable, compared to around 60% for *cstr-kf*. Of the author phrases, 79% were acceptable. This is less than the 80%

achieved by *cstr* phrases of length 2-3, and is reflected in Figure 4 where the *cstr* curve is higher than the author curve when seven phrases are extracted.

In Barker and Cornacchia's study [4], twelve subjects rated phrases produced by their *B&C* system and Turney's Extractor. They used 13 documents, nine from Turney's corpora, and four of their own choosing. Phrases were judged on a 'bad', 'so-so' or 'good' scale that was mapped to values 0, 1 and 2 respectively for analysis. The average score of an Extractor keyphrase was 0.56 (s.d. = 0.11) and of a *B&C* phrase was 0.47 (s.d. = 0.1). Subjects were more negative than indifferent about the phrases produced by both systems. By this measure, Kea compares favourably to either system, with phrases, on average, receiving positive judgements.

Chen's study [5] required assessors to choose appropriate 'subjects' to represent a document—effectively a keyphrase assignment task—in the Chinese language. Description of the captured data is limited, but it is clear that for individual subjects, intersection with the automatically extracted set ranged from 1 to 13 keyphrases, with a mean of 5.25 (s.d. = 4.18). Just over half (42 of 80) of the automatically extracted keyphrases were also selected by 8 subjects across 10 documents. This appears slightly worse than the 62% achieved by Turney, and worse than the results achieved by Kea, although direct comparison is difficult as the experimental texts are very different.

7.6 Limitations

The results reflect positively on the performance of Kea, both independently and relative to other systems. However, there are some limitations to our study. First, due to resource limitations common to evaluations of this type, the number of subjects (28) and papers (6) is limited. This is comparable with similar studies. Tolle and Chen had 19 subjects view 10 abstracts and phrase lists [23]. Barker and Cornacchia had 12 judges view 13 documents [4]. Chen used eight subjects and 10 texts [5]. In Turney's study [25] 205 users assessed keyphrases for a total of 267 documents.

We chose to maximise the number of subjects and the number of assessments of each phrase list to minimise the effect of assessor subjectivity. In this respect our study is more robust than those described above. Although Turney reports large numbers of assessors and documents, the Web-based mechanism by which this was achieved necessarily relinquished control over subject and document selection. Such control was important for our study because of the domain-specific nature of Kea's extraction algorithm. Due to resource and time constraints, the number of documents considered was smaller than we would have ideally chosen. The documents that we used are from a particular domain (computer-human interaction) and of a particular style (conference research paper). It is clearly difficult to assert that the results that we have observed for Kea can be generalised beyond such papers. However, this is actually a moot point, because Kea is a *domain-specific* system, trained on collections of documents that are similar to those from which keyphrases are to be extracted.

A second limitation of the study is the narrow profile of the subjects. To ensure accurate assessment of keyphrases, subjects must be conversant with the domain of the documents under consideration. We have attempted to ensure this in our study to improve the integrity of the assessments. Tolle and Chen also adopted this approach, using strongly matched subjects and documents in a restricted domain [23]. The assessors in Turney's Web-based study were anonymous but submitted their own document for processing, and neither Chen nor Barker and Cornacchia describe their subject populations. Kea is intended for use on restricted-domain document collections, and consequently its users will likely be conversant with that domain. This is the scenario that has been modeled by our study.

This evaluation measured the quality of individual keyphrases. We have also compared sets of keyphrases by combining the scores of individual phrases. However, in some of the uses of keyphrases that we described earlier in this paper, keyphrase *groups* are presented to users. Barker and Cornacchia [4] captured assessments of groups of keyphrases produced by both *B&C* and *Extractor*. They found that *B&C* groups were preferred more often that *Extractor* groups (47% versus 39% of preferences). This is at odds with judgements of individual keyphrases, which reflected a preference for those produced by Extractor. This suggests that evaluations of individual keyphrase quality do not generalize to keyphrase sets.

8. Conclusions

This study has shown that Kea extracts good keyphrases, as measured by human subjects. Their assessments were uniformly positive, and with the exception of very short keyphrase lists, Kea keyphrases were almost as good as those specified by authors. These results corroborate evaluations of Kea based on author keyphrases, and suggest that Kea ranks keyphrases in a sensible way. We are confident that Kea keyphrases are suitable for use in the interfaces described in this paper.

Previous studies have used author keyphrases as a gold-standard, against which other keyphrases are compared, but offered no evidence that author keyphrases are good keyphrases. Our results show that authors do provide good quality keyphrases, at least for the style of documents in our study. They also indicate that author keyphrases are listed with the best keyphrases first, which may have implications when author keyphrases are used to measure keyphrase quality.

9. References

[1] *Proceedings of CHI'97: Human Factors in Computing Systems*, ACM Press, 1997.

[2] Anick, P. and Vaithyanathan, S. Exploiting Clustering and Phrases for Context-Based Information Retrieval. In *Proceedings of SIGIR'97: the 20th International Conference on Research and Development in Information Retrieval*, (Philadelphia, 1997), ACM Press, 314-322.

[3] Arampatzis, A.T., Tsoris, T., Koster, C.H.A. and Van der Weide, T.P. Phrase-based information retrieval. *Information Processing & Management. 34*, 6 (1998); 693-707.

[4] Barker, K. and Cornacchia, N. Using Noun Phrase Heads to Extract Document Keyphrases. In *Proceedings of the Thirteenth Canadian Conference on Artificial Intelligence (LNAI 1822)*, (Montreal, Canada, 2000), 40-52.

[5] Chen, K.-H. *Automatic Identification of Subjects for Textual Documents in Digital Libraries* Los Alamos National Laboratory, Los Alamos, NM, USA, 1999.

[6] Croft, B., Turtle, H. and Lewis, D. The Use of Phrases and Structured Queries in Information Retrieval. In *Proceedings of SIGIR'91*, 1991), ACM Press, 32-45.

[7] Domingos, P. and Pazzani, M. On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. *Machine Learning 29*, 2/3 (1997); 103-130.

[8] Dumais, S.T., Platt, J., Heckerman, D. and Sahami, M. Inductive Learning Algorithms and Representations for Text Categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management*, 1998), ACM Press, 148-155.

[9] Frank, E., Paynter, G., Witten, I., Gutwin, C. and Nevill-Manning, C. Domain-specific Keyphrase Extraction. In *Proceedings of the Sixteenth International Joint Conference on Aritificial Intelligence*, 1999), Morgan-Kaufmann, 668-673.

[10] Gutwin, C., Paynter, G.W., Witten, I.H., Nevill-Manning, C. and Frank, E. Improving Browsing in Digital Libraries with Keyphrase Indexes. *Journal of Decision Support Systems 27*, 1-2 (1999); 81-104.

[11] Jones, S. Design and Evaluation of Phrasier, an Interactive System for Linking Documents Using Keyphrases. In *Proceedings of Human-Computer Interaction: INTERACT'99*, (Edinburgh, UK, 1999), IOS Press, 483-490.

[12] Jones, S. and Mahoui, M. Hierarchical Document Clustering Using Automatically Extracted Keyphrases. In *Proceedings of the Third International Asian Conference on Digital Libraries*, (Seoul, Korea, 2000), 113-120.

[13] Jones, S. and Paynter, G. Topic-based Browsing Within a Digital Library Using Keyphrases. In *Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries*, (Berkeley, CA, 1999), ACM Press, 114-121.

[14] Jones, S. and Staveley, M. Phrasier: a System for Interactive Document Retrieval Using Keyphrases. In *Proceedings of SIGIR'99: the 22nd International Conference on Research and Development in Information Retrieval*, (Berkeley, CA, 1999), ACM Press, 160-167.

[15] Kosovac, B., Vanier, D.J. and Froese, T.M. Use of Keyphrase Extraction Software for Creation of an AEC/FM Thesaurus. *Electronic Journal of Information Technology in Construction 5* (2000); 25-36.

[16] Krulwich, B. and Burkey, C. The Infofinder Agent - Learning User Interests Through Heuristic Phrase Extraction. *IEEE Intelligent Systems & Their Applications 12*, 5 (1997); 22-27.

[17] Larkey, L.S. A Patent Search and Classification System. In *Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries*, (Berkeley, CA, 1999), ACM Press, 179-187.

[18] Paynter, G.W., Witten, I.H. and Cunningham, S.J. Evaluating Extracted Phrases and Extending Thesauri. In *Proceedings of the Third International Conference on Asian Digital Libraries*, (Seoul, Korea, 2000), 131-138.

[19] Paynter, G.W., Witten, I.H., Cunningham, S.J. and Buchanan, G. Scalable Browsing for Large Collections: a Case Study. In *Proceedings of Digital Libraries'00: The Fifth ACM Conference on Digital Libraries*, (San Antonio, TX, USA, 2000), ACM Press, 215-223.

[20] Pedersen, J., Cutting, D. and Tukey, J. Snippet Search: a Single Phrase Approach to Text Access. In *Proceedings of the 1991 Joint Statistical Meetings*, 1991), American Statistical Association,

[21] Siegel, S. and Castellan, N.J. *Nonparametric Statistics for the Behavioral Sciences (2nd edition)*, McGraw Hill College Div, 1988.

[22] Smeaton, A. and Kelledy, F. User-Chosen Phrases in Interactive Query Formulation for Information Retrieval. In *Proceedings of the 20th BCS IRSG Colloquium*, (Grenoble, France, 1998),

[23] Tolle, K.M. and Chen, H. Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools. *Journal of the American Society for Information Science 51*, 4 (2000); 352-370.

[24] Turney, P.D. *Learning to Extract Keyphrases from Text*. Technical Report ERB-1057 (NRC #41622). Canadian National Research Council, Institute for Information Technology, 1999.

[25] Turney, P.D. Learning Algorithms for Keyphrase Extraction. *Information Retrieval 2*, 4 (2000); 303-336.

[26] Witten, I.H., McNab, R.J., Jones, S., Apperley, M., Bainbridge, D. and Cunningham, S.J. Managing Complexity in a Distributed Digital Library. *IEEE Computer 32*, 2 (1999); 74-9.

[27] Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C. and Nevill-Manning, C.G. KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of Digital Libraries '99: The Fourth ACM Conference on Digital Libraries*, (Berkeley, CA, 1999), ACM Press, 254-255.

[28] Zamir, O. and Etzioni, O. Grouper: A Dynamic Clustering Interface to Web Search Results. *Computer Networks and ISDN Systems 31*, 11-16 (1999); 1361-1374.

Community Design of DLESE's Collections Review Policy: A Technological Frames Analysis

Michael Khoo
Department of Anthropology CB 233
University of Colorado
Boulder, CO 80309-0233, U.S.A.
E-mail: michael.khoo@colorado.edu

ABSTRACT

In this paper, I describe the design of a collection review policy for the Digital Library for Earth System Education (DLESE). A distinctive feature of DLESE as a digital library is the 'DLESE community,' composed of voluntary members who contribute metadata and resource reviews to DLESE. As the DLESE community is open, the question of how to evaluate community contributions is a crucial part of the review policy design process. In this paper, technological frames theory is used to analyse this design process by looking at how the designers work with two differing definitions of the 'peer reviewer,' (a) peer reviewer as arbiter or editor, and (b) peer reviewer as colleague. Content analysis of DLESE documents shows that these frames can in tum be related to two definitions that DLESE offers of itself: DLESE as a library, and DLESE as a digital artifact. The implications of the presence of divergent technological frames for the design process are summarised, and some suggestions for future research are outlined.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, standards, user issues.*

General Terms

Design, Human Factors.

Keywords

Content Analysis, Decision Making, Design, Digital Library, Ethnography, Peer Review, Technological Frames.

1 INTRODUCTION

The Digital Library for Earth System Education (DLESE: www.dlese.org) is a National Science Foundation funded project whose mission is to provide searchable access to online earth science resources for everybody from elementary school children to university professors. DLESE is structured around a distributed 'participatory framework' consisting of three main 'primary functions'—Policy, Operations, and Community—located in different parts of the United States. The DLESE Program Center, part of the 'Operations' function, is headquartered in the offices of the University Corporation for Atmospheric Research, Boulder, Colorado. The distributed structure of DLESE means that, apart from scheduled meetings that require members to travel across the country, the great majority of communication between DLESE's various project functions is carried out via various communication technologies.

One such project function involves designing a review policy that sets out criteria for accessing items to DLESE's catalogue. In this paper I examine the ongoing development of this policy, through an analysis of messages posted to the DLESE Collections Working Group discussion forum. I will suggest that the discussion surrounding the review policy is situated within a wider set of negotiations concerning exactly what a 'digital library' is. In these discussions DLESE, as a digital library, appears to be defined in at least two ways: firstly, as a library that that is being digitised; and secondly, as a digital object that has library-like functions. As DLESE aims to be a heterogeneous community, the presence of diverse definitions of technology should come as no surprise, and could be seen as a desirable characteristic. However, for diversity to be used constructively, it has to be recognised as such; if left unrecognised there is a possibility that such differences have the potential to impact DLESE in unforeseen ways. Using technological frames theory (TFT) I will explore how these definitions are present in DLESE documents, mission statements, and discussions of the review policy, and what impact they might be having on policy formulation.

2 TECHNOLOGICAL FRAMES THEORY

Technological frames theory (TFT) studies the shared frames of reference underlying individual and collective perceptions of technology. I will focus on the work of Orlikowski and colleagues [35, 36, 37, 38], who use a form of technological frames theory derived from partly Goffman's theory of frames [24] and partly from Giddens' theory of structuration [22]. Orlikowksi and Gash [37] argue that

> an understanding of people's interpretations of technology is critical to understanding their interaction with it. To interact with technology, people have to make sense of it; and in this sense-making process, they develop particular assumptions, expectations, and knowledge of the technology, which then serve to shape subsequent actions toward it. While these interpretations become taken-for-granted and are rarely brought to the surface and reflected on, they nevertheless remain significant in influencing how actors in organizations think about and act toward technology. (175)

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

```
INCOMMENSURABLE    |  INCONGRUENT    |   CONGRUENT     |   INCONGRUENT   |   INCOMMENSURATE
                            FRAME A <----------------->|
                                     |<----------------->  FRAME B
                 FRAME C <---------------------->      |
                                     |    <--------------------->  FRAME D
 FRAME E <-------------------------------------------->|
                                     |<-------------------------------------------->  FRAME F
```

Figure 1

Orlikowski and Gash refer to these located 'sense-making' interpretations as 'technological frames of reference,' "built-up repertoire[s] of tacit knowledge that [are] used to impose structure upon, and impart meaning to, otherwise ambiguous social and situational information to facilitate understanding" (176). They note parallels between their concept of technological frames, and other theories of shared cognitive structures, including cognitive maps [11, 19], frames [24], interpretive frames [6], interpretive schemes [22], mental models [3, 39], paradigms [27, 40], scripts [1, 23], and thought worlds [17, 18]

There are three aspects of Orlikowski and Gash's conceptualisation of technological frames that I would like to emphasise here: their emergence, their identification with particular groups, and their congruence. First, Orlikowski and Gash stress that technological frames are not set, but *emergent*, contingent and evolving: "Frames are flexible in structure and content, having variable dimensions that shift in salience and content by context and over time. They are structured more as webs of meanings than as linear, ordered graphs" (176). These emergent properties have important implications for practice, making possible ongoing observation, analysis, and potential intervention in technology implementation processes.

Second, Orlikowski and Gash state that "technological frames are shared by members of a group having a particular interaction with some technology" (203); the implication is therefore that frames can be associated with particular stakeholder groups, communities of practice, etc.

Third, Orlikowski and Gash use the notions of 'congruence' and 'incongruence' to describe the nature and extent of differences among frames. While congruent frames are not identical, they are related in ways that imply similar expectations of a technology. Incongruent frames, on the other hand, imply "important differences in expectations, assumptions, or knowledge about some key aspects of technology ... We expect that where incongruent technological frames exist, organizations are likely to experience difficulties and conflicts around developing ... and using technologies" (180).

The upper part of Figure 1 shows two sets of incongruent frames—A and B, and C and D—that overlap to greater and lesser degrees. Orlikowski and Gash imply that these two sets of frames can be mapped onto a continuum leading from congruence to incongruence (Figure 2):

```
CONGRUENT                             INCONGRUENT
<--------●-----------------------●------------>
   FRAMES A & B             FRAMES C & D
```

Figure 2

Note that this assessment is not quantifiable: we cannot say that frames A and B are 'three times more congruent' than frames C and D.

I want to suggest instead that we see technological frames as composed of disparate knowledges and practices and therefore as being constituted in heterogeneous rather than homogeneous ways. My suggestion is outlined in the lower part of Figure 1. Consider frame E as a pedagogical frame. It might contain pedagogists who are also familiar with technological model being implemented (in the case of DLESE, they might be informed about system privacy and security, the differences between search and browse strategies and architectures, GUI design, etc.). Frame E would also contain however people who know something about search and browse strategies, but who do not understand the technical implications of the distinction; and others who think that computers will just somehow make their pedagogical experience more efficient and more fun. Similarly, frame F considered as a designer's frame might contain a proportion of people who have experience with teaching and designing K-12 class projects. It would also have members who might be involved in educational activities but who are not really knowledgeable pedagogists, as well as others who think that the system architectures of the tools they provide will be transparently clear to teachers, and will automatically make the latter's lives easier.

In addition to congruent and incongruent frames, therefore, I propose a third level of technological frame interaction termed, after Kuhn [27], 'incommensurate.' This incommensurate level of interaction is one in which the concepts of one frame can not be understand in terms of the concepts of the other frame. Although Orlikowski and Gash equate 'technological frames' with Kuhnian paradigms, and therefore draw an analogy between incongruence and incommensurability, they do not describe incongruence as precisely equivalent to incommensurability. Orlikowski and Gash's notion of incongruence allows for a certain degree of commensurability, in that it allows for epistemological and ontological continuities between diverse frames. I argue, however, that while incongruence allows for the same data to be interpreted in different ways, incommensurability stresses that the data are in themselves differently constituted. In Figure 2 there is a greater conceptual distance to cover in going from incommensurate understanding to congruent understanding, than there is in going from incongruent understanding to congruent understanding.

The difference is subtle but I believe important, especially with regard to communicative efficiency. While only a small proportion of a community might espouse a technological frame that is incommensurate with the frame of the group that is being interacted with, the impact of that small group on the overall communicative efficiency might be out of proportion to their size within the group. While there is room for both agreement and disagreement in the interactions between two technological frames, it might pay therefore to examine situations where incommensurabilities might be occurring.

2.1 TFT Methodology

TFT methodology examines frame interactions through the study of the social interactions, social behaviours, and texts in and through which such interactions are materialised. These objects of study can include: interviews with participants; ethnographic observation of group activities (workplace practices, meetings, etc); content analysis of paper and CMC documents (reports, memos, agendas, web sites, e-mail logs, etc.); and metrics (the click-trails of users of a web site, or the pattern and type of calls or e-mails to a help desk). The methodology assumes, time and research circumstances permitting, an iterative relationship between induction and deduction. Empirical data are collected, reviewed, and analysed in order to identify technological frames. At the same time, archive research is used to try and identify the broader contexts within which such frames might be set. Primary and secondary sources are then used to inform analyses, analyses that can then be reapplied to existing data, or used to inform the gathering of fresh data.

3 TECHNOLOGICAL FRAMES IN DLESE

In its documents, DLESE represents itself as many things. The two conceptions I am concerned with here are (a) DLESE as a library that is digital, and (b) DLESE as a digital object that performs certain library-like functions. Unlike Orlikowski and Gash, I do not suggest that these definitions are mappable onto distinct constituencies within DLESE, although this is a situation that may change in the future as DLESE develops. I suggest instead that they are emergent properties of DLESE as a whole, which may at some point stabilise into various communities of perception.

3.1 DLESE as Library

The first definition concerns that which DLESE appears as a digital version of—the 'traditional' library. This definition has its origins partly in ideas of what a library should be, and how it orders its contents.

It is always difficult to discern an artifact's point of emergence, but places to look for the roots of modern libraries might include: the development of alphabetisation, the decimal system, and indexing in Europe from the thirteenth century on [15]; the technologies that culminated in Gutenberg's development of the movable type printing press [20]; the development of a scientific community that recorded its results in peer reviewed publications [43]; the growth of a secular public sphere both stimulated by and embodied in public institutions [2]; and the development of systems of classification [16], including the Dewey Decimal System [32, 33].

The DLESE material I have examined presents DLESE more as a digital than a traditional library. Nevertheless, a number of features of "traditional" libraries are referred to—for instance, catalogued materials that are high quality, reliable, and peer-reviewed—and the breadth, depth and quality of DLESE's collection is particularly emphasised. According to the DLESE Community Report [31]:

> A primary goal is to collect and make available the large body of quality educational materials developed with federal funding ... One of the central functions envisaged for DLESE is the development of a comprehensive system for reviewing and evaluating the diverse materials in its collections. This review system will provide a mechanism for recognizing high-quality work analogous to the peer review process traditionally used for journal articles. (18)

As was the case with the development of the first peer-reviewed scientific journals, "the review process should provide users with a high degree of confidence that they will be able to find the high-quality instructional materials that they need, and creators of new materials will benefit from incentives that will result in national recognition for their contribution" (18-19) [c.f. 43].

DLESE materials will be 'rigidly' catalogued along 'a number of axes,' using both traditional classification categories, as well as the new sets of categories emphasising pedagogical efficacy and the earth systems thinking expected to emerge from the DLESE community's requirements. Extensive cataloguing is seen as one way in which the vast amounts of earth science data that already exist can be made more meaningful. Where further assistance is required, a "human-mediated" help desk "will provide user-assistance in finding resources" (32).

The role of the librarian as community delegated curator is emphasised: "Guaranteeing the integrity and stability of DLESE materials will provide one of its key distinctions from the current World Wide Web. Integrity includes accessibility, currency, correctness; stability includes operability, periodic upgrades, and monitoring of product usage. This is a fundamental and non-negotiable function of the library" (32).

3.2 DLESE as Digital Artifact

DLESE documents are particularly rich in discourses concerning the possibilities of technology. The following examples are taken from two main sources: the online proceedings of a DLESE seminar held in 1999, entitled "Portal to the Future" [30], and DLESE's current public mission statement, the DLESE Community Plan [31]. In these documents, DLESE describes itself in a number of ways; for instance as a 'Memex,' a 'portal,' a 'search engine,' a 'network,' and as a holder of 'digital artifacts.'

First, DLESE explicitly references Vannevar Bush's 1945 vision of the Memex as a technology that promised instant control of large amounts of data [13]. According to the DLESE Community Report, "The idea of a system that organizes information and integrates services is not new, dating back at least as far as Vannevar Bush's vision of the Memex ... Nor is it a large step from this idea to its application to science education" (13). The theme is reflected in a number of places, for instance in a vignette that describes two professors producing class projects, who turn to DLESE and rapidly find precisely the information that they are looking for (2). The ease with which they do this recalls Bush's imaginary user manipulating his Memex (2).

Second, DLESE emphasises its role as a 'portal.' The portal metaphor has been central to DLESE design for a while: a 1999 report [30] describes DLESE as a "Portal to the Future." Within the report DLESE is described as "enhancing the collection, distribution, and service functions of a traditional library" (1), by providing a portal to earth data (2), as well as being "a dynamic identity establishing new communication links between new users and experiences users of library materials, between the creators and users of materials, and between learners, educators and scientists" (2). The report contains a quote by Goodchild (a participant) that reads: "The Moonshot: By 2005, to design and implement a digital library for earth science education that serves educators and students by facilitating a new era of sharing materials, and by being a reservoir of choice for those materials" (3). By the time this quote was incorporated into DLESE's Community Plan it read: "The Challenge: By 2005, to design and implement *the Portal*, a digital library for earth science education..." (7; emphasis added).

Third, DLESE presents itself as a search engine: "The DL should provide easy and individualized access to all through a search mechanism and profile of user.... This should be done through computer search, with an option for a human help desk" [12]. In an "Example User Scenario" [25], the two professors mentioned above "[go] back to the search engine [and] ... ask for a global warming theme ... This search led to several hundred results, but luckily the search engine sorted them into three categories..."

Fourth, DLESE will deal in distributed digital artifacts: "The Library's 'holdings' are a federated collection of digital artifacts from many sources, and only a fraction are acquired and maintained in central repositories' [21]. The placing of holdings in quotation marks indicates that DLESE's holdings are not those of the traditional library, but rather provided through connection to high-speed electronic communication networks. Thus, fifth, "The DL will provide seamless links to other relevant libraries" [12]; "[N]ew information technologies provide direct linkages between people (e.g., educators, researchers, and the community at large) and information about the earth (e.g., instructional materials, data sets, images) that make new ways of learning possible" [31]; and "The work of [DLESE] advisory groups should be complemented with broad-based, electronically facilitated input directly from the community of Library users/creators" [21]. Our two imagined users described above, for instance, before performing their search, have sat down "at a nearby PC, connected through high-speed Internet-2 connections [and] quickly moved to the DL's search interface" [24].

These and other descriptions of the potential of DLESE's digital technology are widely distributed throughout DLESE publicity material; the general tone may be summed up as follows: "Taking advantage of new information technologies, DLESE can enhance the collection, distribution, and service functions of a traditional library. Serving as a community center for Earth system education at all levels, DLESE creates the possibility of a seamless, lifelong learning experience" [31].

The characteristics of these two emergent framings of DLESE—library, and digital artifact—are summarised in Table 1 (below). It should be emphasised that these framings are general and are posited heuristically. I do not want to propose (as Orlikowski and Gash do) that they are held by separate, identifiable communities, either congruently or incongruently. Rather, I suggest that their articulation in differing circumstances could both drive and hinder DLESE. To examine ways in which this might occur, I have analysed postings from a publicly viewable DLESE online discussion forum.

4 ANALYSIS

The forum of the DLESE Collections Mailing List (CML) was chosen as this was the forum with the most postings (90 messages between 02.03.2000 and 11.09.2000). Compared to other DLESE forums, there were longer threads, and clusters in time that suggested ongoing dialogues. In 90 messages were 35 single postings, 8 threads of 2 messages, 5 threads of 3, 1 thread of 5, 2 threads of 6, and 1 thread of 7 messages. 23 messages were posted between 02/03/00 and 02/23/00, and 21 messages between 11/01/00 and 11/16/00, and these clusters usually contained the longer threads. Postings have been edited as follows: the original headers have been redacted to impart a sense of narrative continuity; I have only reproduced relevant quotes from these messages and not the whole messages themselves, as indicated by the lead ellipses in each message; and I have quoted messages in the order in which they appeared as threads. Pseudonyms have been used.

Table 1

DLESE AS LIBRARY:	**DLESE AS DIGITAL ARTIFACT:**
A library that is digitised	A digital artifact that acts like a library
Underlying model: bricks and mortar library	Underlying model: the technology enabled community
Related spatial model: centralised and enclosed building	Related spatial model: decentralised and open network
Access to local holdings	Network access to distributed resources
Repository	Portal
Catalogue	Search engines
Stable documents	Fungible documents

An emergent theme in the CML concerns the issue of what constitutes peer review for the DLESE collection. On the one hand, participants can see the advantages of a centralised and closed peer review process. Placing the review process in the hands of those thought to be reliable and trustworthy representatives of the geoscience community is seen as a guarantee that the DLESE resources being reviewed are similarly reliable and trustworthy.

```
From: Jackie
Date: 02.03.2000
...[Chris's report states that] It is the respon-
sibility of the Academic Recognition Task Force to
insure that review processes and selection criteria
are academically sound [and] defensible ...
```

On the other hand, there is a recognition that DLESE community members may also have experience, for instance pedagogic experience, of using a resource. This experience might also be of value to DLESE. Further, given the fact that it is planned to place a significant part of DLESE's metadata generation in the hands of the community, it makes sense to establish a relationship of trust between the generators and holders of metadata early on. CML members therefore recognised that while peer review in the traditional sense was important, so was a form of CMC-based decentralised peer review. The concept of the 'recommendation engine,' specifically linked to the model used by Amazon.com, was thus introduced:

```
From: Jackie
Date: 02.07.2000
... This idea differs from the familiar peer-review
system in that the reviewers self-select themselves
as  users  of  the  material,  rather  than  being
selected by an editor. One advantage is that many
reviews could be accumulated, rather than just one
```

> or two. Another advantage is that the reviewers could be people who actually used the material in the classroom, rather than people who just looked the material over for the purpose of writing a review.

Two different concepts of 'peer review' are thus being discussed. While they are seen as complementary—

> From: Gerry
>
> Date: 02.09.2000
>
> ...I think that user reviews and community feedback are an essential service of the library, not instead of the peer review but in addition to the peer review.

—there also seems to be little recognition that they might be based upon two different definitions of 'peer' (the first being the specific use associated with the journal and conference peer review processes, and the second being the more general sense of 'colleague'). Instead, a continuum of peer status is developed:

> From: Chris
>
> Date: 02.10.2000
>
> ... Just thinking ... I imagine several levels of review ... a la Gerry's comments the other day. First would be the review process (yet to be defined) that leads to acceptance of the product for the library. Second would be reviewers/users. These could be given the opportunity to identify themselves to both other users and to the producer - much as seems to be common practice with journals.

Gerry contributes a JPEG image that shows a target set in a field that represents 'the entire Internet—the good, the bad, the ugly.' The target's series of concentric rings correspond to ever more refined levels of review of the resource, with the bull's-eye representing the most rigorous form of peer review. Jackie responds enthusiastically to the idea, but at the same time emphasises the different nature of 'traditional' and 'distributed' peer review processes, and retains the notion of traditional peer review as the ultimate arbiter of a resource's reputation:

> From: Jackie
>
> Date: 02.16.2000
>
> ... This discussion is all with the understanding that the community-review system would be used to judge "pedagogical effectiveness," "ease of use by faculty and students," and "motivational/ inspirational to students." I am not ready to consider dispensing with the recruited review by a scientist evaluating "scientific accuracy."

Following a DLESE meeting in July, a summary of a focus group organised by the DLESE Collections Working Group (CWG) was posted to the CML. The plan envisaged community collection review being open only to a sub-set of the DLESE community, "members who were registered with DLESE as educators." These members would have privileged access to a section of the DLESE site where a series of online submission forms would gather their feedback about resources. Having previously posited a distinction between "scientific peer review" and "community peer review," the report thus now proposed a further distinction, between community members, and 'registered' community members. A select number of these 'registered' community members were in turn to be appointed to an 'educational review board,' with the expectation that these selected members would bring the same expertise and objectivity to the DLESE peer review process that scientists normally bring to journals:

> From: Jackie
>
> Date" 07.06.00
>
> ... Under the proposed plan, only members who were registered with DLESE as educators would be able to access the evaluation area. However, within the group of people in the DLESE membership database as "educators," reviewers/testers would step forward rather than being recruited by an editor. [...] The present plan calls for a review for scientific accuracy by a scientist selected by an editor/gatekeeper (as one of the last steps, following the community review). The focus group favors having an education review board to parallel the science review board, with a review by a selected education specialist paralleling the review by the selected science specialist.

The community review subsection of the DLESE review plan was now beginning to resemble the 'target model' earlier advanced as a model for overall peer review, with an inverse correlation posited between peer review proficiency and the number of community members expected to display such proficiency.

The multiplication of components in any process inevitably complicates the way that process is mapped out. A contributor (Philip) whose message does not appear in the list, but whose content is reported to the list by a CML member, notes that "Throughout this scenario there is reference to various *evaluation levels.* This is, in my mind, another key that could simplify or complicate things depending on how it is handled." In a follow-up comment to the CML, it was noted:

> From: Jackie
>
> Date: 07.14.2000
>
> ... Hi Phil, I agree with your suggestion that it needs to be really clear to everyone (contributors, potential users, and library staff) at what review step each resource sits. It seems certain that there will be a multi-step review process, but what those steps will be and who will implement them remain undetermined (our Collections proposal suggests some steps, but it is only "pending"). We need labels for those steps, and a clearly-articulated broadly-disseminated definition of what each steps means, philosophically and operationally.

Given that DLESE is committed to an implementation time-line, such generality in the review process specifications worried one developer (caps. as in original):

> From: Gail
>
> Date: 07.14.2000
>
> ... Since I am working on the DLESE metadata framework and developing templates for it, it would be VERY helpful to come to a tentative agreement on evaluation levels VERY soon.

Soon, tentative guidelines for a two-tier acquisition process *were* posted. These partly involved renaming DLESE's 'unreviewed' collection as an 'open' collection, and establishing some basic criteria of 'fitness for geoscience pedagogy' for the latter. It is not clear however to what extent establishing this distinction significantly operationalised DLESE's peer review process, given that (a) DLESE espouses a more rigorous 'scientific peer review' model, (b) that DLESE lacks resources to widely implement this model, and that therefore (c) at any point there will always be 'reviewed' and 'unreviewed' materials in DLESE's holdings anyway. One developer subsequently posted the comment:

> From: Henry
>
> Date: 09.30.2000
>
> ... Aside from the Reviewed and Open Collections, then, is there a plan to allow shall we say the Great Unwashed into the library? i.e., resources that undergo even less scrutiny -- some would suggest none at all. I'm not necessarily advocating such a policy; however, I'd be interested in your views on this, so I can respond intelligently when this question comes up (which it does fairly often).

Henry's point, although humorously framed, was pertinent. By November 1, following the presentation of a draft review policy at another CWG meeting (a presentation that included a sophisticated flow chart of the route that any resource would have to take in order to be accepted into DLESE), the 'reviewed' and 'open' collections became the 'broad' collection. It was proposed that access to the broad collection be guarded by three filtering categories, one of which is subdivided into a further three sub-categories. The question still remained, though, operationalisation:

```
From: Jackie
Date: 11.01.2000
... The question for discussion today is how shall
the filters at the gateway to the Broad Collection
be applied?  This was a topic of heated discussion
at the metadata workshop last week, and was left
unresolved.
```

The 'broad collection' concept has subsequently received some debate that focused on the practicalities of the filtering criteria. Contrary to some of the originally stated aims of DLESE, regarding the desirability and necessity of involving the DLESE community in reviewing DLESE holdings, the conversation since November has tended to focus on the desirability having what one poster called 'DLESE aligned' reviewers at the gate of the broad collection, at least initially. At the time of writing (December 2000), no new policy has been advanced, although one of the last messages on the subject is, appropriately, a plea to begin implementation; once in the broad collection, it is argued, the best resources can be exhaustively catalogued for the core collection, while inappropriate resources can be reviewed and on occasion removed.

5 DISCUSSION

What does this discussion of peer review tell us about the digital library design process? I suggested above that DLESE's construction of itself as a digital library is based on at least two premises: DLESE as a library, and DLESE as a digital object. The two concepts often appear undifferentiated within DLESE literature; this is not surprising, as it is DLESE's stated mission to integrate the two. However, if the two concepts appear within the same discourse without reflexive differentiation, there is potential for some confusion.

I wish therefore to draw attention to two emerging synecdoches in the CML. A synecdoche is a figure of speech that conflates a subset with the overall set, or vice versa (for instance reporting that "The United States won ten medals in today's Olympics," rather than "United States athletes won ten medals...").

First, it is interesting to note that topic under current discussion, of how to review possible accessions to the broad collection, is only a sub-set of the overall peer review scheme. If we consider the concentric target model that was present in early exchanges, this part of the peer review scheme would be represented by just the outer ring of the target; yet somehow it seems to be starting to stand in for the target as a whole.

Second, the discussion concerning the development of a process for peer reviewing accessions to the DLESE collection, is in some senses becoming displaced by a discussion concerning the development of a process for peer reviewing the peers who would then carry out the peer reviewing process. This displacement is exemplified in the exchanges regarding who might be fit to be a gatekeeper to the DLESE collection—community members, 'registered' community members, etc.

In the collection design process, these two synecdoches—that of the broad collection, and that of the peer reviewer selection process—seem to have served as rhetorical devices around which the discussion in the CML has focused. Why might this displacement be occurring? A technological frames theory analysis suggests that in discussing peer review within the context of an institution considered both as a library and as a digital artifact, the CML is in effect working with two alternate frames for peer review. I have tabulated some of these distinctions in Table 2:

Table 2

DLESE AS LIBRARY	DLESE AS DIGITAL ARTIFACT
Library Models:	
A library that is digitised	A digital artifact that acts like a library
Peer Review Models:	
Peer reviewers as 'registered' members	Peer reviewers as community members
Peers as gatekeepers	Peers as colleagues
Peer review as contribution	Peer review as potential hindrance
Peer review as objective fact	Peer review as subjective opinion (recommendation engine)
Reviewed collection	Unreviewed/open/broad collection

In the presence of these two alternate frames, the synecdoches outlined above appear to function as 'boundary objects,' conceptual objects that lie on the boundaries of various communities of perception, that serve as orienting points for discussion between those communities [11]. Although boundary objects are often considered in the context of objects that enable and guide the translation of concepts between different groups, it is possible that they can also emerge unguided. I suggest that the synecdoches identified above constitute unguided boundary objects, that serve as informal conceptual tools by which CML members attempt to reconcile two alternate enframings of the concept of 'peer review.' In such situations there is potential for discussants to assume that an issue has been discussed when in fact it hasn't, or rather, it has been reduced to a synecdoche *that represents a solution to both sides*. This is a possible explanation for the fact that in devising a review policy for the broad collection (a category of DLESE holdings that did not exist at the beginning of the CML discussion) some CML members seem to be thinking that they are devising a review policy for the DLESE collection as a whole.

6 SUMMARY

The presence of diverse technological frames within a community is not in itself counterproductive. The existence of diverse technological frames should however be acknowledged and accounted for in the design process. Identification of mediator roles and of 'guided' boundary objects, for instance, relies in part on an understanding of the technological frames present in a given situation. In attempting such an accounting, this study has borrowed from two different strands of research. The first concerns the emergence of technological frames in the context of technology implementation. Studies have included Orlikowski's work on the introduction of new groupware, Lotus Notes, across a large organisation [36], and Tracy's account of conflicting frames of reference between callmakers and calltakers at 911 dispatch center [42]. The second concerns the implications of digitising 'traditional' libraries [7, 11, 14, 26, 28, 29, 32, 33, 34, 41]. These researchers and others have problematised the assumption that building a digital library is just a case of translating the components of traditional libraries into their digital equivalents. The differences between catalogues and search engines, solo research and CMC mediated collaborative projects, paper and digital media, and so on, are not just issues of design, but also of frame translation. Digital library construction is not just a case of converting traditional library structures and functions into digital form; neither is it just a question of building digital tools to perform library-like functions.

Operationalising these two perspectives in the context of the DLESE community-led design process has been achieved through the use of (a) ethnographic observation and (b) fine-grained interaction and content analysis. The longitudinal and iterative methodology used makes not only for nuanced research, but allows research results to be folded back into the project. Research can make important recommendations in a number of areas, including training opportunities for DLESE staff and community members, documentation and FAQs, and so on.

In terms of design principles, it makes economic and social sense to build good practice into collaborative architecture at the beginning, rather than trying to achieve it after the fact. More recent research is looking therefore at the nature of 'guided' boundary objects, and how these may be supported in the future. Research on the user community of the Water in the Earth System (WES) collection is looking at how experienced WES members explain to less experienced community how WES works. Aware both of where WES community members are coming from, and where WES itself is going, experienced members can mediate and translate between the two technological frames outlined above. In effect they act as 'guided' boundary objects. New research is therefore focusing on (a) the emergent behaviours of experienced WES community members who take on the role of explaining WES as an institution to novice WES community members, and (b) the ways these roles interact with the community use of tools such as brain-storming sessions and concept maps.

7 CONCLUSION

It is often said that changing people will be the hardest part of the information technology revolution. When we ask what a digital library is (and isn't), we are in many senses also asking who belongs to the digital library community now, who will join in the future—and why they will join. In engaging with these questions, this study has examined the emergent perceptions and roles of the members of a large diverse and complex digital library community. Although the research is based upon observation of one case, ongoing research will implement a parallel longitudinal methodological framework to continue the examination of the parts that technological frames, guided and unguided boundary objects, and mediators, play in the community-led digital library design process.

8 ACKNOWLEDGMENTS

I would like to thank the members of the DLESE Project Center, in particular Tammy Sumner (PI) and Mary Marlino (Director), for encouraging and assisting me in this project.

9 REFERENCES

[1] Abelson, R.P. Psychological Status of the Script Concept. American Psychologist 36(1981:7), 715-729.

[2] Anderson, B. Imagined Communities. Verso, London, 1991.

[3] Argyris, C., and Schon D. Organizational Learning. Prentice-Hall, Englewood Cliffs NJ, 1978.

[4] Barley, S. Technology as an Occasion for Structuring. Administrative Science Quarterly 31(1986), 78-108.

[5] Barley, S. Images of Imaging: Notes on Doing Longitudinal Fieldwork. Organization Science 1(1990:3), 220-247.

[6] Bartunek, J., and Moch, M. First Order, Second Order, and Third Order Change and Organization Development Interventions: A Cognitive Approach. Journal of Applied Behavioral Science 23(1987:4), 483-500.

[7] Bishop, A. (ed.). How We Do User-Centered Design and Evaluation of Digital Libraries: A Methodological Forum. 37th Allerton Institute, 1995. Graduate School of Library and Information Service, University of Illinois at Urbana-Champaign. http://edfu/lis.uiuc.edu/allerton

[8] Bloomberg, J.L. The Variable Impact of Computer Technologies on the Organization of Work Activities. Proceedings of the Conference on CSCW 1986. ACM Press, 35-42.

[9] Boorstin, D. The Discoverers. Random House, New York, 1983.

[10] Bougon, M., Weick, K., and Binkhorst, D. Cognition in Organizations: An Analysis of the Utrecht Jazz Orchestra. Administrative Science Quarterly 22(1977:4), 606-639.

[11] Bowker, G., and Star S. (eds.) How Classifications Work: Problems and Challenges in an Electronic Age. Library Trends Special Issue 47(2).

[12] Burrows, Howard, Shelley Canright, Tim Foresman, Don Johnson, Heather Macdonald, Sudha Ram, and Mike Smith. Panel 1: What are the Appropriate Scope and Focus for an Earth System Education Digital Library? Portal to the Future: A Digital Library for Earth System Education. Coolfont Resort, Berkeley Springs, West Virginia, August 8-11, 1999. Ms. http://www.dlese.org/panel/reports/panel_1.html

[13] Bush, V. As We May Think. The Atlantic Monthly 176(1945):101-108.

[14] Crabtree, Andy, Michael B. Twidale, Jon O'Brien, and David M. Nichols. Talking in the Library: Implications for

the Design of Digital Libraries. ACM Digital Libraries 1997:221-227.

[15] Crosby, Alfred W. The Measure of Reality. Quantification and Western Society, 1250–1600. Cambridge University Press, Cambridge, 1997.

[16] Dolby, R.G.A. Classification of the Sciences: The Nineteenth Century Tradition. In: Roy F. Ellen and David Reason (Eds.), *Classifications in Their Social Context.* Academic Press, London, 1979.

[17] Dougherty, D. Interpretive Barriers to Successful Product Innovation in Firms. Organizational Science 3(1992:2), 179-202.

[18] Douglas, M. How Institutions Think. Routledge and Kegan Paul, London, 1987.

[19] Eden, C. On the Nature of Cognitive Maps. Journal of Management Studies. 29(1992:3), 261-265.

[20] Eisenstein, E. The Printing Press as an Agent of Change. Cambridge University Press, Cambridge, 1979.

[21] Fulker, D. Panel 2: How Should the Library be Governed, Managed, and Funded? Portal to the Future: A Digital Library for Earth System Education. Coolfont Resort, Berkeley Springs, West Virginia, August 8-11, 1999. Ms. http://www.dlese.org/panel/reports/panel_2.html

[22] Giddens, A. The Constitution of Society: Outline of the Theoryructure. Berkeley CA, University of California Press, 1984.

[23] Gioia, D.A. Symbols, Scripts, and Sensemaking: Creating Meaning in the Organizational Experience. In: *The Thinking Organization.* Jossey-Bass, San Francisco CA, 1986.

[24] Goffman, E. Frame Analysis. Harper and Row, New York, 1974.

[25] Kastens, K., Ciccione, K., Duff, B., Goodchild, M., Gordin, D., and Ruzek, M. Panel 3: Collections. Portal to the Future: A Digital Library for Earth System Education. Coolfont Resort, Berkeley Springs, West Virginia, August 8-11, 1999. Ms. http://www.dlese.org/panel/reports/panel_3.html

[26] Kilker, J., and Gay, G. The Social Construction of a Digital Library: A Case Study Examining Implications for Evaluation. Information Technology and Libraries 18(1998:2), 60-70.

[27] Kuhn, T. The Structure of Scientific Revolutions. Chicago, University of Chicago Press, 1970.

[28] Levy, D. I Read the News Today, Oh Boy: Reading and Attention in Digital Libraries. Proceedings of ACM Digital Libraries 1997:202-210

[29] Levy, D., and Marshal, C.M. Going Digital: A Look at Assumptions Underlying Digital Libraries. Communications of the ACM 38(1995:4), 77-84.

[30] Manduca, C.A., and Mogk, D.W. Portal to the Future: A Digital Library for Earth System Education. Coolfont Resort, Berkeley Springs, West Virginia, August 8-11, 1999. Preliminary Report. Ms. http://gdl.ou.edu/report/reportnew.html

[31] Manduca, C.A., and Mogk, D.W. DLESE: A Community Plan. Boulder: DLESE Program Center, Boulder CO, 2000.

[32] Miksa, F. The Development of Classification at the Library of Congress. University of Illinois Graduate School of Library and Information Science Occasional Papers, #164, 1984.

[33] Miksa, F. The DDC, the Universe of Knowledge, and the Post-Modern Library. Forest Press/OCLC Online Computer Library Center, Albany NY, 1998.

[34] Neves, F., and Fox, E. A Study of User Behavior in an Immersive Virtual Environment for Digital Libraries. Proceedings of ACM Digital Libraries 2000:103-111.

[35] Orlikowski, W. The Duality of Technology: Rethinking the Concept of Technology in Organizations. Organization Science 3(3), 398-427. (1992a)

[36] Orlikowski, W. Learning From Notes: Organizational Issues in Groupware Implementation. MIT Sloan School Working Paper #3428-92. (1992b) http://ccs.mit.edu/papers/CCSWP134.html

[37] Orlikowski, W., and Gash, C. Technological Frames: Making Sense of Information Technology in Organizations. ACM Trans. Inf. Systems, 12 (April 1994), 174-207.

[38] Orlikowski, W., Yates, J., Okamura, K., and Fujimoto, M. Shaping Electronic Communication: The Metastructuring of Technology in Use. MIT Sloan School Working Paper #3611-93, 1996. http://ccs.mit.edu/papers/CCSWP167.html

[39] Schutz, A. On Phenomenology and Social Relations. Chicago, University of Chicago Press, 1970.

[40] Sheldon, A. Organization Paradigms: A Theory of Organizational Change. Organizational Dynamics 8(3, Winter 1980), 61-80.

[41] Theng, Y., Mohd-Nasir, N., and Thimbedy, H. Purpose and Usability of Digital Libraries. ACM Digital Libraries 2000:238-239.

[42] Tracy, K. Interactional Trouble in Emergency Service Requests: A Problem of Frames. Research on Language and Social Interaction 30(1997:4), 315-343.

[43] Zuckerman, H., and Merton, R. Patterns of Evaluation in Science. Institutionalisation, Structure and Functions of the Referee System. Minerva IX(1971:1), 66-100.

Legal Deposit of Digital Publications: A Review of Research and Development Activity

Adrienne Muir
Department of Information Science
Loughborough University
Loughborough LE11 3TU, UK
+44 (0)1509 223065
A.Muir@lboro.ac.uk

ABSTRACT

There is a global trend towards extending legal deposit to include digital publications in order to maintain comprehensive national archives. However, including digital publications in legal deposit regulation is not enough to ensure the long-term preservation of these publications. Concepts, principles and practices accepted and understood in the print environment, may have new meanings or no longer be appropriate in a networked environment. Mechanisms for identifying, selecting and depositing digital material either do not exist, or are inappropriate, for some kinds of digital publication. Work on developing digital preservation strategies is at an early stage. National and other deposit libraries are at the forefront of research and develop in this area, often working in partnership with other libraries, publishers and technology vendors. Most work is of a technical nature. There is some work on developing policies and strategies for managing digital resources. However, not all management issues or users needs are being addressed.

Categories and Subject Descriptors

K.5.0 [**Legal aspects of computing**]

General Terms

Management, Legal Aspects

Keywords

Legal deposit Digital publications Digital preservation

1. INTRODUCTION

The concept and practice of legal deposit is under threat in the digital environment. The main, though not the original, aim of legal deposit is to ensure the preservation of a nation's intellectual and cultural heritage over time. Many countries are extending legal deposit regulations to cover digital publications in order to maintain comprehensive national archives. However, even countries that have been dealing with the legal deposit of digital publications for some time are still grappling with how to collect and manage this material effectively in the long term. Existing collection management principles and practice were not designed with digital information in mind. Online and networked publications pose particularly complex challenges.

Publishers typically have a legal obligation to deliver one or more copies of their publications to deposit libraries. The depositor is usually responsible for the cost of deposit. Legal depositories include national libraries, parliamentary libraries, university libraries and national archives (for non-print material). There is great variety in the types of material collected through legal deposit. The requirement is usually material available to the public whether for sale, hire or for free. Printed publications are really the only common factor. Other types of material collected include sound recordings, audiovisual material and software. For a recent summary of the current status of legal deposit around the world see [1].

There is currently a great deal of research and development activity in this area. Early work focused on identifying issues and problems and on gathering information for making the case for extension of legal deposit. Currently deposit libraries are carrying out work, often in collaboration with other deposit libraries, publishers and technology vendors. Much of this work is of a technical nature and focuses on building the basic infrastructure, setting up digital depositories and collecting digital publications. Researchers are also working on metadata and digital preservation issues.

This review of research and development work will focus on activities specifically related to digital legal deposit. However, it will also touch on more generic work that is especially relevant. The review starts with a discussion of the issues identified through research to provide some background and context for the rest of the review. Research activities are grouped into categories for discussion. These categories are: building the infrastructure; pilot projects and digital preservation. The issues that are not currently being addressed are identified and conclusions are drawn.

2. THE ISSUES

The deposit of digital publications raises legal, economic, technical and managerial/organisational issues at all stages of the legal deposit process. For the purposes of this review, these stages are summarised as: identification of publications; selection; acquisition; accession and processing, including storing; preservation; and access. There are a number of fundamental factors that facilitate the legal deposit of digital publications: definitions; metadata; and standards.

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

There are also political issues associated with legal deposit. These arise because there are a number of different actors involved in the legal deposit system, and each of these actors has their own interests. The interests of one group do not necessarily coincide with another. For example, the commercial interests of publishers and the legal requirement to give up several copies of their product combine to cause tension between publishers, deposit libraries and legislators.

An important point that arises from the literature is that decisions taken at one stage affect decisions taken at other stages. Selection policies may need to take into account the ability of the depository to capture and preserve particular publications. Technical, legal, economic and organisational issues may influence preservation choices. Alternatively, different preservation strategies have different economic and management implications.

For the purposes of this review, the issues identified through research are discussed within the framework of the legal deposit process described above. Metadata and standards are also discussed in this way. The issue of definitions also pervades the whole process. Many well-established concepts either do not apply in the digital environment or need redefinition. This issue is so fundamental to legal deposit in the digital age that it is discussed separately.

3. DEFINITIONS

A common theme in the literature is a lack of agreed definitions for various concepts in the digital environment. Terms that are well understood in the print environment are irrelevant or have new meanings in the digital world. Examples include terms relating to documents or publishing such as "publication," "place of publication," "publication date," "publisher," "edition" and "authenticity." Another problem is that the same words have different meanings for different communities. For example, "archives" and "metadata" have different meanings for different professional communities [2]. Researchers working in this area have developed glossaries [3, 4]. Unfortunately, these are of limited use because they have been developed specifically for the use of project participants.

3.1 Documents, publications and publishing

Many of the problems of definition associated with legal deposit stem from the fact that the concept was originally based on mainly textual information first made available in individual nations via a physical carrier, usually a book. Some of the traditional concepts still apply to digital information on physical media. However, telecommunications and global networking have radically changed the nature of information dissemination. Many online publications are frequently, if not continuously, updated and they are globally available. New types of communication have emerged, such as email, mailing lists, chat rooms, personal World Wide Web home pages and dynamic Web pages that are generated 'on the fly' from databases. How much of this information can be called a "publication" in the traditional sense is open to conjecture.

The British Library commissioned a study on the definition of terms. Existing sources of definitions are given in appendices to the study report. The study found that, at the time, existing definitions were not helpful because they did not deal well with new types of material, including digital material [5]. One point made was that definitions should be format or medium-independent to make them "future proof" [5].

As far as the concepts of documents and publishing were concerned, the report did not suggest definitions, but provided "an overall framework of analysis within which to work" [5]. Martin defines a document as "(a) a combination of a work or compilation of works the medium on which the work or compilation is stored and any access technology which is specific to the document or (b) any one of a number of copies of such a combination."

This is a more complex definition than that of Mackenzie Owen and van der Walle. They define electronic publications as "published documents which are produced, distributed stored and used in electronic form" [6].

Martin also defines "published within the United Kingdom." This is "the public to which it is offered or broadcast or made available or before which it is performed includes a part of the United Kingdom ... and the publisher or an importer or distributor or an agent of any of the aforementioned is domiciled in the United Kingdom" [5]. Martin admits that this definition creates a potential loophole for publishers whose entire operation is outside of the UK, but whose offering is directly at a UK audience. He also points out it would be difficult to enforce UK law in this situation. This problem would apply to any country.

Another British Library sponsored study [7] spells out the potential problems associated with depositing online material. The report provides definitions for different types of database. However, a major point made is that the traditional concept of publishing is not applicable in the digital environment. The "publication" process is not the same in the print and online environments and different entities are involved. In the online world, no single entity has overall control of the process and intellectual property rights are created at several points. The entity that owns the rights to the data may be different from the entity hosting it. The entity owning the rights to the retrieval software may be different from the data owner and/or the host entity. Who is responsible for deposit?

3.2 Preservation

There is also confusion in the terminology used for the preservation of digital information. For example, there is a difference between digital preservation and preservation digitisation [2]. Digital preservation is "the storage, maintenance, and accessibility of a digital object over time." Preservation digitisation involves digitising a fragile object to preserve its intellectual content. Preservation digitisation, in contrast, produces a surrogate for the original object. This surrogate will then need to be preserved over time.

4. IDENTIFICATION, SELECTION, ACQUISITION

4.1 Identification

The 1996 ELDEP study reported that the amount of material published only in digital form was quite small compared the volume of traditional publishing output [8]. This view was repeated in 1999 in another study [9]. However, both of these studies stated that the proportion of published output released only in digital form was likely to increase over time.

In order to acquire information, depositories have to identify it. One suggestion here is that legal deposit could require all publishers to register their publications [10]. The existence of publications would then be known, even they were not all collected. It may be impossible to enforce in practice because of the large numbers involved, the ignorance of many Internet "publishers" of the traditional systems, or simply their unwillingness to comply.

Bibliographic control is well developed in the print environment, less well developed for non-print formats and, it seems, virtually non-existent for digital publications. One particular area of concern is that of unique identification of digital publications. While identification of offline digital material such as CD-ROMs may be reasonably straightforward using existing identifiers, online material presents problems [9]. There may be different manifestations of the same content. The question is whether each manifestation should have a different identifier, or whether there be one identifier for the underlying work. This raises the further question of how to identify each manifestation and relate this to the underlying work.

New types of identifier being developed include the Digital Object Identifier (DOI) and Uniform Resource Names. The DOI is being developed by the International Foundation to help in the management and exploitation of digital information [11]. Uniform Resource Names are persistent identifiers for online information [12].

4.2 Selection

Mackenzie Owen and de Walle point out that legal deposit laws are often selective in their coverage [6]. Some types of material are included and some are not. They recommend that, with some exceptions, all digital publications should be collected, including those published in parallel with print equivalents. An important point made is that deposit libraries will have to accept that they may never be able to collect all digital publications. There will be too many publications, too many publishers, and the rate of technological change is too fast.

There are some attempts at comprehensive collection of material. In Sweden, the National Library is attempting to capture the Swedish portion of the World Wide Web [13]. The aim of the Internet Archive is to archive the entire Internet [14]. The comprehensive approach may be feasible for countries with a relatively small digital publishing output, but it may turn out to be impossible for countries with bigger outputs.

Different depositories have different collecting policies, but selection often involves quality judgements: the importance of a particular publication, or its future research value. Technical issues can potentially distort digital selection policies [10]. It may well be that for crucially important material that is, for technical reasons, difficult but not impossible to acquire, access and preserve, expense is not a factor for consideration. There is the problem of moderately important "difficult" publications or crucially important publications that are impossible to deal with. These documents may end up being lost.

The acquisition of dynamic digital information is particularly problematic. Many writers comment on the impracticality or even impossibility of capturing every version of databases that are amended very frequently or in real-time [15]. The acquisition method here would be samples or snapshots. However, there is no commonly accepted practice and little practical experience of sampling techniques.

Hyperlinked documents present problems of deciding where the boundaries of the documents are. For example, there are questions about which is the appropriate level for archiving - a Web page, or an entire Web site just one big document? There is a question as to which linked sources should also be selected. Should only internal links within a document be maintained or should links to other documents be archived along with the original document?

With traditional publications, deposit usually means that some responsible entity sends physical objects to depositories. The situation is more complex in the digital environment. Online publications are not available in physical form so they cannot just be sent through the post. At present, there are three main options for acquiring online information. Publishers can transfer the information onto a physical medium and send that to the depositories. Publishers can arrange to transfer, or "push," information to depositories via networks. Alternatively, libraries can "pull" from publishers' sites themselves. A variant of this activity is "harvesting." This is usually done for Internet information, where the depositories use software to identify and pull in information from sites.

Several depositories are working with harvesting software to acquire publications, including the National Libraries of Australia [16], Finland [17] and Sweden [18]. Harvesting information is problematic in that existing tools do not entirely meet the needs of the depositories. Legal issues also arise, in that the depositories often need to negotiate permission from the publishers to copy their material. The push option may not work well on the Internet because it is populated with a huge number of publishers, some of whom are small organisations or even individuals. It would be impossible to set up relationships with all of them.

5. ACCESSION AND PROCESSING

Mackenzie Owen and Walle [8] recommend that quality checks and functional tests should be carried out for all items received. The purpose of such procedures is to check that the item is:

- The correct version
- In the required medium and format
- Complete
- Undamaged
- Error free and fully functional
- Not copy protected

Ensuring the authenticity of digital documents that are fluid by nature and capable of being changed very easily becomes a headache in the digital environment [15]. Some techniques for checking authenticity include time stamping and digital signatures.

The amount and types of information gathered at the accession stage will affect preservation of and long-term access to digital material. Deposit libraries will need more and better information, or metadata, from publishers from publishers at the point of accession than is necessary for printed information. There is also a need for some standardisation in this metadata.

The European Commission ELDEP study particularly focused on the bibliographic control of digital legal deposit collections [8]. There are questions as to whether current cataloguing rules can deal adequately with offline publications. The view emerging from the literature is that current rules may not be able to deal with online material at all.

6. PRESERVATION

Items in legal deposit collections are usually kept forever, therefore preservation is a central issue. If legal deposit collections are to include digital publications, solutions have to be found to the problems of digital preservation.

6.1 Media stability

Early preoccupations in this area were with the longevity of digital media. Estimates of the likely life expectancy of various storage media vary from around 1 to 100 years. Rothenberg gave some low estimates, including as little as two years for magnetic tape in some circumstances [19]. Unfortunately he gave no explanation for the low estimate and did not source his figures. The US National Media Laboratory contested Rothenberg's estimates [20]; it cites a 10-30 life expectancy for magnetic tape. Even so, this projection does not compare well with established archival media such as permanent paper or preservation microfilm. For these carriers, life expectancy is hundreds of years with optimal conditions.

As well as having inherent instabilities, the physical carriers used for digital information also react to environmental factors. These factors include both extremes of, and fluctuations in, temperature and relative humidity. Physical media also suffer from wear and tear and incorrect handling. Van Bogart produced a report on the storage and handling of magnetic tape, which is widely cited in the literature [21].

6.2 Technological obsolescence

Media instability is not the main problem as far as the preservation of digital information is concerned. The main problem is that viewing and using digital information requires the aid of equipment. The biggest threat to long-term survival is that of technological obsolescence of the hardware and software used to create and use digital information. Technical obsolescence is not a new problem for the preservation of information. Earlier examples of this include the Sony Betamax video recording format and Readex Microcards.

Recognition of technological obsolescence as the main threat to the long-term survival of digital information becomes prominent in the library and information science literature from the mid-1990s. Lehman sets out some of the aspects of technological change [22], including changes in coding and formats, software, operating systems and hardware. These changes can render digital material unreadable.

6.3 Preservation strategies

There are a number of possible strategies for digital preservation. A key question in deciding what strategy to use is what is to be preserved. Saving artefacts will not necessarily mean that the information itself is also preserved. Merely refreshing media will not overcome technological obsolescence. There are also problems associated with deciding exactly what the information or intellectual content is. This is especially problematic for multimedia or highly interactive information. Text, sound and pictures may be integrated; the software associated with the information may allow interaction between the user and the information. What has to be determined is whether it is the look and feel and functionality of the information product that is to be preserved, or just the raw information. [2].

There are a number of potential preservation strategies that address different preservation requirements and timeframes. The main preservation strategies are technology preservation, migration and emulation.

6.3.1 Technology preservation

Technology preservation is really a short-term strategy. This involves preserving the information in its original form and also the original software and hardware used to create and run the information. The strategy is likely to also involve media refreshment, especially for information stored on media with very short lifetimes [23]. However, hardware can only be maintained in working order for a finite period.

6.3.2 Migration

The Task Force on Archiving of Digital Information favoured the migration approach. The Task Force report defines migration as "the periodic transfer of digital material from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation." [24]. There are several migration strategies [25].

Adopting migration strategies means making long-term commitments to unknown future activities and unpredictable costs. Webb makes the point that most successful emulation work has been carried out with large amounts of homogenous data [10]. This is certainly not the situation for deposit libraries. While there have been some small-scale experiments with migrating publications from floppy disks [26], it is not known how well complex material stored on optical discs will migrate, if at all.

6.3.3 Emulation

The aim of the emulation concept is to allow long-term preservation of digital material while retaining the functionality and look and feel of the material. The main proponent of emulation as a preservation strategy for digital information is Jeff Rothenberg [27].

The idea of emulation is to view a digital document by using the software that created it. This does not necessarily mean that the software has to be run. The behaviour of the software could be described and the description saved so that its behaviour can be re-created in the future. The requirements for this approach would be to save the digital documents, the programs that were used to create the documents and all software required to run the documents. Software is dependent on the hardware it is created for, so the behaviour of an obsolete hardware platform would have to be emulated too. This would need the development of emulators, or software programs to mimic this behaviour [19].

Hardware emulation is potentially a simpler proposition than software emulation. The reasons for this is that there are fewer hardware platforms than operating systems and application software, so fewer emulators would have to be specified. Secondly, writing specifications for hardware is a better-developed practice than for software, so it would be easier to do [28].

In a paper for the Council on Library and Information Resources, Rothenberg set out the requirements for implementing emulation of hardware [28]. These include: techniques for specifying emulators; techniques for saving the necessary metadata (for finding, accessing and recreating documents) in human-readable form; and techniques for encapsulating documents, attendant metadata, software, and emulator specifications in a coherent and incorruptible way.

6.4 Disasters and rescue of digital information

Ross and Gow investigated approaches taken to access digital information when media are damaged or software and hardware are unavailable or unknown [29].

Things that can go wrong with digital information are:

- *Media degradation* – unfavourable environmental conditions in storage, disaster and manufacturer defects
- *Loss of functionality of access devices* – technological obsolescence, wear and tear of mechanical parts, lack of support for device drivers in newer software
- *Loss of manipulation capabilities* – due to changes in hardware and operating systems
- *Loss of presentation capabilities* - change in video display technologies, particular application packages may not run in newer environments
- *Weak links in the creation, storage and documentation chain* – data recovered but unreadable because of encoding strategy cannot be identified, loss of encryption documentation, use of unusual compression algorithms

Possible techniques for data recovery include heat and chemical treatments for soiled or damaged media, searching binary structures to identify recurring patterns and reverse engineering of content. One of the findings of the study is that there is a distinction between data recovery and data intelligibility. While it may be possible to recover data through searching binary structures, technological developments mean it will become harder to read the recovered data.

A future possibility is the use of magnetic force microscopy to read damaged media. Another is cryptography to help in the interpretation of recovered data. Ross and Gow also suggest an alternative to migration and emulation strategies for preservation. Retargetable binary translation involves "translating a binary executable programme from one machine … running a particular operating system … and using a particular file format … to another platform … running a different operation [sic] system … and using a different file format" [29].

6.5 Authenticity

Migration strategies potentially pose authenticity problems because they cause changes in the publications being migrated. Authenticity means that an object is the same as that expected based on a prior reference or that it is what it purports to be.

There is a range of strategies for asserting the authenticity of digital resources. The strategy used depends on the purpose for which authenticity is needed. These include unique document identifiers, the use of metadata to document changes, hashing, digital stamping, encapsulation techniques, digital watermarks and digital signatures.

During 1998, the CERBERUS project investigated the authenticity and integrity of electronic documents in digital libraries with a deposit task. The project partners were the Dutch Koninklijke Bibliotheek, the Technical Universities of Eindhoven and Delft and the University of Amsterdam. The project was co-funded by Innovation of Scientific Information Supply. There is a brief description on the KB Web site [30], but there is little written in English on the findings of this project.

6.6 Management of digital preservation: life cycles, stakeholders and rights issues

The concept of the life cycle of digital resources has been put forward as a tool for looking at the challenges of digital preservation. This was developed in the context of a study to develop a strategic policy framework for the creation, management and preservation of digital resources. Hendley added to the original model developed by the UK Arts and Humanities Data Service. The life cycle breaks down into several stages. These are: resource creation; selection and evaluation; management; disclosure; use; preservation; and rights management.

Data archives can often dictate requirements at the creation extent to a great extent. This is not the case for legal depositories. Deposit material and depositors will be diverse. In many cases, the main priority of depositors is commercial gain and this can conflict with preservation interests. The creators of the framework acknowledge this.

Rights issues are important because preservation strategies involve copying, and possibly changing, the original information in some way. The life cycle framework illustrates how the stages are interrelated and how decisions taken at one stage impact on other stages. The aim is to help in the policy and decision making process and help identify where collaboration efforts would help preservation. The framework is supported by a number of case studies. One case study is legal deposit libraries.

Haynes et al. examine the attitudes of "originators and rights holders" towards to the issue of their responsibility for digital preservation [31]. The study report lists various stakeholder groups. These are: libraries; publishers; archive centres; distributors; IT suppliers; legal depositories; consortia; authors; and networked information service providers.

The consultants made a number of recommendations in their report. One was that a body should be established to co-ordinate digital archiving activities - a National Office of Digital Archiving. This suggestion is similar to that of a national digital preservation officer from Matthews, Poulter and Blagg [32]. This idea has since been taken forward in the UK. The Joint Information Systems Committee of the Higher and Further Education Funding Councils has appointed an officer to develop digital preservation strategies and work with other bodies to establish a so-called Digital Preservation Coalition [33].

Another suggestion was for a distributed approach to digital preservation. Distribution could have a number of bases, including regional, format and ownership. The suggested National Office of Digital Archiving would coordinate the development of standards and guidelines in cooperation with other agencies. The consultants suggest that legal deposit should be used as a mechanism for acquiring material, but that publishers should only have to contribute one copy of each publication.

Users should not be charged for access, but costs should be shared between research funders, the public (through government funding), and research communities. There is no mention of any responsibility falling on publishers for maintaining archives.

More recently, researchers from AHDS have carried out a study on the preservation management of digital materials and have produced a draft workbook for preservation managers. The workbook brings together research findings and available guidelines and augments this through some original research with some case study organisations [34].

NORDINFO supported a study on the copyright questions related to the legal deposit of online material. This study concentrated on the European and Nordic legal environments. The study report concludes that there is a gap between copyright provisions and legal deposit objectives. Digital preservation requires copying, and copyright exceptions should allow this. Another point that comes up is that there may be moral rights issues arising from migration activities if they result in changes to the migrated material. The report suggests that technology, including electronic copyright management systems, can contribute to solving problems. There also needs to be some investigation into how depositories can cooperate with publishers to solve problems [35].

6.7 Costs

There is a serious problem with identifying the costs associated with digital preservation. Until full-scale operational systems have been running for some time, the nature and extent of costs cannot be known for certain. Some organisations have estimated the cost of caring for digital information. The British Library included some alternative costings in its proposal for the extension of legal deposit in the UK [36].

Hendley's study of different preservation methods and associated costs took into account diversity in digital materials [23]. The study drew on the work of a related set of studies, in particular the work carried out by the AHDS. However, it also reviewed other cost models and visited digital libraries and archives.

7. ACCESS

Traditionally, deposit libraries provide access to deposit publications without charge. It is clear that this situation may not be accepted by rightsholders in the digital environment [37]. It is likely that access to such deposited digital publications will be governed by licence agreements. Williamson gives a flavour of the potential complexity of providing access to digital information [38].

There is nothing in the literature that directly reports the views of users, so it is not clear what kind of access they would require. Access to printed legal deposit publications is often on a reference only basis. It is not clear whether users would be happy with the digital equivalent of this, or if they would want remote access. Neither is there any evidence that users would object to limitations on access rights, say limited or no printing or downloading of material.

Another legal issue is that of security. Authentication of users and the setting up of access rights are security measures that have implications for users and libraries. This issue is applicable to all types of digital library.

New technology provides the means of closely monitoring information use. The purpose of monitoring use may be to police user behaviour to ensure access agreements with publishers are not breached. Williams points out that logging use may be burdensome for deposit libraries [38]. There is also the potential for other uses of data, for example providing feedback to publishers, which may have data protection implications.

While stating that libraries accept that "the legitimate interests of publishers require that access is limited and controlled," Mackenzie Owen and Walle suggest that access should be encouraged to facilitate preservation of digital publications [8]. Their reason is that if electronic publications are not used for some time, they may be found to no longer work. On the other hand, a high level of access will "check the operability of electronic publications and ... identify and remedy any access problems that may occur."

Before using digital information, users have to be able to find it. In the digital environment, bibliographic records can have direct pointers to the material. However, as Mackenzie Owen and Walle point out, there can be several pointers, including to the original storage and the archival storage location [8]. Both of these may be physically located in the library. In the case of online publications, the original location may be a network address.

8. BUILDING THE INFRASTRUCTURE

8.1 Open Archival Information System reference model

A major initiative relevant to digital legal deposit collections is the development of the Open Archival Information System (OAIS) Reference Model [39]. The Consultative Committee on Space Data Systems is drafting this standard for the International Standards Organisation. An OAIS archive preserves information for access and use by a so-called Designated Community. This model is not just applicable to an organization that stores digital records; it can be applied to any type of digital paper material. The OAIS models the functions involved in the long-term storage of and access to digital information. These functions include acquisition and processing (ingest), archival storage, access, data management and administration of the archive.

The development of the OAIS has influenced other work being carried out in exploring the development of digital deposit collections. The CEDARS (CURL Exemplars for Digital Archives) project in the UK has used OAIS in developing its preservation metadata specification. The European NEDLIB (Networked European Deposit Library) project is working on developing an infrastructure for a European digital deposit collection [40]. British Library and the Koninklijke Bibliotheek in the Netherlands have used OAIS as the basis of their recent tenders for systems to manage their digital collections.

8.2 The NEDLIB Project

The NEDLIB project started at the beginning of 1998 and finishes at the end of 2000. Funding comes from European Commission, and the project leader is the Koninklike Bibliotheek in the Netherlands. The project partners are eight European national libraries, one national archive, three publishers and two information and communication technology companies. This project should be very influential in helping depositories cope with digital material.

The stated project aim [40] is to "develop a common architectural framework and basic tools for building deposit systems for electronic publications." The project deals with the technical issues involved in extending legal deposit to digital material. A great deal of detailed project material is available on the project Web site [41].

The project consortium adopted the OAIS model, but the NEDLIB Deposit System for Electronic Publications (DSEP) will be narrower in scope than the OAIS model. This is because some of the OAIS functions, such as Data Management and Access, are part of the general digital library environment and not specific to the digital depository. The aim is to link the functions of the deposit system and the digital library environment through interfaces.

NEDLIB is working with Jeff Rothenberg on an emulation experiment. The plan is that the first stage will result in a design for the whole experiment, a plan for testing and comparing the results of the emulations with the original works and a framework of preservation criteria and authenticity characteristics. The second stage involves modelling the emulation process and identifying metadata and functionality requirements. The last stage of the emulation experiment will be the implementation and evaluation of the emulation process in the testbed developed by the NEDLIB project

The first stage of the emulation experiment is now complete. Rothenberg concludes that "The results of this study suggest that using software emulation to reproduce the behaviour of obsolete computing platforms on newer platforms offers a way of running a digital document's original software in the far future, thereby recreating the content, behaviour, and 'look-and-feel' of the original document" [42]. This claim seems somewhat inflated since the actual experiment actually involved running Windows 95 publications on an Apple Mac using Connectix VirtualPC software as the emulator. The most Rothenberg can claim is that this particular software does what it says it does.

8.3 The BIBLINK Project

BIBLINK started in 1996 with support from the European Commission. Although the project came to an end in 1999, work on the initiative is ongoing under the aegis of the Conference of Directors of National Libraries.

The original aim of BIBLINK was to help develop and improve national bibliographic services, focusing on digital publications, especially online publications. Potentially all libraries would be beneficiaries of the project. However, the main perceived benefit was that national libraries would not miss the publication of significant publications [43].

The BIBLINK project developed a prototype demonstration system, called the BIBLINK Workspace. The demonstrator provides a virtual workspace or "computer mediated work environment" for participating parties. It allows publishers to create records and allows participants to access the system to retrieve, update and delete records in the workspace. BIBLINK is developing an Exploitation Plan [43]. This will provide a framework for library partners to assess the possibility of incorporating the system into operational procedures.

8.4 CURL Exemplars for Digital Archives (CEDARS) Project

The CEDARS project in the UK is funded by the Joint Information Systems Committee of the Higher and Further Education Funding Councils through the eLib Programme. CEDARS began in April 1998 and is due to finish in March 2001. The aim of CEDARS is to 'address strategic, methodological and practical issues and provide guidance in best practice for digital preservation' [44].

The CEDARS team is a partner in a new project on emulation for preservation funded through the JISC/NSF (US National Science Foundation) International Digital Libraries Programme [45]. The other project partners are based in the University of Michigan. The project will 'develop a small suite of emulation tools, evaluate the costs and benefits of emulation as a preservation strategy for complex multi-media documents and objects, and develop models for collection management decisions that would assist people in making 'real life' decisions.

9. PILOT DEPOSITORIES

9.1 National Library of Canada

The National Library of Canada ran such a project between 1994 and 1995. The purpose of the Electronic Publications Pilot Project (EPPP) was to pilot the acquisition, cataloguing, preservation and provision of access to a few Canadian electronic journals and other publications available via the Internet [46]. The National Library of Canada is now building a full-scale electronic collection.

9.2 Koninklijke Bibliotheek, Netherlands

The Koninklijke Bibliotheek in the Netherlands took the decision to collect digital publications in 1994 [47, 48]. Offline publications, such as CD-ROMs were stored on the stacks with the books. From 1995, the KB experimented in handling online publications on a small scale. Three publishers cooperated with the KB by agreeing to deposit some of their electronic publications with the KB. In 1996, the KB reached a provisional agreement with publishers to widen deposit. This small-scale deposit system was based on the IBM Digital Library system and became operational in 1998 [48]. The KB is now setting up a full-scale system [49].

9.3 National Library of Australia

The National Library of Australia set up the PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) project in 1996. The aim of the project was to "develop policies and procedures for the selection, capture, archiving and provision of long-term access to Australian electronic publications." The Library developed a proof-of-concept archive of Australian Internet material, which has been used to develop policies and procedures for the long-term preservation and access to digital publications [50].

The National Library of Australia realised that it needed integrated systems for managing all parts of its collections, including digital material. The NLA is taking this forward with its Digital Services Project. This project will provide storage for its digital material, but it will also provide management systems for most of the Library's collections.

9.4 Helsinki University Library (National Library of Finland)

EVA was originally an eighteen month Finnish project that started in June 1997. The aim of the project was "to test methods of capturing, registration, preserving and providing access to ... online documents ..." [51]. EVA was used to test tools being developed by various Nordic projects, including a Dublin Core metadata template and converter, URN generator and a harvesting and indexing application. Documents were harvested from the World Wide Web using the harvester. Once captured, the documents were analysed, indexed then archived. The EVA II and EVA III projects have been building on this work [52].

9.5 Kungliga Biblioteket, Sweden

The Kungliga Biblioteket in Sweden is currently running the Kulturarw3 project. The aim of this project is to "to test methods of collecting, preserving and providing access to Swedish electronic documents which are accessible on line in such a way that they can be regarded as published." [18]. The project aims to collect Swedish material available on the Internet according to specified selection criteria and to automate the collection through the use of robots.

10. CONCLUSIONS

There is a great deal of activity in this area worldwide. While exploratory work has identified many problems arising from the legal deposit of digital publications, a great deal more work is needed to solve these problems. Much of the work has been small-scale or national, yet the problems transcend national boundaries. Therefore, initiatives such as NEDLIB are to be welcomed.

The OAIS reference model provides a conceptual outline of the processes involved in a digital depository. However, by its nature it does not consider how theseprocess will be carried out. The NEDLIB project is attempting to develop a generic technical infrastructure for digital depositories. However, the project is limited in its scope and does not deal with the interface between the depository and the depositors.

Much of the activity in this area is concerned with technical issues. There is little evidence of any work taking a wide view of managerial or organisational issues. These issues include the management of workflows in depositories, staffing and skills requirements. It is also likely that there will have to be a lot more cooperation between publishers and depositories to facilitate deposit, preservation and access. Access especially may require negotiation. While there is an increasing interest in user needs in digital library research, there is little evidence of this in the legal deposit context.

Current work assumes that deposit systems will be organised in a similar way to current systems in that material will be physically deposited in deposit libraries. Physical deposit may be necessary to ensure long-term preservation, but alternatives to the current system could be considered. While the concept of a comprehensive archive of the national intellectual output may remain an ideal in an increasingly knowledge intensive world, it is not yet clearwhether this is technically and organisationally feasible or affordable.

11. ACKNOWLEDGEMENTS

I wish to acknowledge part-funding of this research by the Loughborough University Faculty Studentship scheme.

12. REFERENCES

[1] Muir, A. Legal deposit of digital material in the UK: recent developments and the international context. Alexandria, 12(3), (2000), 151-165.

[2] Russell, K. Digital preservation: ensuring access to digital materials into the future. http://www.leeds.ac.uk/cedars/Chapter.html.

[3] Working Definitions of Commonly Used Terms (for the purposes of the Cedars Project). http://www.leeds.ac.uk/cedars/documents/PSW01.htm.

[4] Borbinha, J., Cardoso, F., and Freire, N., NEDLIB glossary. http://www.kb.nl/coop/nedlib/glossary.pdf.

[5] Martin, D. Definitions of publications and associated terms in electronic publications. British Library Research & Development Department, 1996.

[6] MacKenzie, G. Searching for solutions: electronic records problems worldwide. Managing information, (July/August 2000), 59-65.

[7] EPS Ltd. The legal deposit of online databases. British Library Research & Development Department, 1996.

[8] Mackenzie Owen, J.S.and Walle, J.v.d. Deposit collections of electronic publications. Office for Official Publications of the European Communities, 1996.

[9] Bide, M., Potter, E.J., and Watkinson, A. Digital preservation: an introduction to the standards issues surrounding the archiving of non-print material. Book Industry Communication, 1999.

[10] Webb, C. Long-term management and preservation of publications on CD-ROMs and floppy disks: technical issues. http://www.nal.gov.au/niac/meetings/tech.html.

[11] The Digital Object Identifier System. http://www.doi.org/.

[12] Uniform Resource Names (urn). http://www.ietf.org/html.charters/urn-charter.html.

[13] Arvidson, A. and Lettenstrom, F. The Kulturarw3 Project - the Swedish Royal Web Archive. The Electronic Library, 16(2 April 1998), 105-108.

[14] Kahle, B. Archiving the Internet. http://www.archive.org/sciam_article.html.

[15] Mandel, C.A. Enduring access to digital information: understanding the challenge. LIBER quarterly, 6 (1996), 453-464.

[16] Digital Services Project. http://www.nla.gov.au/dsp/.

[17] Lounamaa, K. and Salonharju, I. EVA- the acquisition and archiving of electronic network publications in Finland. Tietolinja news, 1 (1999). http://www.lib.helsinki.fi/tietolinja/0199/evaart.html.

[18] Kulturarw3 heritage project. http://kulturarw3.kb.se/html/projectdescription.html.

[19] Rothenberg, J. Ensuring the longevity of digital information. http://www.clir.org/programs/otheractiv/ensuring.pdf.

[20] Van Bogart, J.W.C. Mag tape life expectancy 10-30 years. http://palimpsest.stanford.edu/bytopic/electronic-records/electronic-storage-media/bogart.html.

[21] Van Bogart, J.W.C. Magnetic tape storage and handling. http://www.clir.org/pubs/reports/pub54/.

[22] Lehmann, K.-D..Making the transitory permanent: the intellectual heritage in a digitized world of knowledge in Books, bricks & bytes. American Academy of Arts and Sciences, Cambridge, Mass., 1996, 307-329.

[23] Hendley, T. Comparison and methods & costs of digital preservation. British Library Research and Innovation Centre, 1998.

[24] Task Force on Archiving of Digital Information. Preserving digital information: report of the Task Force on Archiving of Digital Information. RLG, CPA, 1996.

[25] Feeney, M. Towards a national strategy for archiving digital materials. Alexandria, 11(2), (1999), 107-121.

[26] Woodyard, D. Farewell my floppy: a strategy for migration of digital information. http://www.nla.gov.au/nla/staffpaper/valadw.html.

[27] Rothenberg, J. Ensuring the longevity of digital documents. Scientific American, 272(1), (1995), 42-47.

[28] Rothenberg, J. Avoiding technological quicksand: finding a viable technical foundation for digital preservation. Council on Library and Information Resources, 1999.

[29] Ross, S.and Gow, A., Digital archaeology: the recovery of digital materials at risk. British Library Research and Innovation Centre, 1999.

[30] CERBERUS. http://www.kb.nl/kb/sbo/dnep/cerberus-en.html.

[31] Haynes, D., et al. Responsibility for digital archiving and long term access to digital data. Library Information Technology Centre, 1997.

[32] Matthews, G., Poulter, A., and Blagg, E., Preservation of digital materials policy and strategy issues for the UK. British Library Research and Innovation Centre, 1996.

[33] Digital Preservation. http://www.jisc.ac.uk/dner/preservation/.

[34] [34] Jones, M.and Beagrie, N. Preservation management of digital materials workbook: pre-publication draft, . 2000.

[35] Mauritzen, I.and Solbakk, S.A. A Study on copyright and legal deposit of online documents. NORDINFO, 2000.

[36] British Library, Proposal for the legal deposit of non-print publications to the Department of National Heritage from the British Library. British Library, 1996.

[37] Wille, N.E. Legal deposit of electronic publications: questions of scope and criteria for selection in Legal deposit with special reference to the archiving of Electronic publications: proceedings of a seminar organised by NORDINFO and the British Library (Research and Development Department) (Windsor, England 27-29 October 1994). NORDINFO, 1994.

[38] Williamson, R. Access and security issues from a publisher's perspective in Legal deposit with special reference to the archiving of Electronic publications: proceedings of a seminar organised by NORDINFO and the British Library (Research and Development Department) (Windsor, England 27-29 October 1994). NORDINFO, 1994.

[39] Consultative Committee for Space Data Systems. Reference model for an Open Archival Information System (OAIS). http://ftp.ccsds.org/ccsds/documents/pdf/CCSDS-650.0-R-1.pdf.

[40] Werf-Davelaar, T. NEDLIB: Networked European deposit library. Exploit interactive, 4 (2000). http://www.exploit-lib.org/issue4/nedlib/.

[41] NEDLIB. http://www.kb.nl/coop/nedlib/homeflash.html.

[42] Rothenberg, J. An Experiment in using emulation to preserve digital publications. http://www.kb.nl/coop/nedlib/results/emulationpreservationreport.pdf.

[43] Noordemeer, T.C. A Bibliographic link between publishers of electronic resources and national bibliographic agencies: Project BIBLINK. Exploit interactive, 4, (2000). http://www.exploit-lib.org/issue4/biblink.Cedars project summary. http://www.leeds.ac.uk/cedars/documents/MGA04.htm.

[44] JISC/NSF International Digital Libraries Programme: Emulation for Preservation: Project Summary. http://www.leeds.ac.uk/cedars/JISCNSF/summary.htm.

[45] Electronic Publications Pilot Project Team. Electronic Publications Pilot Project (EPPP): final report. http://www.nlc-bnc.ca/pubs/abs/eppp/ereport.htm

[46] Noordemeer, T.C., Deposit for Dutch electronic publications: research and practice in the Netherlands in Research and advanced technology for digital libraries: first European Conference, ECDL'97 (Pisa, Italy, September 1-3, 1997). Springer, 1997.

[47] Steenbakkers, J. Developing the Depository of Netherlands Electronic Publications. Alexandria, 11(2), (1999), 93-105.

[48] Werf-Davelaar, T.v.d. DNEP 1995-2000: the Dutch Deposit of Electronic Publications. http://www.bic.org.uk/Titia.ppt.

[49] Relf, F.A. PANDORA - towards a national collection of Australian electronic publications. http://www.nla.gov.au/nla/staffpaper/ashrelf1.html#abst.

[50] Eva - the Acquisitions and archiving of electronic network publications. http://renki.lib.helsinki.fi/eva/english.html.

[51] Eva - elektronisen verkkoaineiston hankinta ja arkistointi. http://www.lib.helsinki.fi/eva/index.html.

Comprehensive Access to Printed Materials (CAPM)

G. Sayeed Choudhury[1]
sayeed@jhu.edu
Mark Lorie, Erin Fitzpatrick, Ben Hobbs[2]
{mlorie,bhobbs}@jhu.edu
Greg Chirikjian, Allison Okamura[3]
{gregc,aokamura}@jhu.edu

Nicholas E. Flores
Department of Economics
University of Colorado at Boulder
Nicholas.Flores@colorado.edu

ABSTRACT

The CAPM Project features the development and evaluation of an automated, robotic on-demand scanning system for materials at remote locations. To date, we have developed a book retrieval robot and a valuation analysis framework for evaluating CAPM. We intend to augment CAPM by exploring approaches for automated page turning and improved valuation. These extensions will results in a more fully automated CAPM system and a valuation framework that will not only be useful for assessing CAPM specifically, but also for library services and functions generally.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval – *Digital Libraries.*

General Terms

Measurement, Design, Economics, Experimentation

Keywords

Information economics, evaluation methods, browsing, digital conversion, digital preservation, robotics, paper manipulation.

1. INTRODUCTION

Libraries face both opportunities and challenges with the development of digital libraries. While some digital library initiatives focus on materials that are "born digitally," most libraries continue to acquire materials in print format and must consider methods to manage their existing, substantial print collections. This approach of developing digital library capabilities while also managing large print collections has led to space pressures. Given the relatively high costs of building new facilities at central campus locations, many libraries have either built, or considered plans to build, off-site shelving facilities to accommodate their growing print collections.

While moving materials to off-site locations mitigates space pressures, patrons lose the ability to browse these materials in "real-time." Given the relative ease of accessing electronic resources, it is possible that patrons are less likely to access printed materials in remote locations. With the ongoing migration of materials to off-site facilities, there is a risk that a growing body of knowledge will be considered less frequently, and perhaps even ignored.

The CAPM Project began with the goal of restoring browsability through an automated, robotic system that would allow patrons to browse, in real-time, materials shelved at off-site facilities through a Web interface. We envision a patron, upon noting that an item is shelved off-site, will choose the CAPM option through their Web browser. Subsequently, a robot will retrieve the requested item and deliver it to a scanner. Another robotic system will turn pages at the patron's request. The patron will either view or print pages, and eventually "return" the item or request the item for physical delivery. Once the text is scanned, the patron may also perform automated text analysis options, such as keyword searches on the full text. In each case, the system will respond accordingly via remote control.

To date, we have built a book retriever robot that is being tested in a small-scale version of the shelving arrangement used at Moravia Park off-site shelving facility of the Johns Hopkins University. Additionally, we have conducted a valuation analysis to evaluate the potential costs and benefits of CAPM.

2. PROJECT BENEFITS

The most obvious benefit of CAPM is the ability to restore browsability for materials shelved in off-site locations. The possible risk of reduced usage for off-site materials might be mitigated with this remote browsing capability. Essentially, CAPM will "elevate" the status of materials in off-site locations towards that of electronic resources.

While achieving browsability is a noteworthy benefit of the CAPM project, it is one only facet of the project's full potential. By combining OCR software and a search engine being developed at Johns Hopkins, it will be possible to search requested texts and generate keywords through natural language processing of full text. Additionally, the CAPM system will generate tables of contents. The combination of full text, tables of contents and keywords will be cross-referenced within a database to identify items with similar content. This capability, combined with traditional metadata, will emulate the serendipitous discovery of open-stack browsing.

While these benefits are related to access, CAPM could also provide preservation-based benefits. Preservation copies will be created by a batch-scanning mode when patrons do not require the system. Because of the enhanced access offered by CAPM, libraries and patrons might be more amenable to the transfer of

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

items to off-site facilities which generally offer superior conditions for materials (e.g. temperature, humidity, static shelving).

3. BENEFITS VALUATION

The benefits outlined above need to be properly delineated and evaluated before libraries will adopt CAPM. Consequently, in addition to the engineering development of CAPM, there has been a concurrent, valuation analysis. The analysis has two emphases: an engineering cost analysis and a valuation of potential benefits for patrons. The cost analysis has focused on the following cost categories: labor, increased electricity usage, maintenance of equipment, book containers, loss of storage capacity associated with book containers, equipment and setup or implementation costs. The costs were calculated as a levelized average cost per use over a ten-year period. Preliminary results indicate an average cost per use of between $3.50 to $24.52, depending on level of use. These costs compare favorably to costs associated with interlibrary lending services [3].

These potential costs will be compared to the potential benefits, as estimated through a contingent valuation methodology (CVM). CVM represents a "stated-choice" technique where individuals express a willingness-to-pay for benefits or a willingness-to-accept compensation for costs, using dollar values. CVM has been used widely for evaluation of non-market goods, especially in the environmental field. A previous application of CVM in the library environment provides evidence of its utility in decisions regarding resource allocations for libraries [2]. For the CAPM project, patrons were queried regarding hypothetical choices (related to potential features of CAPM) through an online survey. The survey design was developed with the help of a workshop that convened economists and librarians and was pre-tested with a small sample group. The usefulness of this survey results and overall valuation analysis will be assessed during a workshop in May 2001.

4. FUTURE WORK

Both the engineering and valuation activities of CAPM will be extended. We will test the retriever robot at Moravia Park and explore the development of page-turning devices. There are page-turning devices but they are used in controlled settings, such as for patients undergoing rehabilitation [1, 4]. We will investigate and build prototype page-turning devices that will accommodate a range of paper types.

The valuation analysis will be augmented in two ways. We will conduct a post-survey investigation to assess the effectiveness of online surveys. Additionally, there is some evidence that library patrons have difficulty with dollar values for library services [5]. Consequently, we will explore the use of multi-criteria decision-making techniques (MCDM) to evaluate CAPM. The MCDM techniques will involve assessing patrons' and library administrators' reactions to questions involving tradeoffs of service attributes (such as time) and comparing these reactions to questions involving dollar values. This exploration will provide a more comprehensive evaluation of CAPM and a generalized framework for evaluating library services, both from the perspective of patrons and library administrators.

5. CONCLUSIONS

The CAPM Project represents an innovative response to fundamental issues related to digital library development. The system will provide an automated system for on-demand and batch scanning of materials in off-site locations. Additionally, the valuation analysis will result in a rigorous assessment of CAPM and a general framework for evaluation of library services.

Other libraries could adopt CAPM either by implementing a CAPM system in their facilities or by shelving their materials in a facility with CAPM or by accessing CAPM-scanned materials through the Web. Even at this intermediate stage of development, libraries within the United States, Europe and Asia have expressed interest in CAPM for both research interest and potential implementation.

6. ACKNOWLEDGMENTS

The CAPM Project has been supported by grants from the Council of Library and Information Resource and the Mellon Foundation. The Minolta Corporation has provided scanners for the project. An individual donor has also provided financial support.

7. REFERENCES

[1] Danielsson, C., and Holmberg, L. Evaluation of RAID workstation in Proceedings of the RESNA '94 Annual Conference (Nashville, TN, June 1994), 451-453.

[2] Harless, D.W., and Allen, F.R. Using the contingent valuation method to measure patron benefits of reference desk service in an academic library. College and Research Libraries, 60 (1), 59-69.

[3] Jackson, M.E. Measuring the Performance of Interlibrary Loan Operations in North American Research & College Libraries. Association of Research Libraries, 1998

[4] Neveryd, H., Bolmsjö, G., and Eftring, H. Robotics in Rehabilitation. IEEE Transactions on Rehabilitation Engineering, 3 (1), 77-83.

[5] Saracevic, T., and Kantor, P.B. Studying the value of library and information services. Part I. Establishing a theoretical framework. Journal of the American Society for Information Science, 48 (6), 543-563.

[1] Digital Knowledge Center, MSE Library, Johns Hopkins University

[2] Department of Geography and Environmental Engineering, Johns Hopkins University

[3] Department of Mechanical Engineering, Johns Hopkins University

Technology and Values: Lessons from Central and Eastern Europe

Nadia Caidi
Faculty of Information Studies
University of Toronto
Toronto, Ontario M5S 3G6 (Canada)
(416) 978 4664
caidi@fis.utoronto.ca

ABSTRACT

Technology does not develop independently of its social context. Rather, there is a range of social, cultural and economic factors (in addition to technical factors) that define the parameters for the development and use of technologies. This paper presents a case study of the social shaping of one aspect of digital libraries, the development of national union catalogs (NUC), in four countries of Central and Eastern Europe (CEE). It examines the specific choices and values that are embedded in the design of a NUC, and how these might be transferred to other cultural contexts.

Categories and Subject Descriptors

[Social Issues/Implications]: Social Informatics. **[Design-Implementation/Evaluation]**: Case Studies, DL development. **[Policy Issues]**: Transborder/International Issues.

General Terms

Design.

Keywords

Information Infrastructure. Central and Eastern Europe. National Union Catalogs. Social Shaping of Technology.

1. INTRODUCTION

The study of digital libraries (DLs) in Central and Eastern Europe makes a fascinating case study of a socio-technical system in the making. This paper builds on prior work on DLs in the CEE region ([1] [2]), and constitutes a preliminary report of a large research project on the development of information infrastructures and DLs in four CEE countries (Czech Republic, Slovakia, Hungary and Poland) [3]. Developing DLs (understood here in its wider sense, as a mixture of content, technology and people) is essentially a collaborative activity in which various actors are pooling resources and efforts to devise solutions to a specific problem. In the four CEE countries studied, the most interesting debates around digital libraries revolved around the development of national union catalogs (NUC) as a means to creating a shared cataloging system, and making their country's bibliographic records available online. The NUC emerges as a result of the "work" of the various coalitions (libraries, system vendors, funding agencies, policy makers).

Here, we report results from a series of interviews conducted in four CEE countries with key players in the development of NUC projects. The outcomes of these projects show mixed results primarily because of the different approaches being used and the various assumptions embedded in the designs of these systems. Our study shows that beyond the technical and design issues, there are socio-cognitive and socio-political dynamics that contribute to (or hamper) the development of an NUC.

2. METHOD

The choice of these four countries was motivated by their similar level of economic development, the advanced state of reforms undertaken, the political stability, as well as the level of support they receive from Western countries. The focus of the study was on the major academic and research libraries (including the national libraries, and libraries of the national academies). These libraries have shown leadership in adopting information technologies and deep concern with information access. As such, they play important roles in policy decisions regarding library services in their countries.

Most of the data collected were in the form of in-depth interviews, in which a broad agenda of research questions was presented to respondents about their visions and experiences with DLs, NUCs, and the information infrastructure of their country. Face-to-face interviews were conducted between March and June 1999 with 49 library leaders and policy-makers, in 37 institutions (10 in Czech republic, 10 in Hungary, 13 in Poland and 4 in Slovakia). A dozen more informal discussions with other respondents, library vendors, project managers were also undertaken.

3. FINDINGS

These CEE countries faced with the challenges of integration into a global economy and with domestic social and economic transformation from the socialist system. Similarly, the story of CEE libraries is one of institutions striving to survive the social transformations of their countries, while adjusting their organizational structure to the new reality and needs of society and of their user community. The data collected showed evidence both of the presence of common frames (e.g., a set of shared assumptions and beliefs) and of power struggles between various

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

social actors with conflicting interests and agendas (e.g., between types of libraries; between libraries and state agencies; between libraries and the private sector).

3.1 Boundaries and Negotiation

Transition involves the building of new institutions and the creation of new linkages between organizations, which is an eminently political process. The power struggle that ensues was evident in the ways in which most respondents framed their world in terms of binary oppositions (e.g., 'Us' vs. 'Them,' consortia 'insiders' vs. 'outsiders,' East vs. West) and what divides rather than what unites libraries. Similarly, the respondents' rhetoric around the development of an NUC revolved around the obstacles and barriers of its development (e.g., lack of funding and leadership, poor coordination between governing agencies, conflict between libraries, rigid management and structure).

Power is also expressed in the ways in which meanings and definitions are assigned to a technological artifact. For instance, the various conceptions of the nature of 'information' shape the social construction of both digital libraries and the NUC. In the countries studied, 'information' was often viewed as a resource and a public good. The view of information as a commodity that could be sold, bought or exchanged was not the most prevalent among these respondents, who often resented the economic aspects associated with it. Most important to them was the need to create new organizational relationships in this transitional environment.

3.2 Cooperation and Resource Sharing

Establishing a climate of cooperation appeared throughout the data as a prerequisite for any further attempts at developing DLs and NUCs. Means to enable cooperation include shared cataloging cooperatives (e.g., networks, consortia) and the adoption of common standards. The earlier initiatives were mostly fostered by the academic and research library communities, with the help of western library-oriented philanthropic foundations. These foundations were instrumental in enabling and encouraging the transfer of technology. More importantly, these foundations (soon followed by governing agencies in the country) strongly encouraged libraries to form consortia. The financial incentives and the realization of the benefits of resource sharing led to a series of alliances such as the Krakow Library Project and NUKAT (Poland); CASLIN (Czech Republic and Slovakia), HUSLONET and MOKKA (Hungary).

However, after decades of centralization that led to 'artificial' library networks organized by subject areas and/or types of libraries (and imposed on libraries by the ministries), respondents resented the 'forced' cooperation model newly imposed on them. Rather, they expressed more interest in experimenting, in a grass-root manner. The variety of consortia created denotes this trend: while a centralized model was opted for in Slovakia, respondents in Czech Republic, Hungary and Poland favored a mixture of centralization and decentralization for the development of NUCs: tight or loose alliances (CASLIN), resource-sharing consortia (e.g., HUSLONET), highly distributed model (VTLS Group).

3.3 Global vs. Local Considerations

The rush toward adopting internationally recognized standards in the field of networking, technology, and library services was associated with the need to prepare for the integration to the European Union, and to "catch up with the West." In the respondents' accounts, and in observing the societies, there seems to be an uncritical glorification of the Western way of life, economic and political arrangements, and technology. Western integrated library systems were often preferred to local systems developed in the country. Similarly, most standards were often borrowed from foreign sources (e.g., USMARC, the French RAMEAU, Library of Congress Subject Headings, etc.). The language of the West was often adopted and adapted: for instance, some respondents used such terms as "information have and have-nots," "Telematics," "intelligent cities," "virtual libraries," etc. It is not clear, however, whether the use of these terms was fully promoted nationally, or just geared toward westerners.

This interest in everything "Western" has to be contrasted with a growing sentiment of cultural identity and nationalism that has appeared throughout the CEE region. In the library field, this has led to decisions to maintain HUNMARC (the Hungarian version of MARC) over USMARC or UNIMARC in Hungary, as a means to keep the distinctiveness of the country. Too often, the introduction of foreign cataloging and classification systems is viewed merely as a technology transfer, and the emphasis on the cultural values is usually neglected. Reconciling the national tradition with the universal mechanisms of globalization is key.

4. DISCUSSION AND IMPLICATIONS

Digital libraries are a complex arrangement of people, technology, institutions, and content. As such, they have to be understood as more than the conduits (pipes) or the content that flows through these pipes, to include elements of power and culture. DLs cannot be replicated. Rather, they fit to the unique situation of the environment. The CEE countries studied are at the important stage of positioning or redefining the role of the various players. Rather than asking whether these countries have or not DLs, it may be more useful to ask when and who takes part in their development; and how this relates to the free flow of information, and the advancement of democratic values in the country. These issues will be developed further in future articles.

5. ACKNOWLEDGMENTS

Thanks to the A.W. Mellon Foundation for funding this project.

6. REFERENCES

[1] Borgman, C.L. From Gutenberg to the global information infrastructure. Access to information in the networked world. Cambridge, MA, London: MIT Press, 2000.

[2] Lass, A. and R. Quandt. (eds.). Library automation in transitional societies. Oxford University Press, NY, 2000.

[3] Caidi, N. The information infrastructure as a discursive space: A case study of the library community in Central and Eastern Europe. PhD dissertation, Department of Information Studies, UCLA, 2001. (Adviser: C.L. Borgman).

Use of Multiple Digital Libraries: A Case Study

Ann Blandford
Middlesex University
Bounds Green Road
London, N11 2NQ, U.K.
+44 20 8362 6163
A.Blandford@mdx.ac.uk

Hanna Stelmaszewska
Middlesex University
Bounds Green Road
London, N11 2NQ, U.K.
+44 20 8362 6905
H.Stelmaszewska @mdx.ac.uk

Nick Bryan-Kinns
Icon MediaLab London
1, Martha's Buildings, 180 Old Street
London EC1V 9BP, U.K.
+44 20 7549 0206
nickbk@acm.org

ABSTRACT

The aim of the work reported here was to better understand the usability issues raised when digital libraries are used in a natural setting. The method used was a protocol analysis of users working on a task of their own choosing to retrieve documents from publicly available digital libraries. Various classes of usability difficulties were found. Here, we focus on *use in context* – that is, usability concerns that arise from the fact that libraries are accessed in particular ways, under technically and organisationally imposed constraints, and that use of any particular resource is discretionary. The concepts from an Interaction Framework, which provides support for reasoning about patterns of interaction between users and systems, are applied to understand interaction issues.

Keywords

Digital Libraries, video protocols, interaction modelling, HCI.

INTRODUCTION

Digital libraries are moving from research and development into commercial use. If they are to realise their full potential, however, the experiences of end users need to be taken into account from the earliest stages of design. Those end users are typically individuals who have no particular skills in information retrieval, and are accessing library resources from their own desks, without support from a librarian.

These factors clearly have implications for design. In particular, libraries need to be "walk up and use" systems that are easily learned; while more experienced users may require powerful features that support them in performing focused searches, novice users need to get early results for minimal effort [5]. The study reported here investigated how comparative novices work with existing digital libraries, focusing particularly on the patterns of interaction and difficulties they experienced. Many of these difficulties are not apparent when a library is tested in isolation. Whereas users of physical libraries have to move very deliberately from one library to another – often with a substantial time interval between the use of one and the use of the next – users of digital libraries can move almost seamlessly between them, sometimes without even noticing the transition. The focus of this study is on use of multiple digital libraries within a single session, and on user experiences of the interaction.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

BACKGROUND

We briefly review related work on use in context and on types of use. These other studies provide the academic context for the study presented here. We then present an overview of methods used in previous studies, relating the methods to the kinds of findings those studies could establish. In particular, an understanding of the alternatives was used to guide the design of the study reported here.

1.1 Use in context

Several studies have looked at digital libraries in context – that is, not just the library itself, but also how it sits within a larger frame of use. One example of such studies is Bishop's [3] consideration of the use of digital libraries by people from different social and economic backgrounds. Her studies indicate that people from different backgrounds (low-income and academic) can easily be put off using digital libraries – small problems tend to be magnified until they deter potential users, and lack of awareness of library coverage often prevents users from understanding what they could get out of the libraries. As libraries are becoming increasingly available for general use, the finding that people are easily deterred has to be taken seriously; our results, as presented below, highlight some of the deterrent factors, including poor reliability, inadequate feedback and the time taken to familiarise themselves with a new library.

Covi and Kling [7] investigated patterns of use of digital libraries by different groups of users, and how they vary across academic fields and universities. They focussed on interviewing potential users and, moreover, were concerned primarily with university members, rather than considering the population at large. Their study led them to conclude that the development of effective (useful and used) digital libraries needs to take account of the important roles played by other people within the broader system of use (notably colleagues and librarians), and that the views of end users, as well as those of librarians and computer specialists, need to be understood for effective design. Whereas Covi and Kling were concerned with how library use fitted in with overall working patterns, the study reported here is looking at how a particular interaction between a user and digital libraries evolves – i.e. the patterns of interaction, rather than the patterns of use. As shown below, the decisions taken by computer scientists and librarians have substantial impact on the user experience, in ways that are hard to anticipate.

1.2 Types of work with digital libraries

One approach to understanding users' views is to analyse the kinds of work people perform with library resources. In this section we summarise studies of the kinds of things people get up to, or might get up to, in a digital library.

Searching for interesting articles is often the first thing that comes to mind when considering "digital libraries". However, as is being increasingly recognised, searching is not just a case of entering a search term and viewing a list of results. Furnas and Rauch [9] found that in searching for information a "one-shot query" is very rare. More typical is an extended and iterative search which involves opportunism; that is, the searching evolves over a period of time and relies on users being able to follow new (interesting) paths as they appear, which may not necessarily have been specified at the start of their search.

These notions of extended searching are supported by research carried out in conventional libraries [14]. As with digital library studies, O'Day & Jeffries found that one-shot searches were rare. Rather, they found that single searches evolved into other kinds of searching which they identified as: monitoring a topic over time; following an "information-gathering plan"; and exploring a topic in an undirected way. The findings of the current study are consistent with those of others, but illustrate that many searches are unsuccessful.

People do not just search for items in digital libraries, but also browse for them. Jones *et al.* [12] characterise this distinction as follows:

1. **Browsing** – users traverse information structures to identify required information
2. **Searching** – users specify terms of interest, and information matching those terms is returned by an indexing and retrieval system. Users may, in turn, browse these results in an iterative manner as discussed above.

Gutwin *et al.* [11] discuss the browsing in digital libraries, but tend to focus on how user interfaces they develop can support browsing, rather than considering what browsing is. However, in their discussion of browsing support they do categorise the purpose of browsing as follows:

1. **Collection evaluation:** What's in this collection? Is it relevant to my objectives?
2. **Subject exploration:** How well does this collection cover area X?
3. **Query exploration:** What kind of queries will succeed in area X? How can I access this collection?

These can all be considered aspects of familiarisation: with the type of information in collections and how the library works. In the study reported here, many of the purposes associated by Gutwin *et al.* with browsing arise in situations where there is no significant "browsing" activity – i.e. no substantial traversal of information structures. The purposes of an interaction and the types of behaviour (e.g. searching vs. browsing) do not appear to be closely linked. Nevertheless, the identification of these purposes is helpful: in the discussion below, these kinds of purposes are discussed under the heading of "familiarisation".

Once a document has been located a user typically does some work with it (not necessarily immediately afterwards) – otherwise, there would be no point in finding it in the first place. Amongst the activities associated with working with documents are reading and annotating them. One important finding about reading activity [1, 2, 13] is that people do not simply read articles from beginning to end, but rather move between levels of information – for example, from authors and titles to reading the conclusion. Current digital libraries available via the web are, for various reasons, limited in terms of the kinds of reading they support. Even with such technical restrictions, there were substantial variations between subjects in this study as regards the reading processes they adopted.

As well as reading, people also create, update, and annotate documents. O'Hara *et al.* [15] focussed on such writing activities in their studies of PhD students' use of libraries. They found that reading and writing were inextricably intertwined. Existing web-based library interfaces do not support any writing activities, so subjects in this study saved and printed relevant articles, for future organisation and annotation.

1.3 Techniques Used in Studies

Many different techniques have been used to study people's use of digital libraries. This section outlines the range of techniques used and their applicability. An understanding of these past studies and their scope was used to design the study described here.

Adler *et al.* [1] and O'Hara *et al.* [15] asked people to keep **daily notes** of their document activities, and followed up these descriptions with structured interviews. Such a technique has the advantage of low effort on the part of the analysts. However, a large load is placed on the participants as they have to keep notes of their activities. Moreover, as individuals keep their own diaries there will be differences between their note taking styles which may be significant and cause problems in generalising results. However, such an approach helps us to understand what people do and why they do it, which can give useful input to design or some qualitative evaluation results.

Questionnaires, on-line form filling, or registration documents can provide simple feedback. Bishop [3] used registration documents to build up an understanding of the different backgrounds of digital library users. In contrast she used surveys to find out information about users' use of the system after a period of time. Advantages of using such techniques include the ability to get a large response, but the information returned is often shallow – typically just simple answers to questions asked with no explanations of answer rationale. Theng *et al.*'s work [17] contrasts these approaches by using extensive questionnaires with a small group of users after they have completed tasks with digital libraries. These questionnaires elicit users' perceptions of the digital libraries used but, again, they provide no means of assessing why users felt as they did.

Several studies, e.g. [2, 3, 13], have employed **transaction logs** to gain an understanding of the activities users were engaging in with digital libraries. These logs give quantitative accounts of user actions and so can be used to make statements such as "(about 5%) took advantage of the ability to search for terms in individual components of articles" [2]. However, such logs do not provide an understanding of why users use particular features of systems. Understanding why things have happened is typically tackled by interview and possibly diary studies.

Several studies have investigated how people use particular interfaces and how their use differs between interfaces and between tasks. For example, Bishop [3] and Park [16] used **experimental design** to compare the applicability of user interfaces for digital libraries. The use of experimental design gives statistically significant results, but such studies are costly to develop and run, and can only answer specific questions. In these cases, the aim was broadly to inform design and redesign of a particular interface. Therefore, the user and a single interface could be considered as a "closed" system, independent of external influences. This contrasts with the present study, for which use in context is the primary concern.

Bishop [2] used three **focus groups** to elicit understandings of how faculty members used journal articles, and what requirements such use placed on design of digital libraries. Following on from that, Bishop [3] used focus groups to help understand different socio-economic backgrounds of digital library users. Although focus groups are useful for gaining an overview of the issues and problems, they tend to produce information which is often sketchy and in outline form.

Bishop [2, 3] used **interviews** to follow up on topics that were raised in her focus groups. This approach allows the analysts to develop a fuller understanding of the issues raised in the focus group, but interviews are time consuming and are, again, producing qualitative results which feed into design. Other studies such as those conducted by Covi and Kling [7] relied solely on interviews to assess people's perceptions and use of digital libraries. Approaches such as those employed by Furnas and Rauch [9] involved a combination of techniques as they used interviews to inform further observation of people using digital libraries. In contrast Marshall *et al.* [13] used interviews to follow up other techniques such as examination of transaction logs; their interviews allowed analysts to probe why users were performing certain patterns of interaction identified from the logs.

Observing what people do as they use systems is a time consuming activity. However, it can provide useful insights into the usability of systems. Bishop [3] discusses her use of observation to gather information on engineering work and learning activities. This information can then be used to inform (re)design, and/ or followed up in other ways such as interviews, as illustrated by Furnas and Rauch [9].

From the view of the studies presented here it is clear that although there has been work on studying the usability of specific user interfaces for searching, and to a lesser extent browsing, in digital libraries there has been little work on understanding the nature of these tasks or how libraries are used in a natural setting.

METHOD

The study reported here aimed to achieve a better understanding of how users interact with digital libraries within a single session, but not necessarily using a single library.

Because we wished to gather detailed interaction data, techniques such as diary-keeping, interviews, transaction logs and focus groups were inappropriate. A video-based observational study with think-aloud commentary was selected as the most appropriate means of gathering data, with a short debriefing interview a few days later to clarify any issues raised by the video data.

Five users were recruited for the study. Three of these were first year PhD students (referred to below as "A", "B" and "C"), one a final-year PhD student ("D") and one an experienced academic ("E"), all computer scientists. A – D were recruited as subjects specifically for this study; E, aware that this study was being conducted, offered to participate while performing a self-defined library searching task.

The aim was not to give users artificial tasks, which are liable to be either too precisely defined to be natural or too meaningless for participants, but to ask participants to select their own tasks to work on. Therefore the task defined for participants A – D was simply to obtain at least one paper on their own research topic to help with their literature review, using their choice of libraries from a given set (easily accessed via bookmarks in a web browser). They were asked to think aloud while working. They were provided with a little information about each library, as shown in Table 1. Access rights for each library were defined by the subscription held by the organisation in which they are based. Note, in particular, the restriction on ACM access – a cause of difficulties as discussed below.

Table 1: bookmarked libraries for users A – D.

ACM Digital library www.acm.org/dl/	Full text access only to journals and magazines (not conference proceedings)
IDEAL www.idealibrary.com	Access only to articles prior to 1998
NZDL www.nzdl.org	Full text articles
EBSCO www-uk.ebsco.com	Full text articles
Emerald www.emerald-library.com	Full text articles
Ingenta www.ingenta.com	Full text articles

As noted above, E was not recruited in the same way, but offered to participate. She was planning to search for articles on particular topics to help with writing academic papers, and volunteered to do this with a video camera running, and to "think aloud" while working on her self-defined task. Consequently, she used digital libraries of her own choosing, and did not have explicit information about limitations on access. She held a personal subscription to ACM, so she was not subject to the same restrictions as other users.

Users A, B, C, D and E worked with the digital libraries for 57, 62, 62, 51 and 80 minutes respectively. The video data was then transcribed, including speech and some description of interaction between user and computer system. It was analysed using the Interaction Framework [4]. Extracts from these transcripts are used in the following sections as source materials for examples.

1.4 Interaction Framework: overview

The Interaction Framework is an approach to describing actual or possible interactions between agents (users and computer systems) in terms of the communicative events that take place between those agents, and the patterns of interaction. Using this neutral language, which aims to take neither a user- nor a computer-centred view of the interactive system, we can identify and discuss properties of the interaction that might or might not be desirable. For example:

- **Blind alleys** are interactions such that the objective is unachievable, but where that fact does not become apparent until some way into the interaction.
- **Discriminable events** are ones that another agent can easily choose between.
- A **canonical interaction** is one that achieves all its objectives as efficiently as is theoretically possible. In the case of digital libraries, such objectives will include accessing particular papers, accessing papers on a particular topic or general subject area, and gaining **familiarisation** with a collection content, type or features.

Central to any successful interaction is the idea that users must have an adequate understanding of the state of the system. Put in neutral terms, the user and system must share "common ground" [6] – that is, each interactional event has to communicate sufficient information to enable the agents to maintain common ground.

SUMMARIES OF INTERACTIONS

Before presenting detailed results, summarising the important difficulties found, we present a brief overview of the interaction of each participant with the available digital libraries.

1.5 User A

User A was interested in papers on electronic commerce. As she started working, she spent a while browsing unrelated material, as if orienting herself to browsing, before selecting a link from the bookmark list. She selected the ACM digital library, and spent some time reading through the introductory page. She then searched for "the best of electronic commerce", but seemed rather confused by the results returned. She selected an alternative search mechanism and repeated the search. She found various articles that were "interesting" but did not print them. Although she limited her search in ACM to "journals only", conference articles were listed among the search results; when she tried to download one of these articles, she was asked to enter a user name and password, which she did not have, resulting in an authorisation failure.

She moved to the Computer Science Technical Reports link in NZDL, and got over a thousand hits on her search. She reformulated her search several times, and still too many items were returned. Eventually, she found an article she wished to view and download, but failed to download it, for reasons discussed in section 5.1 below. She then moved on to EBSCO. When her search results were returned, she commented that "this is more readable than from other libraries". She successfully found, saved and printed one article.

1.6 User B

User B was looking for material on knowledge management, text mining and link analysis. Although he has used other browsers, he did not have previous experience of using Netscape, so he starting by browsing Netscape pages before connecting to the ACM digital library via the bookmarks. Before specifying the search terms, he spent some time "trying to understand how to make the search". When searching the ACM library, he did not restrict his search to journals only, and consequently received many hits that were conference proceedings; like user A, he got authorisation failure when he tried to print any of these.

Although new to Netscape, user B was a relatively sophisticated user of information retrieval systems. For example, he understood how to use quotation marks selectively in search queries and also resorted quickly to using two browser windows so that he could continue working in one window while a document was downloading in the other. Using this strategy, he explored the EBSCO library, then NZDL, then Ingenta – continually flicking from one window to the other as pages were loading. One apparent consequence of this is that his behaviour was more reactive than that of the other users: he appeared not to form clear beliefs about the state of the system, but to simply respond to whatever was currently displayed, and the interaction appeared relatively unstructured and haphazard.

1.7 User C

User C was searching for material on Growing Cell Structures (GCS), text classification and Self Organising Maps (SOM). He started by accessing Emerald, but rapidly switched to Ingenta, when his first few query formulations returned no results. Results from Ingenta were more promising; in fact, there were so many hits he appeared to be overwhelmed: "I found over three hundred documents here". Several times, he selected an article with a promising title and followed the link "full text at Science Direct", but was then refused access. [The user organisation had a subscription to Ingenta journals but not to Science Direct.] He saved several abstracts and printed one full text article from Ingenta. He accessed the IDEAL and ACM libraries twice each, but each time judged them "too slow". In EBSCO, he failed to find any matches to his search terms. He returned to the Emerald library, and reformulated his query: "I think I made some mistakes last time. I searched for GCS as the keyword so this time I searched for GCS for full text and I found something". Later on, his search took him to a link "order the book", which took him from the library to an internet bookseller. He conducted further searches, using Emerald, Ingenta and NZDL, usually receiving either no matches or too many to deal with. In NZDL, he found an interesting article, but failed to download it. Although he could read the article on screen, he tried three times to download it, without success. By the end of his interaction, he had found some relevant material on text classification, but none on GCS or SOM.

1.8 User D

User D looked for different things in different libraries. He started with ACM, searching for articles on usability evaluation; a large number of results were returned. He tried to save a selected article to a "binder", but received "authorisation failure". He moved on to use NZDL, now searching for articles on "musical timbre"; although he found one that he skim-read on screen, he moved on quickly to search for "timbre perception", which returned results that he judged "more or less similar to my previous query". He opened second and third windows to access EBSCO and Emerald, but never switched between windows (they seemed just to be a historical record of where he had got to in a particular library). He browsed and submitted search queries in a relatively unstructured way, apparently trying to familiarise himself with the content and structure of the library. When he tried to follow links to particular journals, an error message, "Type mismatch" was displayed. Subsequent tests indicate that this was a temporary error, but it had a strong influence on this particular interaction.

He tried to access Ingenta, but received an error message: Ingenta was unavailable. He moved on to Emerald, about which he commented: "It's actually good to have some basic description of the journals. It's not my area." After a quick look, he returned to ACM to search for articles on visual texture; on selecting a "Find related articles" link, he found something quite different that interested him – a case of serendipity in the search. He viewed the abstract and tried to download the paper, but got an "authorisation failure" – he "wasn't paying attention" when earlier told that he could download journal papers but not conference proceedings. He went on to successfully identify and download a relevant journal article.

1.9 User E

User E required information on diary systems, cognitive modelling and usability of Artificial Intelligence systems. In meeting these objectives she used three libraries: ACM, IDEAL and the New Zealand Digital Library (NZDL). These were used in a relatively orderly sequence of ACM, followed by IDEAL, and finally the NZDL. The search objectives were repeated to some extent with each information source as opposed to meeting each objective in sequence.

In detail, she spent most of her time using the ACM DL. Any articles she found that seemed relevant were printed for further review. Her first few searches were unsuccessful, yielding "no matches". A search for information on diaries and calendars was more fruitful. She then switched to browsing various collections, with mixed success. When she found an article that appeared interesting, she always printed it out, rather than trying to read it on screen. Overall, this user's interaction was the most fruitful, and by the end she had printed about a dozen articles.

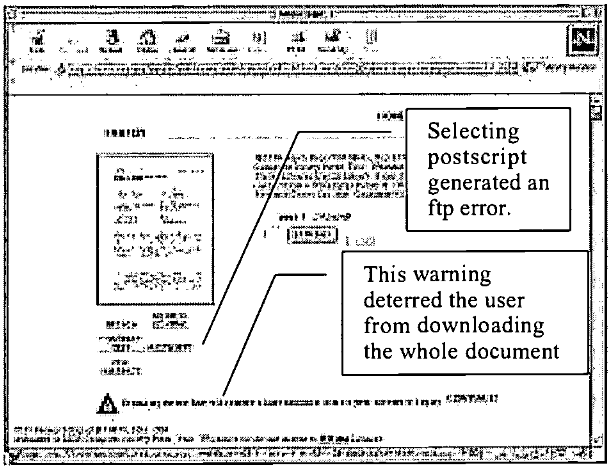
RESULTS

Clearly, each of the five users in this study behaved in quite different ways when navigating the libraries, and the sample size is far too small to make any general claims about usage patterns or typical behaviours; this was not the purpose of the study. Similarly, the purpose has not been to conduct usability evaluations of particular libraries, or to pit them against each other. Rather, the aim has been to identify core usability issues that arise through the details of user interaction with digital libraries, where the users are not assumed to be interaction retrieval experts, and where use of any particular library is discretionary.

We discuss the main usability issues raised within this study under two headings: deterrent factors (and encouraging ones) that may influence future acceptance of digital libraries; and usability issues that arise as a direct consequence of libraries not being "closed" – i.e. that the interface between a library and other resources is easily traversed.

1.10 Deterrent factors for use in context

The users in this study all experienced substantial difficulties with using one or more of the libraries they accessed. Many of these difficulties resulted in blind alley interactions – that is, interactions that did not achieve the user's objectives. Some of these blind alley interactions took place over several minutes. For example, user A spent over six minutes trying to download an article from one library before giving up, commenting that "I've tried all, but I can't download". Some of her sources of difficulty are illustrated in Figure 1, where the warning message reads: "Expanding the text here will generate a large amount of data for your browser to display".



Selecting postscript generated an ftp error.

This warning deterred the user from downloading the whole document

Figure 1: difficulties with downloading in NZDL

Most users suffered from a "triumph of hope over experience", repeating the same unsuccessful action several times without modification, and each time receiving the same result. For example, user B tried eight times to download full text of an article from Science Direct, while user C did the same six times. User C also tried to download the same postscript file three times, each time receiving the error message "Netscape is unable to find the file or directory named /pub/techreports/1993/tr-93-025psZ. Click the file name and try again". Similarly, user D tried three times to access the Ingenta library, each time receiving the error message "TfeLogin: Data decoding / encoding error". It is too early to follow up these users to find out whether or not their experience working with these libraries on this occasion has deterred them from any future use, but the productivity of all participants in the study was low (one document successfully retrieved after an hour of use for all except user E). Even the most successful user, E, commented at the end of the interaction that:

> ...I don't think I could cope with doing anything more. I don't feel that I've really found much, apart from on calendars, which I could focus. I haven't found very much that was actually very helpful.

The downloading difficulty of user A can be understood in terms of poor "common ground" between user and system – that is: the user did not fully understand the state of the system, or the meaning of the warning message, and consequently could not find a way around her difficulty.

The errors experienced by users C and D can be understood in similar terms: that the event communicating the fault to the user is insufficiently expressive for the user to comprehend and respond appropriately (see Figure 2). Accepting that occasional errors such as network faults are unavoidable, the challenge to the digital library designer is to communicate the status of the system (which may include multiple servers) effectively to the user.

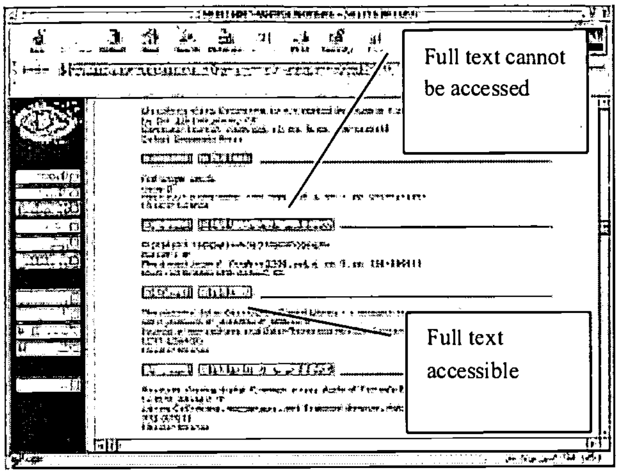


Figure 2: users did not understand the difference between the two sources of full text

Gould [10] discusses the importance of reliability when considering usability: a computer system that cannot be relied on to work predictably well is difficult to learn and use. Digital libraries accessed using standard web technologies are complex systems that are vulnerable to many kinds of failure. Some of the failures affecting basic accessibility of the library resources have already been discussed. Others have a more subtle effect on the interaction. For example, user A, aware that she could only access full text of journal papers (not conference proceedings) from ACM, conducted a search restricted to journals only. The following transcript shows the user's words, her [>actions] and the [<system feedback]:

> ... so I will search in "all journals and proceedings"...no, "journals only"...
> [> selects "journals only"]
> [> clicks on "search" button]
> [< ACM DL – search result page]
> ...I found 23 results... money in electronic... not that interesting. Electronic markets and intelligent systems.... This is interesting ...
> [...]
> ... Building bridges with practice... I will have a look at this
> [> clicks on the document link]
> [< article abstract page appears]
> ... this is very close to what I am interested in
> [> scrolls down the page]
> [< bottom of abstract article page displayed]
> [> clicks on the "full text" button]
> [< the pop-up window "User name and password" appears]
> Ow yaa. You have to get a username and the password because I haven't registered

Her confusion lasted some time longer. What appears to have happened here is that there was a temporary bug in the ACM library, or in the means by which the user's search specification was transmitted to the search engine, such that the search was not restricted as specified. Consequently, many of the results returned were articles from conference proceedings, but the user did not check this, believing that she had restricted her search to eliminate such results. A few days later, when we tried to reproduce the interaction as recorded on videotape, the problem had been corrected. Accepting that such errors are not always avoidable, it should be possible, as discussed above, to give more focused feedback, so that the user can understand and respond to the error in an informed way. In this particular case, the user was restricting her search because she was aware that she had authorisation (based on her host organisation's subscription) to access full text of journal articles, but not conference papers. Arguably, the system also has access to this information, and might have been designed so that the user's access rights were clearly indicated before she tried to download an article.

As well as failures, the user's choice of libraries is clearly affected by the quality of their experience with a library, and also their sense of familiarity with that library. Within the study reported here, the three dominant factors that determined perceived quality were: a sense of making progress (whether in finding relevant articles or in understanding a library better); a sense of the task being manageable (in particular, not an overwhelming number of alternatives with poor discriminability); and system response time being acceptable. We consider each of these aspects in turn.

1.10.1 A sense of progress

All participants had the experience of issuing search requests that returned no matches. Particularly for user C, this became very frustrating. E.g.:

> ...I searched this author in several digital libraries and always cannot get anything ... this professor has published several papers in computer journals.

And:

> This haven't found anything about "GCS" too.

Where users have a sense that they are learning from their null results, they still seem to have some sense of progress. E.g. user A felt able to modify her search:

> [> clicks on the "search" button]
> [< "search result" page appears]
> ...no matches... I haven't found any matches...Perhaps I shouldn't put trans... publications.,
> [> clicks on the browser "back" button]
> [< ACM DL – search page]
> ... so I will search in "all journals and proceedings"...no, "journals only"...

Similarly, user E modified her search, but was still unsuccessful:

> [< search results page replaces search formulation page]
> No matches. Great.
> [> clicks back]
> [< search page replaces search results page]
> So even on full text. Of course, its possible, no.
> What happens if I turn off the human and try again?
> [> removes human from subject search terms]
> [< subject search is now "artificial intelligence" (note – author search is still "Hollnagel")]
> I would expect to get a fair amount.
> [> clicks search]
> [< search results replace search formulation page]
> No matches. Ha ha ha.

In this case, the source of the problem was that this search was conducted some time after an earlier one that has specified a particular author, but the user had not noticed that there was an author entry in the search specification. From the user's perspective this was a new context, a new query; from the system's perspective, this was an elaborate query (with no matches). Improved feedback – helping the user to understand exactly what the search results referred to – might well ease such situations and improve the user's familiarity with the system.

Conversely, users appeared to have a particularly strong sense of making progress when serendipity worked for them – that is, when they came across an interesting item that was unexpected in the current context. This happened twice to user E; e.g.:

> [- ACM DL list of papers in most recent DIS conference]
> Triangulation.
> That actually looks quite interesting.
> [> clicks on paper's link]
> [...]
> So I'll have a look at that one.
> [> clicks acrobat print button]

It also happened to user D:

> I'm looking into a visual texture related papers but I actually found something that doesn't seem to be related to that at all. It's about comparing two different methods of evaluations: empirical testing and walkthrough methods which is also interesting to me so... I'll just jump to that from the search .

While serendipity is difficult to design for (by definition), it can be supported through discriminability: it is important that it is obvious to a user when such items come into view – that the descriptions of items make their nature clear.

1.10.2 A sense of the task being manageable

While no matches is clearly not a desirable result, there has been little discussion of the effects of too many results. Within the Information Retrieval community, one focus of research, and one criterion by which the quality of a search engine is assessed, is the quality of search results returned, as measured against suitable metrics. In practice, the users in our study appeared to place more store by quantity, or at least discriminability, than quality. For example we have the following comments from users in response to various search results:

User A (having specified "some" of the search terms to be used in a search for "electronic commerce"):

> Here I found one thousand seven hundred twenty two commerce and two thousand...
> ...OK, I will have to go back and say all
> [> clicks on "back" button]

User C commented:

> Sometimes the search finds too many articles so it is a little bit boring to read all these article titles.

More encouragingly, user E was pleased when fewer results were returned:

> 16 documents, that not too bad.

As was user B:

> ...only seven, that's fine...

Repeatedly, users commented positively when the number of results was small (less than twenty) and negatively when it was large (typically over two hundred). We can understand this in terms of the discriminability of events and the number of alternatives: with a large number of results, it is generally not possible to discriminate between those possible future interactions that are likely to be successful and those that are blind alleys (relative to the user's objectives).

1.10.3 Response time: the pace of the interaction

Due to factors such as network bandwidth limitations, the geographical locations of servers, network loading, etc., the response times of libraries were very variable in this study. Dix [8] discusses the importance of the *pace* of the interaction being appropriate to the task properties. Two of the users in this study (B and E) responded to the pace being too slow by opening a second browser window and interleaving interaction with the two windows. (D also opened multiple windows, but only used them sequentially.) Both B and E commented on download time. For example, user E noted that

> We're going to work much more efficiently if we have a new navigator window...
> [> clicks browser file -> new navigator window]
> [< New window (2nd window) appears with home page]
> ... and have that one running in the background.

As switching between windows, B commented that

> ...it's taking quite a long time to download [...] ...so I can continue working with ACM...

C also reacted to the response time of certain servers, but responded in a different way – by avoiding those libraries completely:

> [> clicks on IDEAL link]
> It seems to be very slow. So I will stop and try another.
> [> clicks on browser "Stop" button]
> [> selects browser "bookmarks" folder]
> [< list of links displayed]
> [> selects ACM DL from the bookmarks]

While the cognitive demands of multitasking in this environment are not excessive (because neither task is time-critical), the similarity of the two tasks is liable to cause interference. Both B and E appeared to lose track of the state of the second window at times.

1.10.4 Familiarity

As noted above, user E made her own selection of libraries; for the other four users, information was intentionally re-ordered, both on the bookmarks list and on the accompanying paper documentation, so that users who followed the list order of libraries would naturally start with different libraries. Nevertheless, four of the five users started by working with the ACM library. Two (D and E) stated explicitly that this was because they were familiar with that library. For example, D commented:

> ...I've used this library in the past so more or less I know that I can find relevant papers ...what I'm looking for.

B commented about a library that he had not used before:

> ...let's try this one: EBSCO on line. I don't know it but it's always good to know new things.
> [...]
> ...what shall I do. ... I am trying to find out how should I operate this... favorite journals, let's try it.
> [> clicks on "Favorite journals" link]

[<" Manage Favorites" pop-up window appears on thc top of "EBSCO page"]

...I am trying to understand ...how to operate this...

These two examples illustrate familiarity (or lack of it) with different aspects of library use. D is concerned with the type of content (and his expectation of being able to locate relevant papers), while B is concerned with how to work with the library system and the navigation features it offers.

Other aspects include familiarity with content of collections or articles – type (e.g. the kinds of papers that are published in a particular journal), structure and detailed content. For each aspect of familiarity, the design challenge is to support recognition (e.g. E, reading the title of a paper: "I happen to know what that work is") and incremental learning. For example, user E had difficulty understanding the "binders" feature in the ACM library:

[- ACM DL search results page]

I don't know what happens when I tick them.

[...]

I've clicked these things now and I haven't got a clue what I'm meant to do with them. Hmm. I don't know what a binder is...

[> clicks and holds on binder link]

[< browser pops up menu of possible actions]

... help...

[> releases mouse]

[< page is replaced with page containing list of articles selected]

... please choose a binder from your bookshelf.

I don't have a binder, and I don't have a bookshelf as far as I'm aware.

One of the substantive challenges in designing usable, useful digital library systems is enabling users to learn features incrementally, so that they are not overwhelmed at the outset by a large number of alternative actions that are indiscriminable (because the user has no way of predicting the effects of any actions). This would include distinguishing between core and discretionary features (e.g. document management features can be learned after document retrieval features), and presenting information *in context* – at the time it is needed – as well as making novel features self-explanatory as far as possible.

1.11 Using multiple libraries

The main purpose of this study has been to focus on use in context. Many aspects of context have been discussed above while considering use of individual libraries and deterrent factors. In this section we consider additional difficulties that arise because libraries are accessed by users from a particular organisation, using particular technologies that interact in ways that may not have been anticipated.

Users were all working within a University setting, using a PC running the Netscape browser to access library resources. The setting and the task were designed to be as natural as possible. Two main classes of difficulty emerged in the sessions: understanding the access rights they had based on organisational subscriptions and working across boundaries (between libraries, other software, other systems). We consider each of these aspects separately.

1.11.1 Access rights

As described above, user A had difficulty downloading papers from ACM because she believed she had limited her search to those items to which she had access rights, but others were listed in the search results. User B also tried to download conference papers, having ignored the details on the user instructions. Similarly, as described above, users had difficulty understanding the difference between papers that could be downloaded via Ingenta, and papers that could not be downloaded because they were only available via Science Direct. A further potential difficulty of the same kind resides in the IDEAL library, where users have access rights only to journals from one publisher, but not others. These distinctions are poorly understood by users and inadequately explained within the libraries. Also, in the current dynamic environment where new resources are coming on-stream rapidly and spending decisions (e.g. on subscriptions) need to be reviewed frequently, users have great difficulty keeping track of what resources are available to them. Particularly as users are often working in various locations, without easy access to the information about permissions settings, that information needs to be made easily accessible at the point of need. Of the five libraries used in this study, one (IDEAL) has recently added a feature to give any user individualised information on their access rights, which is clearly recognition of the difficulty and a first step towards addressing this particular problem, which is just one aspect of the broader challenge of enabling user and library system to establish common ground.

1.11.2 Working across boundaries

Within this study, there were many cases of users traversing boundaries – between collections such as Ingenta and Science Direct, as discussed above, or between NZDL and various ftp sites, as exemplified in user A's interaction:

[< CSTR – browse result page displayed]

[...]

[< top of the page displayed]

...let's try the first one

[> clicks on the first link]

[<ftp://actor.cs.vt.edu page appears]

... so I can't find easily what I am looking for.

Such transitions, which are often not well marked for the user, demand that the user switch to a new way of working, with different navigation and other features, in an environment that looks and behaves suddenly differently. While transitions may be necessary, even desirable in some circumstances, their implications for users need to be better understood.

Because libraries are commonly made available via the world wide web, there are many cases where users can accidentally leave the library, following links to other web based resources. User C experienced this when following a link from Emerald that referred to a book: further information about the book could only be accessed via internet book stores:

[> presses "Review" button]

[< "review article" page displayed]

[> scrolls down the page]

[< bottom page revealed]

[> scrolls up the page]

[< top page displayed]

I'm trying the book.

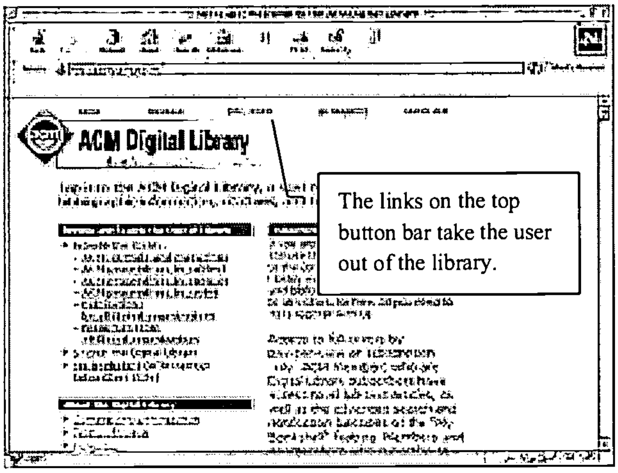
[> scrolls down the page]

[< bottom page displayed]

[> clicks on "order the book"] [amazon]
[< "amazon" page displayed]

User A also accidentally left a library – twice – as she searched for library resources. The first time was almost immediately after she had followed the ACM link from the bookmarks provided:

...what's new. Maybe if this is on the top then it will be more convenient for me.
[> clicks on the "search acm" button]
[< ACM search page displayed]
...are not included. What is not included?
[> types "the best of electronic commerce" into search box]
...it's a key word? Yes.
[< "key word" pre- selected]
[> clicks on "search" button]
[< "Search Result" page appears]
I have 83 pages matching my query and I have here 1 to 12
The score here is showing what is most related to my search
Strategic analysis reports...social impact...
[> scrolls down the page]
[< more results appear on the page]
... I will have a look at the "strategic analysis reports"



The links on the top button bar take the user out of the library.

Figure 3: identifying the exit routes from a library

She is unaware of the fact that she is not using the digital library, but searching the ACM web site. This is the behaviour of a novice user who has just entered the ACM digital library for the first time, but she has clearly failed to discriminate between "search ACM" and "search the Digital Library". This is partly an artifact of the study, since she entered the library directly from "outside" without passing through the ACM web site, but such situations are common, particularly with the increasing provision of digital library portals that provide links to various libraries from one web site.

Later in the interaction, she again left the library – rather more briefly – as she tried to return to the digital library home page:

[> ACM DL search page displayed]
[> click on "home" button]
[< ACM – home page]
So, I would like to find some journals, articles, magazines...
... So where can I find them...
[> clicks on "digital library" link]
[< ACM digital library page]

Figure 3 illustrates the source of these interactional detours: the user has failed to discriminate between the type of links on the top links bar (which are generic ACM links) and those on the body of the page, which are digital library links.

As well as boundaries with other web based resources, users also have to work across boundaries with other software and the operating system. For example, user C had some difficulty saving a file, shifting attention from library use to working with the operating system. User B accidentally closed his browser session, and had to start again:

[< further pages displayed]
...not interesting...
[> closes the Netscape window]
[< "word" document page displayed]
[> closes the Microsoft "word" application]
[< pop-up window "Microsoft Word" appears "Do you want to save changes in the document"]
[> clicks "no" button]
[< computer desktop appears]
...I got out of the "Netscape"
[> clicks on "Netscape" icon]
[< "Netscape" window opens]
...bookmarks...now again..

As the only user in this study who printed files, rather than saving them, user E also had to integrate library work with using other resources – notably a printer:

It's all tied up for the minute, so I'll go and collect some things from the printer.
[- ACM DL 1st window showing ontology paper]
OK. So I don't know how I will know if I've got everything I've printed out. Because I'm losing track of it completely. But I think that's been sent and not needed anymore.

Such boundaries cannot be easily eliminated, but the consequences of making transitions across boundaries can be minimised; for example, user B had difficulty restoring the state before he closed the browser, while user E had difficulty establishing exactly what state the system was in (including what had and had not been printed or fully downloaded): common ground needs to be restored.

CONCLUSIONS

We have considered deterrents to the use of digital libraries, and usability issues that are raised by the pragmatics of use in a natural setting – taking account of organisational and other concerns. One way of considering usability is as the absence of any undesirable features of the interaction; therefore, the focus in this paper has been mainly on difficulties experienced rather than successes, although all users achieved some success in their interactions. If some libraries have been discussed more than others, it is largely because those libraries were used more extensively – a consequence of use being discretionary.

In terms of the Interaction Framework concepts introduced earlier, the most important design issues have been found in this study to relate to:

1. **Familiarity**: users need to be able to rapidly acquire understanding of core library features, content and

structures, if they are to gain a sense of progress while working with the library.

2. **Blind alleys**: many interaction sequences did not achieve the user's objectives. This is most obvious when a search returns "no matches", but occurs in some more extended interactions. Some interactions that have no material outcome achieve improved familiarisation (the user learns more about the library structure or contents), but others do not. Perhaps more surprisingly, some interactions that resulted in a large number of hits failed to achieve the user's objectives because the user was (apparently) overwhelmed by choice and retreated from the search results page, due to poor discriminability.
3. **Discriminability**: forming understandings of the content and possibilities in a collection relies on being able to discriminate between possibilities. Therefore we need to ensure that the potential events are easily discriminated.
4. **Serendipity** – finding unexpected interesting results – seems to give users a particular sense of making progress. Serendipity depends on users being easily able to identify interesting information, which is one aspect of discriminability.
5. **Working across boundaries**: transition events, where one agent (user or computer system) changes context, can often be a source of interactional difficulties; measures need to be taken to ensure agents maintain common ground and understand the consequences of transitions.

This study has identified many cases where decisions taken by computer scientists and librarians have had unanticipated consequences. The incremental approach to delivering library resources has many positive advantages (e.g. ease of introducing new features) but can also result in inconsistency as users move between libraries, particularly where that move is made so smooth that users are often unaware it has even happened. The use of site licenses that permit partial access to resources, or access that is temporary, makes systems unpredictable to users, so that they cannot develop an adequate (usable) understanding of what detailed goals are and are not achievable. If theoretical possibilities are to become practical solutions, then the kinds of pragmatic issues raised by this study need to be taken into account in the design of future libraries.

ACKNOWLEDGEMENTS

This work is funded by EPSRC Grant GR/M81748. We are grateful to the anonymous participants who took part in this study, and to George Buchanan and anonymous reviewers for comments on earlier versions of this paper.

REFERENCES

[1] Adler, A., Gujar, A., Harrison, B. L., O'Hara, K., & Sellen, A. (1998). A Diary Study of Work-Related Reading: Design Implications for Digital Reading Devices. In Proceedings of CHI '98, pp. 241-248.

[2] Bishop, A. P. (1998). Digital Libraries and Knowledge Disaggregation: The Use of Journal Article Components. In Proceedings of ACM DL '98, pp. 29-39.

[3] Bishop, A. P. (1999). Making Digital Libraries Go: Comparing Use Across Genres. In Proceedings of ACM DL '99, pp. 94-103.

[4] Blandford, A. E., Harrison, M. D. & Barnard, P. J. (1995) Using Interaction Framework to guide the design of interactive systems. International Journal of Human-Computer Studies, 43, 101-130.

[5] Carroll, J.M. and Carrithers, C. (1984). Blocking learner errors in a training wheels system. Human Factors, **26,** 377-389.

[6] Clark, H.H. and Brennan, S.E. (1991). Grounding in Communication. 127-149 In Resnick, L.B., Levine, J and Behrend, S.D. (Eds.) Perspectives on Socially Shared Cognition. Washington DC.: APA.

[7] Covi, L. & Kling, R. (1997). Organisational Dimensions of Effective Digital Library Use: Closed Rational and Open Natural Systems Model. In Kiesler, S. Culture of the Internet, Lawrence Erlbaum Associates, New Jersey. pp. 343-360

[8] Dix, A. J. (1992). Pace and Interaction. In A. Monk., D. Diaper and M. Harrison, Eds. People and Computers VII, Cambridge: CUP.

[9] Furnas, G. W., & Rauch, S. J. (1998). Considerations for Information Environments and the NaviQue Workspace. Proceedings of ACM DL '98, pp. 79-88.

[10] Gould, J. D. (1988). How to design usable systems. In M. Helander (ed.) Handbook of Human-Computer Interaction, pp. 757-789. Elsevier : Amsterdam.

[11] Gutwin, C., Paynter, G. Witten, I., Nevill-Manning, C., & Frank, E. (1999). Improving Browsing in Digital Libraries with Keyphrase Indexes. Journal of Decision Support Systems, 27(1-2), 81-104.

[12] Jones, S., McInnes, S., & Staveley, M. (1999). A Graphical User Interface For Boolean Query Specification. International Journal on Digital Libraries, Volume 2 Issue 2/3, pp 207-223

[13] Marshall, C. C., Price, M. N., Golovchinsky, G., & Schilit, B. N. (1999). Introducing a Digital Library Reading Appliance Into a Reading Group. In Proceedings of ACM DL '99, pp. 77-84.

[14] O'Day, V. L., & Jeffries, R. (1993). Orienteering in an Information Landscape: How Information Seekers Get From Here to There. In Proceedings of InterCHI '93, pp. 438-445.

[15] O'Hara, K. O., Smith, F., Newman, W., & Sellen, A. (1998). Student Readers' Use of Library Documents: Implications for Library Technologies. In Proceedings of CHI '98, pp. 233-240.

[16] Park, S. (1999). User Preferences When Searching Individual and Integrated Full-text Databases. In Proceedings of ACM DL '99, pp. 195-203.

[17] Theng, Y.L., Duncker, E., Mohd Nasir, N., Buchanan, G. & Thimbleby, H. (1999), Design guidelines and user-centred digital libraries. In Abiteboul, S. & Vercoustre, A. (Eds.), Proc. ECDL'99, 167 – 183.

An Ethnographic Study of Technical Support Workers: Why We Didn't Build a Tech Support Digital Library

Sally Jo Cunningham
Department of Computer Science
University of Waikato
Hamilton, New Zealand
+64 -7-838-4402
sallyjo@cs.waikato.ac.nz

Chris Knowles
ITS Campus Media
University of Waikato
Hamilton, New Zealand
+64 -7-838-4466 x7714
chrisk@waikato.ac.nz

Nina Reeves
School of Multimedia and Computing
Cheltenham and Gloucestershire
College of Higher Education
Cheltenham, Gloucestershire, UK
+44 -1242 -543236
nreeves@chelt.ac.uk

ABSTRACT

In this paper we describe the results of an ethnographic study of the information behaviourss of university technical support workers and their information needs. The study looked at how the group identified, located and used information from a variety of sources to solve problems arising in the course of their work. The results of the investigation are discussed in the context of the feasibility of developing a potential information base that could be used by all members of the group. Whilst a number of their requirements would easily be fulfilled by the use of a digital library, other requirements would not. The paper illustrates the limitations of a digital library with respect to the information behaviourss of this group of subjects and focuses on why a digital library would not appear to be the ideal support tool for their work.

Categories and Subject Descriptors

H.3.7 Digital Libraries: User issues; D.2.1 Requirements/Specifications: Elicitation methods

General Terms

Design, Human Factors

Keywords

Ethnography, user studies, requirements analysis

1. INTRODUCTION

There have been many studies in information science looking at the nature and frequency of information seeking activities by different groups. Many have focused on the differences in how people search, what they search for, and why. They have also identified how people use the information they have obtained and how the original source of this information has been used on subsequent occasions over the longer term. Today, there is a plethora of information sources available in both electronic and paper-based forms to a wide range of user groups. One of the most promising areas for developing information sources to meet the needs of a variety of user groups has been in the construction and accessibility of digital libraries.

Quantitative studies of usage patterns in existing libraries (digital or physical) have, on occasion, contained indications of a less-than-perfect fit between users and libraries; for example, Cunningham and Mahoui [6] note that only 28% of visitors to two computer science digital library make a second visit to those sites. Quantitative studies, such as transaction log analysis, typically give detailed pictures of user actions, but can give little or no insight into users' motivations and needs. For example, we know that the majority of the computer science digital libraries' users do not return to these web sites—but why not? Is it because the users' information needs were perfectly satisfied by a single visit, or because the collections' contents are inadequate, or perhaps because the user interfaces are unacceptable?

Often usability studies, whether quantitative or qualitative, are limited to evaluating the interface to an existing digital library or the ease with which a given user group can find the information in the library. Many of the studies deal with academics or other specific user groups carrying out research activities. The studies do not, however, necessarily evaluate the basic concept of a digital library as a suitable information tool in comparison to other information tools for people who provide a technical support function for those researchers, e.g. a digital library may be constructed for a group of researchers but is developed or maintained by a technical support group. In the case of computer science the researchers may use the library but the specialist computer support staff who themselves 'work' in the same area (computer science) may not actually choose to use the product.

Instead of looking at groups of potential users who have already been studied in detail we identified a specific group who have had little attention paid to their circumstances. This group appear to be a potentially rich source of data for looking at information behaviourss in terms of the types of tasks they perform, the tools they use, their knowledge base, how they interact with one another, etc. The study reported here concerns a group of consultants who provide technical support in a School consisting of computer science, mathematics and statistics departments.

The data gathered during the study was from six members of the group. As the entire technical support group comprised eight members, the number taking part in this study was, we felt, more than satisfactory to get to grips with the essential elements of its operation and behaviours with regard to information seeking and

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

use. The size of a subject group is an issue in any study, but even more so when dealing with specialist groups. Such groups are, by their very nature, different to 'general' groupings. They have a more defined focus, whether that be in terms of the work they carry out, the type of people they are and so on. The size of this group and its uniqueness within the institution does not make them less worthy of investigation, indeed it makes them more attractive as a focus of study. In an institution of around seven hundred staff this support group is unique. There may be similar related groups in other institutions that collectively would form a more significant subject base to study. Whilst an increased subject base may raise some underlying level of reliability in the results, the greater influences of the environment (e.g. the nature of the institution, the country, the profile of the department being supported etc.) may negate this. In such instances, the reasons for the variation would also be of great interest and of relevance for those designing information resources such as digital libraries.

In looking at the value of digital libraries it cannot be desirable that only groups with a large membership be studied in any depth. Such a focus would surely result in an architecture that was limited by the characteristics of those groups? Specialist groups provide different challenges to the norm which may be more useful in identifying different behaviours that should be identified. Levy and Marshall [12] assert that the "highest priority of a library, digital or otherwise, is to serve the research needs of its constituents." Different users within that constituency focus on different elements of resources and those developing and supporting digital libraries may tend to "idealize users and uses, projecting or inventing an incomplete or even inaccurate picture of the real work being done." (p 80)

Their recommendation to avoid this is to adopt a work-oriented perspective and instead look at the users, the work they do and how that work is supported through the use of the technologies and documents. To achieve this they suggest the use of ethnographic techniques of observation. This is exactly what this study did and the results demonstrate how valuable such an approach is in terms of the richness of the results. Identifying differences between different user groups will, hopefully, serve to stimulate ideas about the nature of digital libraries and how they might be used.

In considering the depth and breadth of the different facets of the operation of this group we gave a great deal of thought to how we should go about identifying the information behaviourss of this group. We chose to adopt an ethnographic approach to this research.

Ethnographic methods are being used more and more by researchers and practitioners in human-computer interaction studies and in systems analysis and design in an attempt to become more aware of the work circumstances, personal and social characteristics and knowledge base of potential users of proposed systems [4]. Ethnographic techniques are also used to identify the relevance of different information systems to those people and their roles in an organization [17].

The frequently cited usability maxim of 'know thy users' is seen as a critical part of designing any information system. The ethnographic approach appears to provide an ideal opportunity for us to get to know this group in depth and to identify the various types of information behaviourss they exhibit. Technically oriented investigations of work systems are often reductionist and miss important aspects of work activities [3] [4] [18]. For instance, support staff may often continue to problem solve outside the formal workplace in the tea room, pub or via web access at home. The shortcomings of the reductionist approaches are due, in part, to the 'outsider' perspective of the researcher who, having observed the work processes, then attempts to describe them using their own models and vocabulary. The latter may have little meaning for those participating directly in the work and often focus on those aspects that are most relevant to the researcher's preferred model and solutions, for example the creation of a digital library.

The goal of this ethnographic study was firstly to discover the types of information that these technical consultants used in their work; then to understand how they individually and as a group gather and use information. Although adopting the basic principles behind ethnographic forms of investigations was seen as an interesting and potentially valuable way to look at the situation, we acknowledged at the outset that there might be limitations to the investigation, especially in how we might conduct it or draw conclusions from it. For a particularly thoughtful examination of ethnographic techniques in the context of and information science, see [4]. The aim of the study was not use the results to build an information system to support this group, but it did include some element of identifying the requirements such a system might have to meet and the constraints under which such a system might operate. In this sense, the idea of the digital library might be an option to be considered, should the study identify information behaviourss that would be supported by such a tool.

This paper is organized as follows: Section two describes the methodology used in this ethnographic study: the consulting group is described, the ethnographic techniques employed are listed, and the consultants' range of work activities are described. We then briefly discuss Greenstone, the digital library construction software developed by the New Zealand Digital Library Research Group (http://www.nzdl.org). Greenstone shares many characteristics of other examples of the current crop of digital library architectures; its assumptions about collection design, user interface, and user characteristics are discussed in Section 3. Section 4 then examines the information behaviours observed in the ethnographic study; as indicated by the title of this paper, the information gathering and usage behaviourss of the consulting group were not well-suited to support through a Greenstone-like digital library. Section 5 presents our conclusions.

2. Methodology

The authors conducted an ethnographic study of six members of a technical support group (TSG) serving the School of Computing, Mathematics, and Statistics at the University of Waikato.

To those whose expertise is in quantitative empirical methods, six participants may appear rather few but the aim of naturalistic ethnographic techniques is to discover a rich picture by developing an 'appreciative stance' via the researcher's involvement in the setting. The number of participants is therefore not as important as depth of the enquiry.

The difficulties of interdisciplinary working between ethnographers and computer scientists, in terms of language and natural attitude, have been summarized by Crabtree *et al.*, [4]. The approach in this study has been to allow the researcher to view technical knowledge as a socially distributed resource that is often stored primarily through an oral culture [7]. The technical consultants' war stories therefore become texts for both the ethnographer and the consultants themselves. Thus, data gathering techniques included: semi-structured interviews of participants; 'shadowing' participants as they worked; observation of semi-social discussions in the School tearoom; and examination of various work artifacts (email, bookmarks, webpages, office bulletin boards, etc.).

The data gathering phase of this study ran from May – August 1999.

2.1 Participant Demographics and Description of Consultant Activities

There were six participants in the study. All were male, ranging in age from the mid-twenties to mid-thirties. All have formal tertiary qualifications. Four of the six have bachelor's degrees in computing, and two have Ph.Ds in physics. The group support a range of users of computer systems within the School such as, lecturing staff, administrative and research staff, and university students at undergraduate and postgraduate level. Some members of the group were students at the University, but all have been students, and have worked closely with academic staff both as students and currently as technical support consultants. The group is, therefore, seen as having a special position as a technical support group in the School as well as individually being seen as colleagues by a large number of the staff.

Two members of the group have special roles: the group leader, who holds a managerial position and also provides Unix support; and the undergraduate lab support consultant, who provides primary maintenance for the printers, hardware, and software in the large teaching labs. In terms of physical proximity, the undergraduate lab support consultant has his office in a separate building from the other five consultants, who have offices adjacent to each other in the ground floor of the School of Computing, Mathematics, and Statistics. The separation of the group into two areas has some implications for the degree of 'personal' contact between the technical support staff and the people they support, in terms of 'foot traffic' going past their offices and in terms of the means of contact used (e.g. phone, email, etc.).

Each technical consultant is seen as having a primary area of expertise. Often this is in terms of operating systems (e.g., Windows, Unix, Macintosh, etc.) A consultant may also be identified by responsibility (for example, one consultant is in charge of the first year labs). The group is seen as a central resource for the whole School rather than for individual departments. This means a member of the group tends to be called upon for reasons related to their area of expertise than some organizational role.

The consultants' work is mainly task-based. The nature of the job involves dealing with both poorly and well-defined tasks. In consequence, the consultants often have to employ a range of information seeking techniques. Often projects are relatively flexible in terms of how they may be approached; frequently other members of the group are recruited to deal with specific elements as required, with one person being responsible for the overall project. It is difficult to describe a typical day of a consultant in this group, but depending upon the time of year he may have he might expect to:

- deal with immediate, low-level problems such as fixing a stopped printer queue;
- interact with novice users to answer questions about standard software packages;
- set up new facilities (ranging from complete installation of a 50+ computer lab, down to setting up a single new laptop for an academic);
- proactively investigate potential software and/or hardware problems, and locate solutions to these problems that haven't occurred—yet;
- keep an eye on long term developments in his area of hardware/software expertise, so that he can provide informed advice to decision-makers in the School;
- update information sources used by themselves and their 'customers' in the School; etc.

For some of these tasks there is a level of repetition such that once the initial information seeking activity is completed the results can be applied and re-applied when the situation arises again. Some of the tasks, however, fall into the category of 'one-offs', with the consequence that the problem solving and information seeking behaviourss result in a solution that cannot substantially be used again. It is clear, however, that most of the tasks, despite the level of repetition, share many of the characteristics of information seeking and retrieving activities described in information behaviours studies. In an earlier work [9] we interpreted these activities in terms of a specific information behaviours framework (that of Ellis [7]). In this paper, we examine the potential for supporting these activities through a digital library tailored to the TSG's needs and preferred search behaviourss.

3. THE GREENSTONE DIGITAL LIBRARY ARCHITECTURE

Greenstone (http://www.nzdl.org) is a toolkit for constructing and maintaining digital libraries. A digital library is viewed as a set of collections, where each collection has a focus—typically by type of document (for example, music videos), or by subject (for example, computer science research documents), or by user needs (for example, people working in disaster relief).

Collections can be composed of documents held locally, can be an index to geographically distributed documents, or can be a mix of local and offsite documents. It is expected that a collection will be constructed after a set of documents grows past the point at which a linear search of the set is feasible. A collection, then, is expected to be large: hundreds, thousands, or millions of documents.

Current digital library architectures—of which Greenstone is typical—make a number of assumptions about the documents and users of a collection:

- The primary interaction mode is presumed to be searching, rather than browsing. The collection creator can organize the Greenstone search interface into 'simple search' and 'advanced search' modes. The advanced search mode offers a more extensive set of options for tailoring a search, but at the expense of requiring the user to know more about search strategies, the collection's metadata, and the system's implementation.
- Browsing facilities are relatively crude: the collection builder can specify simple categorizations and sortings of documents (grouped and sorted by author's last name, for example). These browsing facilities cannot be defined in an *ad hoc* manner by users. This situation is common across digital library implementations; while a number of novel browsing

schemes have been prototyped (see, for example, [13]), they are not standard with digital library construction software.

- A document collection can be expected to grow monotonically, with documents remaining in the collection (and assumed to be potentially useful) for the foreseeable future. Discussions in the digital libraries research literature on collection culling have tended to focus on the need to conserve space or to observe memory limitations, rather than on detecting documents whose contents have become obsolete.
- The contents of a collection are relatively static. Groups of documents are added to the collection periodically, rather than having individual documents added in a steady stream, in 'real time'. This limitation arises because rebuilding the collection's index is an expensive operation, and incremental index construction is not a common feature of digital library software. Typically, then, there is a (sometimes significant) gap in time between a document's production and its availability through a digital library.
- Documents can be presumed to be trustworthy, or can be evaluated for trustworthiness solely on their contents. The collection builder is generally, though not always, expected to serve as gatekeeper for the digital library. This role may be served by selecting additions to the collection on a document-by-document basis, or by exercising quality control through choice of reputable document sources from which additions the collection are automatically drawn.
- Documents stand alone, in the sense that a document is viewed with little or no regard to its relationship to other documents. For example, it may be difficult or impossible to view relationships such as a sequential publishing of two documents, or because the documents come from the same source, or because they are distributed by the same mailing list, etc.).
- Documents, once released or published, do not change. Documents are not usually monitored for changes to their contents or metadata, which can mean that the document's description in the collection may be inaccurate if the document is altered.
- The primary scenario for locating information involves a solo user searching a collection of documents. Little, if any, support is provided for collaborative information behaviours.

4. TSG INFORMATION BEHAVIOURS

The properties of the Greenstone Digital Library Architecture offer the potential user all the advantages of a typical digital library. Given the type of activities that the technical support group carry out as part of their work and the associated information seeking behaviourss it is tempting to assume that a digital library would be the most obvious tool for consolidating their information sources. Members of the group share resources, compare information from a number of different sources, keep records of information they have found and search for information often, some times for a specific reason and sometimes for general information gathering to develop specific skills and knowledge in an area. All of these activities would be supported through the architecture of the Greenstone Digital Library. However, the value of adopting an ethnographic approach to the investigation of the group's information seeking and information use, in preference, to other, more structured design based techniques, becomes apparent when looking at how well a digital library would really serve the group's needs.

4.1 Formally Published Documents Usually Aren't Useful

'standard' academic resources such as journals, conference proceedings, bibliographic indexes, etc. were not used by members of the group for work, as their work-related tasks were not seen as 'research'. Interestingly, this was as true of the consultant seconded to an academic research group; he provided programming support, but did not follow the 'research' side of this work.

Popular IT magazines such as MacWorld were irregularly consulted, primarily for pricing information in advertisements. Occasionally the local version of ComputerWorld was read, mainly for New Zealand-specific news or for job adverts: *"I read ComputerWorld for a laugh, because they started sending it to me for no apparent reason. I always hear about non-local things on the web first, the only reason to read it* [ComputerWorld] *is to see what's happening locally. Everything in it is usually old news."* None of the magazines were viewed as core resources.

The printed documentation that accompanied hardware and software purchases in the School was also seen as irrelevant. All members of the TSG had bookshelves full of documentation that they rarely, if ever, consulted. Often the manuals would be saved for years in pristine condition before being thrown away, still shrink-wrapped:

Tom[1] points at his bookshelves *"I've got documentation for the stuff we've bought. It actually gets used less than you'd think, it rarely goes into sort of details I need for the problems I deal with. It's probably useful for tutors to explain software to students."*

This is not that surprising given most of this application documentation is intended for end users rather than technical consultants. For example, the manuals typically focus on how to carry out specific tasks. There is often a simple 'trouble shooting' section that outlines the most frequently encountered problems and how to solve them, but this is generally pitched at a relatively simple level, not necessarily assuming the user has detailed knowledge of the program or the operating system on which it is running. The technical support staff are generally called upon when more complicated or subtle bugs are encountered, and the TSG then require a different level or type of information to solve these problems.

4.2 Many Documents are Ephemeral...

One striking feature of the information sources used by the TSG consultants is that many are extremely ephemeral. The usefulness of a given document tends to be highly time-dependent; for example, they need the latest news about the latest bug found in the latest version of the operating system. The usefulness (and in that sense, the 'quality') of information tends to deteriorate substantially over time.

In many libraries having access to historical records, however, 'recent' that history may be, is often viewed as an advantage for research purposes, where people can go through the background

[1] Consultants' names have been altered for privacy reasons.

material in a subject and use it to study developments in the area. The emphasis on timeliness for the technical support group, however, would necessitate a culling of older documents as they lose relevancy. If this were not done then the older, obsolete, documents would numerically overwhelm the more timely pieces of information.

4.3 But Some Documents Hang Around Forever!

Paradoxically, however, some documents are valuable because they are obsolete! The TSG have to support several 'legacy machines', elderly computers that run long-outdated versions of software that are nonetheless still heavily used by some members of the School. It can be difficult to locate information on the specifications of old computers, or on bugs and fixes for obsolete software. This type of information might not be carried on current websites or other information sources, which tend to cull information on outdated machines; instead, if a spec or bug fix is needed, it is likely to be found on a carefully hoarded document in a consultant's private stash. For example, in trying to locate a part number for an eight year old laptop, the TSG member consulted a long outdated version of a cdrom called 'service Source', distributed to Apple technicians. The TSG member had spotted it in a trash heap in another workshop on campus and retrieved it for his own use, citing the cdrom as "a good starting point for antiques".

At the time of this study, a number of historic documents were held on paper—what one consultant dismissively referred to as *"just rubbish, it's filing cabinet stuff"*:

Chris asks about Tom's filing cabinet *"A lot of what goes in there I hope never to see again. Some stuff comes up yearly for staff, [other documents are] vendor product details that I didn't want to hear about in the first place, copies of everything purchased end up in it because sometimes you have to tell them try again, all the leave that anyone does, changes in salary, I file them forever."*

Chris asks about Tom's bookcase [Those are] *"the folders that Dave* [the former TSG manager] *passed on to me. I don't think I ever looked in them. He did them back in the days when paper really was the best way to pass things on. Now if I print something, it will go out of date, it will go out of date if I print them again. I don't think I've ever looked in those boxes."*

There is an expectation that much of this "stuff" will never be used—but it should be archived, just in case it: *"like somebody buys something and it comes with instructions and a little plastic whatsit, you give it to the punter and he loses it. I keep it and it turns out that no one ever needs it, but at least we've got it."*

Gradually much of this documentation was becoming available for storage electronically, rather than on paper. The university was distributing most of its forms as Word files, and many purchasing records were appearing in digital form. This type of record would be well-suited for inclusion in a digital library: once created, the records are unchanging, and many of the records can be easily categorized and cataloged. The ability to easily search for and retrieve a given document could provide a significant advantage over physical filing systems. It was not clear that it would be easy to locate a given document in the cabinets, boxes, and heaps then in use—or even to know whether or not a particular document had been stored!

4.4 Documents Might Not be Trustworthy

Consider one of the many digital libraries intended for computer science researchers. These collections generally contain conference papers, journal articles, and working papers written by members of the computing field. Many documents have undergone the scientific refereeing process, and others (such as the working papers) generally have the imprimatur of a recognized research institution. Digital library users assume that the contents of the documents have been verified or vouched for, and that the documents are, on the whole, trustworthy. Exceptions may arise, but they are expected to be confined to a small minority of the digital library's contents.

In contrast, much of the information that the TSG gather from websites and mailing lists hasn't been verified, and in fact may be expected to contain errors, half-truths, unsubstantiated advertising claims, and rumors. The TSG members recognize that they will often have to do extensive cross-checking to feel assured that the information is reliable. For example, one consultant was observed to regularly monitor the websites of major software producers for news on upcoming and existing products. This information would then be cross-checked with product reviews. The he validity of individual reviews would also then be cross-checked, depending on the source.

Sometimes the trustworthiness of a site or a particular document cannot be immediately evaluated. One consultant saves particularly interesting WWW articles on stickies on his desktop, and consults them occasionally to see how the document's contents hold up over time. *"As things come to pass on the stickies"* [that is, as the events or trends mentioned in the articles actually occur], the consultant can put more trust in the document and, by extension, its source website.

4.5 A Primary Information Source: Other People

Other members of the TSG are a primary, significant source of information. One common strategy for finding a solution to a problem is to ask another member of the group. These interactions are usually not formally recorded; when asked how communications were managed between members of the TSG, the TSG members clearly preferred immediate, personal contact with each other. The following comment was typical: *"For something important I go next door or ring, otherwise it's email. Ringing or going to see someone is the first thing for communication."* Close physical proximity was seen as a positive advantage in solving problems: *"We used to be in the same room, so we were just a shout away from getting answers. BUT one person's interruption was every person's interruption."*

TSG members also occasionally use each other as information filters:

Bob: *"Occasionally I see something interesting to Dave, or I pass something to someone else."*

Tom, a supervisor: *"There's a Linux kernel development mailing list but that's too high volume, I rely on the guys [to keep up with that]."*

A significant amount of 'passive' or serendipitous information gathering also occurs in face-to-face communication with colleagues. One consultant describes himself as frequently *"gossiping"* with other consultants on campus to find out activities or events that are relevant to his interests but that he's not directly working on. More formally, two of the consultants attended campus-wide informational meetings scheduled and run by the central computing service (irreverently referred to as "*prayer meetings*"). All TSG members attend a weekly local TSG group meeting: "*We send Tom our weeklies [a weekly report on their individual activities]. He goes through them and picks out things he thinks are important and we talk about it in the meetings. When we finish going through that then each person may bring up other stuff and we chat about them. We may talk just between 2 or 3 people in a meeting.* "

The TSG members rarely proactively announce solutions to problems to the group as a whole, possibly because the consultants tend to specialize, and most problems would be of interest only to one or two members of the group; "*Very infrequently someone will go, this is the solution to this problem and then go tell everyone.*"

It appears, then, that critical information gathered by TSG consultants is not digitally recorded, and is difficult to formalize for inclusion in a digital library. This is not an unusual situation, of course! Few digital libraries would claim to be a comprehensive source of information for their users. It is, however, important to be aware of the limitations of a proposed information source, particularly in noting the bounds for assembling a complete and comprehensive resource.

4.6 Local Production of DL Documents

Some of the documents used by the TSG (and which presumably should be included in a TSG digital library) must be constructed by the TSG themselves—who as a group aren't known for their love of documentation! These documents are mainly descriptions of the local system: lab configurations, local network descriptions, instructions for setting up new machines, etc.

These documents differ from the static, unchanging documents that typify the contents of most digital libraries. Locally produced documents require considerable maintenance—and are generally not complete or entirely up to date. The TSG recognize that developing documentation necessarily takes time away from other activities. One supervisor pointed out that, "*some of the NT guys got all keen and set up a web based thing to track their work and reports. More often the report is created from stuff scribbled in a diary. I'm more interested in what's coming up and what they'll have to do, than what they've spent their time doing.*"

This concern is not unique to the local group; for example, in an extensive case study of a work-planning process, Soloman [19] describes an information gathering and documentation process gone astray:

"The staff in the regional offices take time away from their technical assistance project activities to fill out project status forms and the study unit's staff takes time away from their program evaluation aims to maintain the project database. Projects fall behind schedule and the goal of evaluation to make things better is thwarted. The well-intentioned drive for accountability and improvement seems to have made things worse." (p. 1106) On the other hand, inadequate or outdated documentation can cause problems if older versions are used to base decisions on or to solve problems with. Keeping some level of version control also means keeping some record of who produced the documentation and where it is kept. At present, the TSG members attempt to find balance by creating 'just enough' documentation.

Internal documentation (intended primarily for use within the TSG) is by no means complete, but is generally sufficient to jog the memory of the documentation's creator, or to be used as a starting point for exploration by the other members of the TSG team. For example, a work diary—paper or electronic—could be referred to later to recall a problem and its solution.

External documentation—intended to more directly support the consultants' client base—appears to be developed in response to sustained demand from students or School academics. One supervisor notes, "*I'm a big fan of automation. Machines should be able to work by themselves.* [TSG] *Staff turnover is a problem. One of the biggest causes is repetitive work. I tell the guys if you have to tell someone something more than twice, set up a web page and give them the URL, if more people need it try to set up a program to do it automatically. We have developed automated things for costs, measuring web surfing traffic.*"

Some records of problems and solutions are created by enforcing, where possible, a preference for dealing with "the punters" through email. Email correspondence documents the consultant's activities for internal reports, and can be filed away for later use if the problem is likely to recur. One consultant notes that, "*I don't have voice mail. I don't see any point in voice mail when email is much better. The problem with voice mail is you can't file things, you don't have an accurate representation of the problem because you're coping with voice as well as trying to keep your facts straight. With email you can read it over first. And I try to keep things filed with email. I use... what do you call them, folders I guess.*"

In examining the problem of including locally produced documents in a digital library, another point to consider is that the construction of some documents may not be in the individual TSG consultant's best interest (although maintenance and development of these docs may be in the best interest of the group as a whole, and of the university). Chatman [2] identifies *secrecy* as a strategy sometimes employed to give an individual greater control or influence in the communication process. Chatman's study concentrated on various 'outsiders' characterized as members of the 'information poor'—for example, women involved engaged in job training with the CETA (Comprehensive Employment and Training Act) program. These women did not share information about opportunities for permanent jobs, as letting others know about the positions would reduce the probability that they themselves were offered a coveted permanent position.

It is important to note that we did not observe any member of the technical support group consciously employing the secrecy strategy. However, there are obviously opportunities for a technical consultant to create job security by becoming the only person with critical (and undocumented) information about the structure and function of a given system. For example, a supervisor jokingly noted that one consultant had exclusive knowledge of a particular system setup: "*if he leaves, then we will never print again!*" Another supervisor, describing an upcoming lab revamp during a brief semester break, pointed out that, "*Bob has something planned. I know what they are but not the details. If he gets hit by a truck, we're in the poo.*"

The idea of someone no longer being a member of the group because of an 'accident' is something that is perceived to be a risk, albeit a remote one. A far greater risk, and an event more likely to occur, is that of the person leaving for more lucrative employment.

This is a real problem that the group has experienced time and time again. Retaining staff is difficult because of the more highly paid opportunities outside of the University. The need to record a person's 'knowledge' as information available to the group is something TSG members are aware of, but again this has to be weighed against the need to solve the problems at hand and deal with the fact that recorded information will become out of date relatively quickly.

The TSG see themselves as, to some extent, outsiders in the world of the university; one TSG member remarked that academics are "*protected to some extent by tenure*," while "*technical staff are just cannon fodder.*" It would not be unexpected if consultants occasionally—unconsciously or consciously—utilized the outsider's information tactic of secrecy. The dependence of a collection's integrity on the producers of documents—who may have different priorities than the digital library maintainers—is likely to be an intractable problem.

4.7 Creating Individual Information Resources

All of the TSG members created information resources that they stored on their own computer. Email filters were used by all participants, primarily to filter all messages from a mailing list directly into an associated mailbox. These messages were generally never read, and the mailbox would be consulted only if the consultant encountered a problem related to that mailing list's topics. In essence, a consultant uses his mail filters to build up private, searchable document collections based around the mailing lists to which the consultant subscribes. Some of the consultants also created significant resources based around files downloaded from various websites. These files are not cataloged, and are rarely formally organized. One consultant discovered that he had over 79,000 files (of all possible description) on his Macintosh hard drive. He reported using Sherlock (the Macintosh 'find file' utility, which supports searching both by file name and file contents) to locate a specific file. The files were organized, if that term can be used, in a very flat and wide file hierarchy: "*Quite often it* [Sherlock] *will find it on the desktop!*"

These personal resources should logically be included in a digital library to support TSG activities. Greenstone's definition of a digital library as consisting of a set of focused collections, it would be logical to include a personal resource as a distinct collection, accessible only by the individual TSG member. At present, however, the process of creating and maintaining a digital library is not trivial; it would be overly burdensome for an individual to use a full-fledged digital library tool such as Greenstone to maintain these individual resources as a collection integrated into a TSG digital library.

4.8 Browsing and search by location

Searching is certainly an important technique for locating documents, whether the search is conducted over the WWW as a whole generic WWW search engines such as Google, or over Usenet articles with Deja News, or over a TSG member's own hard drive using Sherlock. As noted in Section 3, digital library architectures assume that the primary access mechanism to the collection is searching. The types of search options supported by Greenstone and other digital libraries tend to closely match the standard options on WWW search engines and other search tools, so that movement between a digital library and other commonly used resources should be relatively painless.

Browsing is also an important technique for navigating information sources, with locational cues playing a part in information storage and subsequent retrieval. Frequently used physical documents (in contrast to "filing cabinet stuff") are strategically placed so as to be easily viewable. For example, large items like the network configuration diagram, lab timetable and the holiday rota are drawn on a whiteboard or pinned to a noticeboard and smaller items, such as a list of shortcut commands, are written on a Post-it note and stuck up anywhere handy. Sometimes the back of the hand is used to note down IP addresses and passwords!

Digitized documents stored on the hard drive are also sometimes retrieved (browsed) by location or appearance. This technique has been noted in earlier studies of file organization [1]. TSG consultants might place files that they expected to use regularly or shortly on their computer's desktop or other readily accessible spot (although this technique falls down if the consultant does not engage in regular housecleaning, as evidenced by the difficulties experienced by the consultant who had accumulated 79,000 files). Color is also sometimes used as an adjunct to location, when browsing through a set of files. One consultant uses color extensively with his stickies, and another consultant uses colored nodes in a mind map to organize information.

Digital libraries, as currently designed, offer little support for the user to structure information for retrieval based on appearance or position in the collection. Documents typically have no location, as such, and their appearance through the library interface cannot be altered by the user.

4.9 Information Might Not Look Like a Document

Sometimes the object of an information search is not a document as such. Instead, the consultant may be looking for an example of some sort—a sample of a type of file setup or a piece of code that solves a problem similar to the situation at hand, for instance. Some examples are located in formal sources, and are intended for use as problem solving aids. IT textbooks recognize this preference for learning through example; some of the most popular technical resources are example-heavy 'cookbooks'. One consultant prefers a particular software development kit because the kit and the associated website include a large amount of sample code. Other examples consulted have not been formally prepared as examples, but are simply remembered portions of the local system that a consultant feels might provide insight into a similar situation being faced. This latter type of example might include the contents of a .login file, a piece of code written by one of the local TSG members, or a particular file organization.

It would be difficult to formalize these examples for inclusion in a digital library. When searching a digital library what type of key search terms could the TSG member use to describe the specific problem that would match to the metadata recorded in the library? The local examples are exceptionally problematic: what sort of metadata would describe a file hierarchy? Formal examples in code libraries, 'cookbooks', and developer's websites are generally accompanied by a description of what the example does, but this description might not include the features that a consultant finds most useful in the example. Working from sets of examples generally entails considerable browsing, trying to match features of the solutions to the features of the problem at hand—and again, browsing can be difficult to effectively support in digital libraries.

4.10 Collaboration

Twidale and Nichols [20] describe the process of sense-making through interactions with peers as 'Over the Shoulder Learning' (OTSL). In this scenario, formal information sources such as online help, printed manuals, and training materials are regarded as secondary, rather than primary, sources for coming to an understanding of the overall 'shape' of a system. Instead, in OTSL it is interactions with colleagues that provide a great deal of the context in coming to grips with a complex system. These interactions may be prolonged and intensive, or—more typically—can be informal and short, focused on authentic tasks in the work environment.

Clearly the TSG relies to a large extent on OTSL for bring new staff members up to speed; the convention is to have new staff members share an office with a more experienced consultant who works on the same operating system or who supports the same group of labs. In the past, when the consultants all shared the same large office, this OTSL would have been achieved more naturally, with the new staff member being immersed in an ongoing conversation about all the different systems, user groups, and physical labs. As Twidale and Nichols [20] note, facilitating OTSL in a digital environment is problematic: the system must support a collaborative interaction with the document set, preserve a sense of history in the user's interactions, and maintain a task- or work-related focus in the presentation of information. Ideally, the context of the OTSL would be retained, perhaps in the form of a playback facility that would allow the user to review the concepts and procedures presented in OTSL sessions.

Current digital library architectures such as the NZDL do not provide an effective support for collaborative information behaviours such as OTSL. In a physical library, a reference librarian may serve as the expert 'colleague' in OTSL; good reference librarians do not simply give users answers to specific questions, but also who the users how their information needs can be satisfied by guiding a library user to a greater understanding of the library's contents and organization. Peer-to-peer OTSL may also occur as, for example, students working on the same assignment cluster around the same library catalog monitor [3]. Direct communication between digital library users is not currently a feature of digital library architectures such as Greenstone. 'Ask a Reference Librarian' services have been incorporated into digital library frameworks, including Greenstone [5]. The reference librarian services offered in a digital library have impoverished interfaces, in comparison to face-to-face interactions in a physical library. Typically the digital reference librarian services are based on an exchange of email—but email 'conversations' are generally too slow to support the (sometimes extensive) back-and-forth required to reach a consensus on the problem to be explored. Email is generally adequate for one person (a reference librarian or a OTSL participant) to give an answer to a question or to retrospectively explain how a solution was found, but not to allow two people to explore an information source collaboratively.

Experiments in providing digital reference services via synchronous communication software such as video conferencing [11] or a MOO [16] have yielded mixed results. On the one hand, true conversational interactions are made possible. However, these systems introduce problems of their own: users find it difficult to master the MOO interface, the videoconferencing systems require awkward-to-use hardware on participating sites, and the systems are easily crippled by slow (or even merely moderately fast) connect times. Further, the MOO interface is text-based, and videoconference image resolution may be fairly coarse—both limitations making it difficult for one person to observe another's interactions with the digital library or other information source.

Digital Libraries, as many are currently designed, offer little support for the user to select information for retrieval based on appearance or position in the collection. Although significant research has been carried out in terms of visualizing information spaces [10] [14] current DLs do not provide significant levels of support for either individuals structuring that which they have retrieved or for collaborative groups to structure their combined hits. Of course, a digital library could be used in an OTSL fashion by two TSG consultants physically sharing a monitor—so collaboration between members of the local TSG group is possible to the same extent as any piece of software. Collaboration between physically distant members of the greater TSG community, however, is not well supported. In sum, then, digital libraries with an architecture grounded in a view of the user as an individual searcher, such as Greenstone, are not well-suited to supporting collaborative information behaviours.

5. CONCLUSIONS

There are a number of conclusions that can be drawn from this study concerning the value of the ethnographic approach, the nature of the information gathered, the practical needs of the support staff for information seeking and information use, and how well different types of information system might best support such a group.

In terms of the use of ethnographic methods of identifying the information behaviourss of this group, this study should be viewed as a success. The richness of the data gathered and how it was analyzed enabled us to view the situation from several different perspectives. The information gathered demonstrated how each member of the group interacts with his own information sources and those used by others, including colleagues as information sources in their own right.

The information behaviourss we identified through using the ethnographic methods were primarily ones of browsing, cross checking and verifying, filtering information, monitoring sources, working through related information in steps to get to an end point and extracting information that was relevant from that which was not. These behaviourss correspond well with the framework described by Ellis [7].

Of these behaviourss, working through steps from a start point to an end point in its generic form could easily be supported by the searching mechanisms within a digital library. If the user were to begin with a simple reference to a subject or problem and work through to lists of citations and further references this would be similar in many respects to an academic researcher's use of a digital library. Extracting relevant information and differentiating between what was useful and valid would also be relevant behaviourss within a digital library construction.

In terms of the information needs, rather than actions, of the group there are several areas where a digital library might provide the support required. For example, material that is kept for historical reasons or safety reasons (e.g., supporting legacy systems) and is not actually used on a regular basis would be kept in a secure place, that could accessed easily by any member of the group. This type of material is relatively static.

Documents are generated by different members of the group and as people come into the group or leave it there is a need for continuity of information and recording 'intellectual capital' for use by the rest of the group. There was the recognition by the group that not having central repositories of information meant too much local knowledge resided in individuals and the risk in that was significant. The group members trust each other and respect the level of expertise each person has with respect to a particular area of the job. This would again seem to reinforce the utility of a digital library in reducing the risk of losing local information and the level of trust that the people could place in the content of the library because they would have generated much of it or at least cross-checked the original source with other sources.

Whilst a number of information behaviourss and information needs would, at least on initial inspection, appear to be well served by a digital library, not all would be. This study identified several information behaviourss and several features of the current digital library architectures that conflict with the information requirements or work habits of the TSG consultants.

If we begin with information behaviourss: the activities of browsing are not necessarily well-supported by digital libraries founded on the premise that the basic interaction mode is searching. Browsing is a less well-defined activity but is very relevant when dealing with less well defined problems. Cross-checking different sources is also an activity that is not well-supported in that the relationship of one document to another is only recorded in very limited terms. Often the group members cross-checked reports from different sources, at different times for different reasons. No assumptions about the validity or trustworthiness of the sources would be made until cross checking and verification had been made. In a digital library the need to record the result of this cross checking and validation would be important.

Monitoring sources and filtering information also relates back to the issue of trustworthiness and being able to re-present or re-structure the relationships between documents often to enable the user to deal with timeliness issues. The documents themselves would need to be updated regularly and in doing so this may change the type of relationship the document has to others, again requiring some restructuring or re-presenting of the source.

One critical issue when developing a digital library is defining a set of target users. For the TSG, it is unclear what community should be the focus for a collection. Each member of the TSG participates in a number of communities: the School TSG, the university TSG, the universal set of technical consultants, technical consultants within their individual operating system specialty (Macintosh, Linux, Windows, etc.), and technical consultants by role (security administrator, lab administrator, etc.). A digital library's usefulness would be compromised if it did not support an individual's membership in all of his communities, but at the same time a generic digital library for all TSG members would overwhelm an individual consultant with irrelevant material from communities which he is not a member of. In particular, it would be difficult to seamlessly merge internally and externally produced documents into a single digital library.

During this study a number of very interesting and useful insights into the information behaviourss of this group were gained. While we were aware of previous research on how people use different information sources and how this was partially dependent on the task they were carrying out, the results tested our assumptions about what else the use of such sources was dependent upon. It also enabled us to question assumptions about how appropriate digital libraries are in this situation.

The group studied here does not necessarily have the same mix and proportion of information behaviours exhibited by other groups. It may not be able to make as effective use of a digital library, however broad that definition may be taken to be, as groups such as academic researchers. For a digital library to be viewed by this group as more than just another information source it would need to provide a greater level of support than might currently be seen to be the case.

This group is very adept at using the most efficient technologies or tools to get what they want. These technologies have affordances that digital libraries do not have and may never be able to match. The behaviours of cross-checking, monitoring, re-presenting and re-structuring information have a strong link with the activity of annotation, something that is not that well supported in digital libraries.

There is a relationship between the type of annotation, the tools used to make the annotation and the materials on which the annotation is made. Marshall [15] uses the example of student textbooks where underlining is used to help students identify and reflect on key phrases or terms to demonstrate this relationship. Underlining was used in preference to highlighting (using highlighter pens) in paperback books because the paper is absorbent and the highlighter ink would leak through to the other side of the page. If we look at a simple example of the same type of relationship for the technical support group, the Post-it note or physical yellow sticky for recording useful pieces of information or lists of tasks to remember has value because it can be transported physically from one point to another, over time and edited or amended over time.

The Greenstone architecture can be seen as a good implementation of a typical digital library. At the moment, our results compare the types of information behaviour a typical implementation supports with those undertaken by our subject group. Having reviewed our conclusions, are we being overly harsh in judging the potential value of a digital library for this group? Are we setting a standard that no comprehensive resource for this group could ever meet?

The work context for this group of subjects does provide a picture of information behaviours that perhaps a 'typical' digital library implementation was never designed to support. If that is the case, we should not be surprised the support is not there.

If we adopt a broader view of digital libraries, as advocated by Levy and Marshall [12], would that lessen the significance of our conclusions? The answer is probably yes, but only in part. Re-examining them, in the context of a broader view, still leaves us with the problem that the broader view is not currently the typical implementation. This will change but we would hope that the results of this study may contribute to pushing out those boundaries.

A final lesson to be learned from this study is the rather prosaic observation that the technical consultants use information sources to support their work. Their job is not to consult digital libraries or to gather information, but rather to selectively use resources that will enable them to effectively and efficiently solve problems. This commonsense point tends to be obscured when system developers concentrate on creating an information system, rather than on ensuring that the system created is useful and usable, or even investigating whether a system should be created at all!

6. ACKNOWLEDGMENTS

We gratefully acknowledge the patience of the technical support group that was studied, and thank them for their participation in this project.

7. REFERENCES

[1] Barreau, D., and Nardi, B.A. Finding and reminding: file organization from the desktop. SIGCHI Bulletin 27(3), July 1995, 39 – 43. http://www.cwi.nl/~steven/sigchi/bulletin/1995.3/barreau.html

[2] Chatman, Elfreda A. The impoverished life-world of Outsiders. Journal of the American Society for Information Science 47(3), 1996, 193-206.

[3] Crabtree, Andy, Twidale, Michael, O'Brien, Jon, and Nichols, David M. Talking in the library: Implications for the design of digital libraries. in Proceedings of the Second ACM International Conference on Digital Libraries (Philadelphia, PA, 1997), 221-228.

[4] Crabtree, Andy, Nichols, David M., O'Brien, Jon, Rouncefield, Mark, and Twidale, Michael B. Ethnomethodologically-informed ethnography and information system design. Journal of the American Society for Information Science 51(7), 2000, 666-682.

[5] Cunningham, Sally Jo. Providing internet reference service for the New Zealand Digital Library: gaining insight into the user base for a digital library. in Proceedings of the 10th International Conference on New Information Technology (Hanoi, Vietnam, March 24-26 1998), 27-34.

[6] Cunningham, Sally Jo, and Mahoui, Malika. Search behaviour in two digital libraries: a comparative transaction log analysis. in Proceedings of the European Conference on Digital Libraries (Lisbon, Portugal, September 2000), Springer-Verlag, 418-423.

[7] Ellis, D. A behavioural approach to information retrieval design. Journal of Documentation, 46, 1989, 318-338.

[8] Hideaki Kanai and Katsuya Hakozaki. A Browsing System for a Database Using Visualization of User Preferences. Proceedings of the IEEE International Conference on Information Visualization, 2000. London, IEEE.

[9] Knowles, Chris, and Cunningham, Sally Jo. Information behaviour of technical support workers: an ethnographic study. In Proceedings of OZCHI 2000 (Sydney, Australia, December 2000), IEEE Press, 275-280.

[10] Lamping, J., Rao, R., and Pirolli, P. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. Proceedings of the ACM Conference on Human Factors in Software(CHI '95). ACM. 1995.

[11] Lessick, Susan, Kjaer, Kathryn, and Clancy, Steve. Interactive Reference Service (IRS) at UC Irvine: expanding reference service beyond the reference desk. ACRL '97 National Conference, 1997, http//www.ala.org/acrl/paperhtm.a10.html.

[12] Levy, D. and Marshall, C. C. Going Digital: A Look at Assumptions Underlying Digital Libraries. Communications of the ACM, 38, pp 77-84. ACM 1995.

[13] Lieu, Y-H., Dantzig, P., Sachs, M., Corety, J.T., Hinnesbusch, M.T., Damashek, M., and Cohen, J. Visualizing document classification: a search aid for the digital library. Journal of the American Society for Information Science 51(3), 2000, 216-227.

[14] Mackinlay, J., Rao, R., and Card S. An organic user interface for searching citation links. Proceedings of the ACM Conference on Human Factors in Software (CHI '95). ACM. 1995 pp. 67-73.

[15] Marshall, C. C. Annotation: from paper books to the digital library. Proceedings of DL 97.pp131-140. ACM 1997.

[16] Meyer, Judy. Servicing reference users. 1997 http://www.ala.org/editions/cyberlib.net/4meyer01.html

[17] Orr J.E. Talking about Machines: An Ethnography of a Modern Job. 1996 Ithaca, NY: ILR Press.

[18] Reeves, E.M. A study of usability aspects of a graphical user interface for discretionary users. Ph.D. thesis, University of Bristol, Bristol, UK.

[19] Soloman, P. Discovering information behaviours in sense making: Time and Timing. Journal of the American Society for Information Science 48(12), 1997, 1097-1108.

[20] Twidale, Michael B., and Nichols, David M. Using studies of collaborative activity in physical environments to inform the design of digital libraries. in Proceedings of the CSCW'98 Workshop on Collaborative and co-operative information seeking in digital information environments. Also available as Technical Report CSEG/11/1998 (Computing Department, Lancaster University, http:///www.comp.lancs.ac.uk/ computing/research/cseg/98_rep.htm

Developing Recommendation Services for a Digital Library with Uncertain and Changing Data

Gary Geisler
Interaction Design Laboratory
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360, USA
+1 919 489 2759
geisg@ils.unc.edu

David McArthur, Sarah Giersch
Eduprise
2000 Perimeter Park Drive
Morrisville, NC 27560, USA
+1 919 376 3424
{dmcarthur, sgiersch}@eduprise.com

ABSTRACT

In developing recommendation services for a new digital library called iLumina (www.ilumina-project.org), we are faced with several challenges related to the nature of the data we have available. The availability and consistency of data associated with iLumina is likely to be highly variable. Any recommendation strategy we develop must be able to cope with this fact, while also being robust enough to adapt to additional types of data available over time as the digital library develops. In this paper we describe the challenges we are faced with in developing a system that can provide our users with good, consistent recommendations under changing and uncertain conditions.

Keywords

Digital Library, Recommender System, User Services.

1. INTRODUCTION

Digital libraries—and Web-based resource collections in general—have traditionally enabled their users to locate resources through search and browse services. Over the past decade there has been growing use of recommendation systems as a way to suggest new items of potential interest to people [3]. In designing a new digital library (DL) called iLumina, we are exploring the potential of recommendation services as a way for users to find resources of interest in a digital library. iLumina is a DL of undergraduate teaching materials for science, mathematics, engineering, and technology (SMET) education [1], now being developed by Eduprise, The University of North Carolina at Wilmington, Georgia State University, Grand Valley State, and Virginia Tech. Because the DL is new and non-commercial, the data available to use as input into a recommendation system will be uneven and will change over time. We are particularly interested, therefore, in how to create robust recommendation services for a DL that contains changing and uncertain data.

JCDL '01, June 24-28, 2001, Roanoke, Virginia.
Copyright 2001 ACM 1-58113-000-0/00/0000...$5.00.

2. INFORMATION AVAILABLE FOR RECOMMENDATIONS

The iLumina digital library contains content across a wide range of science, math and engineering disciplines. Our user community includes instructors, students, and resource contributors; some—but not all—will register and provide profile information. From our server logs we expect to have usage data related to both resources and users. These relatively standard characteristics provide us with four classes of data potentially useful for recommendation services:

Resource characteristics: All resources will be described by basic, IMS derived [4] metadata that identifies elements such as title, description, format, and requirements. This metadata will be rich and consistent.

Resource quality judgements: Subject experts will formally review many (but not all) of library resources. All of the content will, however, be available to be reviewed informally by users; similarly, all materials will have passed minimal acceptability standards. But the written reviews, where available, can provide richer and more subjective information about the potential usefulness of the resource for specific situations.

Resource use: Data describing how often a given resource has been downloaded, reviewed, or used as a component in larger resources is available from server logs and the resource database. Data about resources can also be associated with individual users to develop patterns of usage by user characteristics.

User profile: Our database contains descriptive information about registered users, such as status, affiliation, areas of interest, explicit resource ratings, and service preferences. This information can be used to group users based on various similarity characteristics, track resource usage data for a given user, and adjust recommendations based on expressed preferences.

All of this data can be very useful for generating recommendations, but they vary in quality and likelihood of existence. For example, we do not require users to register, so profile information will vary. Reviews will range from expert-level, instructor judgements to student comments. Some registered users will rate resources, others will not. Table 1 summarizes the quality and existence characteristics for each of the basic types of data we expect to use as input for generating recommendations.

Table 1. Characteristics of the types of data available for generating recommendations

Type of data	Quality	Existence
Resource characteristics	High	Always
Resource quality	Variable	Variable
Resource use	High	Variable
User profile	Variable	Variable

3. GENERATING RECOMMENDATIONS

Given the characteristics of the types of data we have available, several schemes for providing recommendations are possible. At one extreme, we could use only data that we know is high quality and always available, such as resource characteristics and resource use. One example of this strategy would be to recommend to users resources that are similar, on specific metadata fields, to ones they have previously downloaded. The obvious downside of this strategy is that other potentially rich sources of data are ignored.

At the other extreme, we can include all potentially useful sources of data in our recommendation scheme, and attempt to compensate for missing inputs and inputs of variable quality. The downside of this strategy is that it is hard to combine multiple information sources in a principled way, and the variable quality of sources threatens the usefulness of recommendations based on them.

Different rules could be used to, in effect, define various points along this continuum of recommendation strategies. These might include (information types used noted):

- If the user suggests a resource she likes, we can suggest structurally similar resources (resource characteristics)
- If profile information for the user exists, we can use it to suggest resources (resource characteristics, user profile)
- If profile information exists and the user has previously downloaded resources, we can suggest resources based on the previously downloaded resources (resource characteristics, user profile, resource use)
- If the user suggests a resource she likes or profile information exists, we can suggest resources based on all available data (resource characteristics, user profile, resource use, resource quality judgments)

4. CHALLENGES

There is evidence that recommendations can be improved by combining methods, such as collaborative and content-based approaches [2]. In iLumina we will collect a range of data that supports such a multiple-source recommendation strategy. However, we can only be certain of the existence and reliability of the basic resource descriptive data; the availability of other data such as reviews, ratings, and user profiles is less predictable.

This uncertainty brings into question the value of a DL recommendation scheme that attempts to use all data sources and adapts when some of the data does not exist or is unreliable. In such cases, will the recommendations still be useful? More generally, do DL recommendations improve as the number of information sources they use increases? Is a strategy that uses all available information, some of which may not be available, more effective than one that uses the high-quality, always available resource characteristics? If so, given that gathering each type of data comes at some cost, is the difference in effectiveness worth the cost? Can we evaluate the cost/benefit ratio of using the different types of data?

Because the iLumina DL is new we will be in a position to address these kinds of questions. We expect to phase-in user services that provide us with relevant data over time; the data available to use in the second year of operation will be much broader than that available in the first year. As iLumina grows, then, we will not only be able to devise recommendation schemes that incorporate new types of information as they become available, but will also be able to employ user opinions to judge how the perceived value of different schemes improves (or not) as the data sources become richer.

5. CONCLUSION

Recommendations can be a valuable service for users of a DL such as iLumina, which will eventually contain many resources. The variety of data associated with both the resources and users of a DL represent a potentially rich source of input for recommendation services. As the iLumina DL evolves, we will be interested in developing both strategies that provide useful recommendations to users even if the data available is uneven and changes over time and methods for evaluating these methods. We expect that the results of our efforts will be informative to developers of many types of DLs.

6. REFERENCES

[1] McArthur, D., Giersch, S., Graves, B., Ward, C.R., Dillaman, R., Herman, R., Lugo, G., Reeves, J., Vetter, R., Knox, D., & Owen, S. (in press). Towards a Sharable Digital Library of Reusable Teaching Resources: Roles for Rich Metadata. *Journal of Educational Resources in Computing*.

[2] Mooney, R. J. and Roy, L. (2000). Content-Based Book Recommending Using Learning for Text Categorization. In *Proceedings of the 5th ACM Conference on Digital Libraries* (San Antonio, TX, June 2000).

[3] Recommender Systems [theme issue]. (1997). *Communications of the ACM* 40(3).

[4] The IMS Global Learning Consortium, http://www.imsproject.org

Evaluation of DEFINDER: A System to Mine Definitions from Consumer-Oriented Medical Text

Judith L. Klavans
Center for Research on Information Access
Columbia University
New York, NY, 10027
klavans@cs.columbia.edu

Smaranda Muresan
Department of Computer Science
Columbia University
New York, NY, 10027
smara@cs.columbia.edu

ABSTRACT

In this paper we present DEFINDER, a rule-based system that mines consumer-oriented full text articles in order to extract definitions and the terms they define. This research is part of Digital Library Project at Columbia University, entitled PERSIVAL (PErsonalized Retrieval and Summarization of Image, Video and Language resources) [5]. One goal of the project is to present information to patients in language they can understand. A key component of this stage is to provide accurate and readable lay definitions for technical terms, which may be present in articles of intermediate complexity.

The focus of this short paper is on quantitative and qualitative evaluation of the DEFINDER system [3]. Our basis for comparison was definitions from Unified Medical Language System (UMLS), On-line Medical Dictionary (OMD) and Glossary of Popular and Technical Medical Terms (GPTMT). Quantitative evaluations show that DEFINDER obtained 87% precision and 75% recall and reveal the incompleteness of existing resources and the ability of DEFINDER to address gaps. Qualitative evaluation shows that the definitions extracted by our system are ranked higher in terms of user-based criteria of usability and readability than definitions from on-line specialized dictionaries. Thus the output of DEFINDER can be used to enhance existing specialized dictionaries, and also as a key feature in summarizing technical articles for non-specialist users.

Keywords

Text data mining, medical digital libraries, natural language processing, automatic dictionary creation.

1. The Digital Library and Text Mining for Definitions: the DEFINDER System

The existence of massive digital libraries containing freeform documents has created an unprecedented opportunity to develop and apply effective and scalable text mining techniques for the automatic extraction of knowledge from unstructured text [1]. Text mining applications raise particularly challenging problems within digital libraries since they involve large collections of unstructured documents. Our approach is to combine shallow natural language processing techniques with deep grammatical analysis in order to efficiently mine text.

Automatic identification and extraction of terms from text has been widely studied in the computational linguistics literature [2], and many systems exist for this task using both symbolic and statistical techniques. The extraction of definitions and their associated terms has been less widely studied, although extraction of lexical knowledge has a rich literature [7].

Through an analysis of a set of consumer-oriented medical articles, we identified typical cue-phrases and structural indicators that introduce definitions and the defined terms. Our system, DEFINDER, is based on two main functional modules: 1) a shallow text processing module which performs pattern analyses using a finite state grammar, guided by cue-phrases ("is called", "is the term used to describe", "is defined as", etc.) and a limited set of text-markers ((), --) and 2) a grammar analysis module that uses a rich, dependency-oriented lexicalist grammar (English Slot Grammar [4]) for analyzing more complex linguistic phenomena (e.g. apposition, anaphora).

2. Evaluation: Users and Uses

In this brief paper, we present the results of three methods to evaluate the output of our system: 1) performance in terms of precision and recall, 2) quality of extracted definitions in terms of user-based criteria of readability, usefulness and completeness and 3) a method to evaluate the coverage of on-line specialized dictionaries. For the first two, we performed a user-centered evaluation using non-specialist subjects. For the latter we chose a set of defined terms extracted by our system and compared them against three on-line dictionaries. The results we have obtained were run over a limited set of articles in order to thoroughly test our methods before moving to a larger scale user-based evaluation of significantly more data. We present the results of three experiments to quantitatively and qualitatively measure DEFINDER output.

2.1 Definition Extraction Performance

The purpose of this experiment was to measure the performance of DEFINDER in terms of precision and recall against a human-determined "gold standard". Four subjects unrelated to the project were provided with a set of nine patient-oriented articles and were asked to annotate definitions and the terms they define. We chose several genres (medical articles, newspapers, manual

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

chapters, book chapters) from trusted resources. The resulting gold standard was determined by those definitions marked-up by at least 3 out of the 4 subjects and consisted of 53 definitions. DEFINDER identified 40 out of these 53 definitions obtaining 86.95% precision and 75.47% recall.

2.2 User Judgements on Definition Quality

In this experiment we asked users to rank definitions to determine if they are readable, useful or complete. The motivation is that there is unlikely to exist a single definition suitable for both specialists and non-specialists. Indeed, specialized on-line dictionaries, while valuable resources, can be too technical for non-specialists. We evaluated the quality of DEFINDER output in comparison with two specialized on-line dictionaries (UMLS and OMD). Eight subjects not qualified in the medical domain participated in the experiment. They were provided with a list of 15 randomly chosen medical terms and their definitions from these three sources. The task was to assign to each definition a quality rating for three criteria: usefulness (U), readability (R) and completeness (C) on a scale of 1 to 7 (1 worst, 7 best). The source of each definition was not given in order not to bias the experiment. Statistical significance tests were performed for subjects and terms using Kendall's coefficient of concordance, W [6] and the sign test [6].

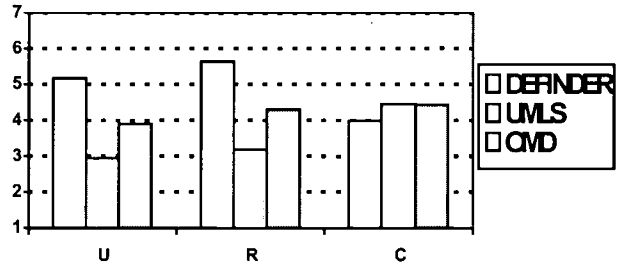


Figure 1 - Average quality rating (AQR)

We first measured the average quality rating for each of the three sources on the three criteria. The results in Fig. 1 show that DEFINDER clearly outperforms the specialized dictionaries for usefulness and readability to a statistically significant degree, given by the sign test (p=0.0003). In terms of completeness both UMLS and OMD performed slightly better (p=0.04).

One question that arises in computing the AQR is whether the high scores given by one subject can compensate for the lower values given by other subjects, thus introducing noise. To validate our results, we performed a second analysis to evaluate the relative ranking of the three definitional sources. Using Kendall's coefficient of correlation, W, we first measured the interjudge reliability on each term, and for terms with significant agreement we compute the level of correlation between them. If W was significant, we compared the overall mean ranks of the three sources. We obtained statistically significant W values for usefulness and readability (W=0.54 and W=0.45 at p=0.01 and p=0.05 respectively), while for completeness the correlation was not statistically significant. Thus Figure 2 shows the results for usefulness (U) and readability (R) for which DEFINDER outranked both UMLS and OMD.

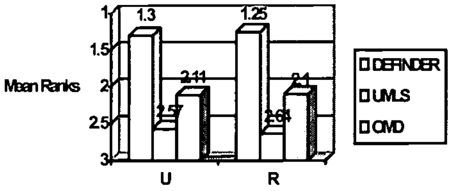


Figure 2 – Ranking

2.3 Coverage of On-line Dictionaries

DEFINDER identifies terms and definitions lacking from existing resources. To evaluate coverage, we choose a base test set of 93 terms and their associated definitions, extracted by our system from text. Three cases were found, as shown in Table 1: (1) the term is listed in one of the on-line dictionaries and is defined in that dictionary (defined); (2) the term is listed in one of the on-line dictionaries but does not have an associated definition (undefined); (3) the term is not listed in one of the on-line dictionaries (absent).

Term	UMLS	OMD	GPTMT
defined	60% (56)	76% (71)	21.5% (20)
undefined	24% (22)	-	-
absent	16% (15)	24% (22)	78.5% (73)

Table 1 Coverage of Existing Online Dictionaries

Table 1 shows that on-line medical dictionaries are incomplete compared to potential DEFINDER output. For example, column two shows that in OMD only 71 terms out of 93 are listed, thus leading to 76% completeness, while GPTMT, a glossary addressed to non-specialists is far from being complete, i.e. only 20 out of 93 terms were present. This proves the ability of DEFINDER to enhance on-line dictionaries with readable and useful definitions.

3. References

[1] Hearst, M. Untangling Text Data Mining. *Proc. of ACL'99* University of Maryland, June 20-26, 1999 (invited paper).

[2] Justeson, J. and Katz, S. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. Natural Language Engineering. Vol 1(1). 1995. pp. 9-27.

[3] Klavans J.L., Muresan S. DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text. *Proc of AMIA 2000*; pp. 1906.

[4] McCord M.C. The Slot Grammar system. IBM Report; 1991.

[5] McKeown K.R et al. PERSIVAL, a System for Personalized Search and Summarization over Multimedia Healthcare Information. *Proc of JCDL 2001*.

[6] Siegal, S. and Castellan, N.J. (1988). Non-parametric statistics for the behavioural sciences (2nd Edition). New York: McGraw Hill.

[7] Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, Seroussi B, Boisvieux JF. From Text to Knowledge: a Unifying Document-Oriented View of Analyzed Medical Language. Proceedings of IMIA WG6. 1997.

Overview of The Virtual Data Center Project and Software

Micah Altman, L. Andreev, M. Diggory,
G. King, E. Kolster, A. Sone, S. Verba
Harvard University
G-4 Littauer Center, North Yard
Cambridge, MA 02138
(617) 496-3847
Micah_Altman@Harvard.edu

Daniel L. Kiskis and M. Krot
University of Michigan
Media Union
2281 Bonisteel
Ann Arbor, MI 48109

ABSTRACT

In this paper, we present an overview of the *Virtual Data Center* (VDC) software, an open-source digital library system for the management and dissemination of distributed collections of quantitative data. (see <http://TheData.org>). The *VDC* functionality provides everything necessary to maintain and disseminate an individual collection of research studies, including facilities for the storage, archiving, cataloging, translation, and on-line analysis of a particular collection. Moreover, the system provides extensive support for distributed and federated collections including: location-independent naming of objects, distributed authentication and access control, federated metadata harvesting, remote repository caching, and distributed 'virtual' collections of remote objects.

Categories and Subject Descriptors

H.3.7. [**Information Systems**] -Information Storage and Retrieval - Digital Libraries (H.3.7); -- *DL System Architecture, Distributed Systems, Information Repositories, Federated Search, DL Impact*

General Terms

Management, Design, Standardization

Keywords

Numeric Data, Open-Source, Warehousing.

1. INTRODUCTION

Researchers in social sciences, and in academia in general, increasingly rely upon large quantities of numeric data. The analysis of such data appears in professional journals, in scholarly books, and increasingly often in popular media. For the scholar, the connection between research articles and data is natural. We analyze data and publish results. We read the results of others' analyses, learn from it, and move forward with our own research.

But these connections are sometimes difficult to make. Data supporting an article are often difficult to find and even more difficult to analyze. Archiving, disseminating and sharing data is crucial to research, but is often costly and difficult. [Sieber 1991] Thus, our ability to replicate the work of others and to build upon it is diminished. Researchers, university data centers, and students all face challenges when trying to find and use quantitative research data.

The *Virtual Data Center* (VDC) software is a comprehensive, open-source digital library system, designed to help curators and researchers face the challenges of sharing and disseminating research data in an increasingly distributed world. The VDC software provides a complete system for the management and dissemination of federated collections of quantitative data.

2. Features, Design, and Implementation

The VDC provides functionality for producers, curators and users of data. For producers, it offers naming, cataloging, storage, and dissemination of their data. For curators, it provides facilities to create virtual collections of data that bring together and organize datasets from multiple producers. For users, it enables on-line search, data conversion, and exploratory data analysis facilities.

More specifically, the system provides five areas of functionality:

(1) Study preparation. (unique naming, conversion tools for multiple data and documentation formats, tools for preparing catalog records for datasets);
(2) Study management. (file-system independent dataset and documentation storage, archival formatting, cataloging);
(3) Interoperability. (DC, MARC and DDI metadata import and export; OpenArchives and Z39.50 query protocol support);
(4) Dissemination. (extract generation, format conversion, subset generation, and exploratory data analysis).
(5) Distributed and federated operation. (location-independent naming, distributed virtual collections, federated metadata harvesting, repository exchange and caching, and federated authentication and authorization)

Each VDC library comprises a set of independent, interoperating components. These independent VDC libraries can also be federated together, sharing collections through harvesting and dynamic caching processes. (See Figure 1)

Many of the core components are modeled after the design described in Arms [1997]: objects are stored in repositories, referenced through names that are resolved by name servers, and are described in index servers that support searching. We also incorporate some features of the NCSTRL [Davis 2000] architecture and particularly Lagoze's [1998] idea of a collection as an object that groups queries against remote index servers.

We have extended this core design in several significant ways. First, to support completely distributed operation of the

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

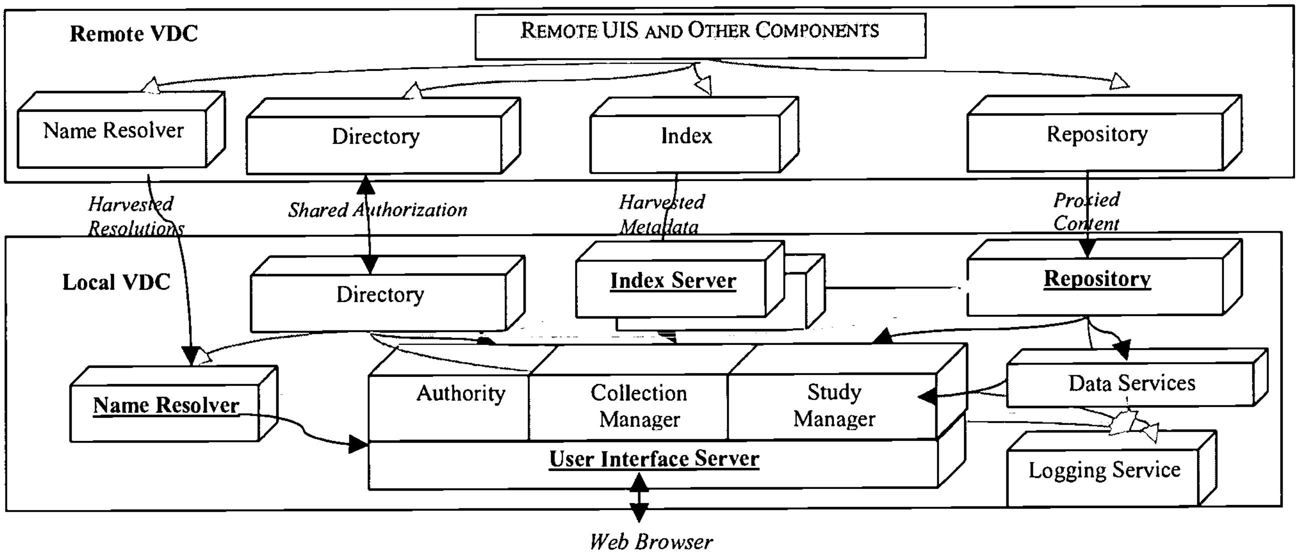


Figure 1. Simplified Representation of VDC

components, we have added to each VDC node a directory server, which allows the components to locate each other, and a centralized logging service, which aids the administrator in tracking usage of the system. Second, the idea of the collection has been expanded to provide independent encapsulation of the specification of virtual content and the 'view', or organization, of that content. Third, the metadata harvester, caching repository proxy, and distributed authentication and authorization components work together support the 'federation' of independent VDC libraries: The result is that content from a group of libraries is made available to their collective users, while each library maintains complete control over how its collections are accessed and how its patrons are authenticated.

Our implementation strategy emphasizes 'open source' development, and integration of the system into a production environment. The director of the *Digital Library Initiative, Phase 2*, of which the VDC is a part, notes the 'unnatural separation' between the producers and consumers of digital libraries, and calls for a balance among research, application, content and collections. [Griffin 1998] In keeping with this admonition, the VDC software system is not simply an isolated research project, it is also a part of Harvard University's first generation *production* digital library system – the VDC software and HMDC site support real use by Harvard library patrons. And this project benefits from taking part in an unusually large and decentralized library system, from cross-fertilization with Harvard's own digital library efforts (see [Flecker 2000]), and from being heavily used by the Harvard research community. In addition, to support the academic norms of openness and accessibility associated with research data, we are in keeping with Lessig's [1999] assertion that the 'code' supporting the fundamental infrastructure for citations must be open. Our code is 'open source', and freely available for modification and use. (see <http://TheData.org>)

3. CONCLUSIONS.

By providing a portable software product that makes the process of data archiving and dissemination automatic and standardized, the Virtual Data Center will help researchers and data archives meet the challenges of sharing and using quantitative data. Consequently, we believe that quantitative research will become easier to replicate and extend.

4. ACKNOWLEDGMENTS

Supported by NSF Award #IIS-9874747.

5. REFERENCES

[1] Altman, M., et al., 2001, "The Virtual Data Center," Working paper. <URL:http://thedata.org/publications/>

[2] Arms, W. Y., et al, 1997. "An Architecture for Information in Digital Libraries," *D-Lib Mag.*

[3] Davis, J. R. and C. Lagoze, 2000. "NCSTRL: Design and Deployment of a Globally Distributed Digital Library," *JASIS*, 51(3):273-280.

[4] Flecker, D. 2000. "Harvard's Library Digital Initiative" *D-LIB Mag.* 6(11).

[5] Griffin, S. 1998. "NSF/DARPA/NASA Digital Libraries Initiative" *D-Lib Mag.*4(7)

[6] Lessig, Lawrence 1999. *Code, And Other Laws of Cyberspace.* Basic Books, NY.

[7] Lagoze, C., D. Fielding, November 1998. "Defining Collections in Distributed Digital Libraries," *D-Lib Mag.*

[8] Sieber, J. E. (ed.) (1991). *Sharing Social Science Data.* Sage Publications, Inc, CA.

Digital Libraries and Data Scholarship

Bruce R. Barkstrom
Atmospheric Sciences
NASA Langley Research Center
Hampton, VA 23681-2199
1-757-864-5676
b.r.barkstrom@larc.nasa.gov

ABSTRACT

In addition to preserving and retrieving digital information, digital libraries need to allow data scholars to create post-publication references to objects within files and across collections of files. Such references can serve as new metadata in their own right and should also provide methods for efficiently extracting the subset of the original data that belongs to the object. This paper discusses some ideas about the requirements for such references within the context of long-term, active archival, where neither the data format nor the institutional basis can be guaranteed to remain constant.

Keywords

Data scholarship, structural data references, object data references, digital libraries, EOSDIS.

1. ON DATA SCHOLARSHIP

In the humanities, there are many mechanisms for identifying and cross-referencing document contents: "chapter and verse", tables of contents, indexes, concordances, glossaries, commentaries. While the table of contents and index now reside in the original document, most of the other scholarly works appear after publication and do not perturb the original document. To be effective, the references need to provide precise pointers to the start and end of the original document segments they point to. Over the years, the humanities have developed important conventions to anchor the references, including structural features such as numbered document sections and non-structural references such as page numbers. We are only beginning to recognize the need for such apparatus in many areas of digital librarianship. In this brief paper, we explore the implications of this need within the context of scientific data collections for the Earth sciences.

The emphasis in current Earth science data collections, such as NASA's Earth Observing System Data and Information System (EOSDIS), has been on preserving collections of files and providing catalogs with some simple metadata to help users find files that interest them. These systems treat files in ways that are similar to journal articles in other kinds of libraries [3, 4]. However, they have not dealt with the issue of recording references to and properties of interesting objects within the files, or of allowing data scholars to extract specific subsets of data that belong to these objects [1]. Thus, we have had little serious study of the problems of identifying, recording, or providing a subset of pixels belonging to a hurricane within an Advanced Very High Resolution Radiometer (AVHRR) or Moderate Resolution Imaging Sprectoradiometer (MODIS) image.

Likewise, there appears to have been little work done on how to allow scientists to publish stable references to objects that occur in collections of files, particularly when the collections come from very different sources. In this case, a scientist might want to bring together data from a field experiment that is contained in one or two files of satellite data (e.g., a MODIS image collocated with an ASTER image) and in many small files of in situ data (such as radiosonde profiles of temperature and humidity).

These problems become still more interesting when we think of the long-term evolution of data systems. Over long time periods, references to structures within files require pointers that can survive structural rearrangement of the file format – because the archive needs to upgrade the original file from one version of a format to another. References to objects in file collections must be able to survive collection migrations from media to media and repository to repository. It would also be helpful for collections to be cognizant of retrospective pointer additions in other locations.

2. OBJECT REFERENCES

It is convenient to categorize Earth science data as either satellite data or in situ data. Satellite data usually come in large, homogeneously structured files. The satellite orbital motion provides an intricate and highly structured sampling of space-time. The instruments on the satellite add spatial or angular sampling patterns that are also quite regular – although there may be large differences in the spatial sampling patterns of different instruments on the same satellite. Satellite data producers work with their data "wholesale." They create large files that have "smooth edges" – all of the files in a data product contain one hour of data or one day of data. In situ data, on the other hand, come from specialized platforms near the Earth's surface, such as Earth-fixed collection sites, ocean buoys, ships, or aircraft. The files containing these data tend to be smaller and more diverse. Practical collections may be heterogeneous: data collected for a field experiment may contain a few, highly regular satellite files and a larger number of smaller, heterogeneous files.

In conventional Earth science thinking, the data in these files come from continuous fields. Recently, some Earth scientists have begun thinking in terms of data as collections of objects. In this new mode of thought, Earth scientists do not want data averaged over Earth-fixed grids. Rather, they want averages for irregularly shaped objects or that recognize the internal structure of object hierarchies. In the conventional approach, scientists might resample the spatial

This paper is authored by an employee(s) of the United States Government and is in the public domain.
JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
ACM 1-58113-345-6/01/0006.

structures of satellite or in situ data onto grids aligned with latitude and longitude. Of course, this can be quite expensive: satellite orbital inclinations do not align with the North-South boundaries of most latitude-longitude grids and resampling can involve numerically intensive spatial filters. In the object-oriented approach, rectangular grids become less important. Rather, references to "the original measurements that belong to an object" must allow for its irregular shape. This means that extracting "just the pixels that belong to the object" from the file may involve more than simply extracting a hyperslab from an array.

What makes this problem rather interesting is that normal rules of good library behavior exclude "writing in the original manuscript." This suggests that the indexes to objects need to be treated as external metadata. They remain outside the file that contains the original data, so that it remains unaltered.

A natural solution is to use a "pointer structure" to index the measurements within the file's data that belong to a particular object. Under the assumption that the data are contained in multi-dimensional arrays, an index needs to know the mapping between the array indexes and the sequential position of the measurements within the array. With this mapping, an object reference becomes a list of one-dimensional array segments, where each segment identifies a starting position in the array and an ending position. If the data in the file are images, it may also be possible to use object descriptions of the type used by geographic information systems, e.g. [2].

This approach suggests that digital libraries need to accommodate "object commentaries" that contain indexing structures with a list of objects and their properties. For each object, the indexing structure needs to include a list of files containing the object. For each file, the index would contain a collection of appropriate object properties and a list of array segments that contain the object. Provided the "list of files" uses a stable referencing scheme, this approach appears to be a satisfactory start for building external references that do not need to "write in the original manuscript."

When there are multiple, geographically dispersed copies of the original files and when the repository sites have independent hardware and software refresh policies, any object index will need additional functions that can identify structural variants. For example, if one copy of the data were written by a FORTRAN program and another copy were written by a C version, the two arrays may contain the same data, but the array indexes would be permuted. Thus, the pointers in an object index would need to identify which variant created a particular file. The pointer software would also need to recompute the array segments to correspond to the appropriate format variant. Alternatively, digital librarians need to create strong array ordering conventions.

3. COLLECTION INDEXES

The second problem for references arises from identifying collections of files that "belong together," particularly if the collection allows data scholars to obtain a homogeneous view of a particular class of objects. The identification mechanism not only needs to identify the files that belong together, it also needs to distribute this identification to other locations that contain copies of the collection. In this way, scholars that find a collection useful for working with a particular class of objects can make those objects more widely available. Such notification also allows a data set's users to identify who else has worked with a data set and to receive notice when the files move to new media or new formats.

Data producers will not have collection indexes available for their data when they create a data set. As time passes, scientists outside the original team will identify interesting objects within the collection. They are likely to identify the same objects in other collections. For example, a scientist interested in hurricanes might notice that MODIS contains data useful for identifying cloud properties (such as cloud top altitude), that CERES contains data useful for examining the hurricane's radiation budget, and that AMSU-E contains data useful for understanding the distribution of liquid water within the storm. These data are stored in different locations and may well have different array structures to record in the object indexes. Collection indexes thus need flexibility to add indexes for other collections – and are prime candidates for agent automation.

Creating collection indexes that contain object references raises some interesting long-term issues. If we knew for sure that the files would never be reformatted and that repositories would remain constant, we could settle for simply keeping track of the "Universal Name Reference" of the files in the collection. However, the problem becomes more difficult when data repositories find it necessary to reformat files because the original formatting software has a new version or because the file format's creators have disappeared. In this situation, a digital library needs to know which other libraries contain file collections that might fulfill searches for objects within the files. Clearly, this need places a requirement on the digital library to add external metadata that will let the file's users know that reformatting has taken place and where alternative files may reside.

Object indexes and collection indexes that contain references to objects in files are a critical component of data scholarship. At the same time, our inability to guarantee absolute stability in information technology raises interesting issues in implementing such indexes.

4. ACKNOWLEDGMENTS

The author is pleased to acknowledge the support of NASA's CERES project for this work.

5. REFERENCES

[1] Barkstrom, B. R. Digital Archive Issues from the Perspective of an Earth Science Data Producer. http://ssdoo.gsfc.nasa.gov/nost/isoas/us12/barkstrom/Archival%20Issues.html.

[2] http://www.gisportal.com provides numerous sources of information.

[3] Lamport, L. The implementation of reliable distributed multiprocess systems. Computer Networks, 2, 1978.

[4] Rosenthal, D. S. H. and V. Reich. Permanent Web Publishing. http://lockss.stanford.edu/freenix2000/freenix2000.html

SDLIP + STARTS = SDARTS
A Protocol and Toolkit for Metasearching

Noah Green
ngreen@cs.columbia.edu

Panagiotis G. Ipeirotis
pirot@cs.columbia.edu

Luis Gravano
gravano@cs.columbia.edu

Computer Science Dept.
Columbia University

ABSTRACT

In this paper we describe how we combined SDLIP and STARTS, two complementary protocols for searching over distributed document collections. The resulting protocol, which we call SDARTS, is simple yet expressible enough to enable building sophisticated metasearch engines. SDARTS can be viewed as an instantiation of SDLIP with metasearch-specific elements from STARTS. We also report on our experience building three SDARTS-compliant wrappers: for locally available plain-text document collections, for locally available XML document collections, and for external web-accessible collections. These wrappers were developed to be easily customizable for new collections. Our work was developed as part of Columbia University's Digital Libraries Initiative–Phase 2 (DLI2) project, which involves the departments of Computer Science, Medical Informatics, and Electrical Engineering, the Columbia University libraries, and a large number of industrial partners. The main goal of the project is to provide personalized access to a distributed patient-care digital library.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering, Query Formulation, Search Process, Selection Process*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Data Sharing, Web-based Services*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.2.4 [**Database Management**]: Systems—*Textual Databases, Distributed Databases*; H.2.5 [**Database Management**]: Systems—*Heterogeneous Databases*

1. INTRODUCTION

The information available in electronic form continues to grow at an exponential rate and this trend is expected to continue. Although traditional search engines like AltaVista can solve common information needs, they ignore the often-valuable information that is "hidden" behind search interfaces, the so-called *"hidden web."*

One way to access the information available in the hidden web is through *metasearchers* [12, 18, 24], which provide users with a unified searchable interface to query multiple databases simultaneously. A metasearcher performs three main tasks. After receiving a query, it determines the best databases to evaluate the query (*database selection*), it translates the query in a suitable form for each database (*query translation*), and finally it retrieves and merges the results from the different sources (*results merging*) and returns them to the user using a uniform interface.

Metasearching is a central component of the Digital Libraries Initiative–Phase 2 (DLI2) project at Columbia University, which involves the departments of Computer Science, Medical Informatics, and Electrical Engineering, the Columbia University libraries, and a large number of industrial partners (e.g., IBM, GE, Lucent). The project is named PERSIVAL (PErsonalized Retrieval and Summarization of Image, Video, And Language resources) and its main goal is to provide personalized access to a distributed patient-care digital library. The information needs vary widely among the users of the system. We have to provide access to all kinds of collections, ranging from Internet sites with consumer health information to the Columbia Presbyterian Hospital information system, which stores patient-record information and other relevant resources. Metasearching is further complicated by the different access methods used by each source, which range from public CGI-based interfaces to proprietary access methods used inside the Columbia Presbyterian Hospital. Key features of the project are the abilities to access all these distributed resources regardless of whether they are available locally or over the Internet, to fuse repetitive and conflicting information from multiple relevant sources, and to present concisely the retrieved information. Exploiting such a variety of information sources is a challenging task, which could benefit from information sources supporting a common interface for searching and metadata extraction.

Rather than defining yet another new protocol and interface that distributed sources should support, we decided to exploit existing efforts for our project. More specifically, we built on two complementary protocols for distributed search, namely SDLIP [20] and STARTS [11], and combined them to define SDARTS (pronounced "ess-darts"), which is the focus of this paper.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

SDLIP (Simple Digital Library Interoperability Protocol) is a protocol jointly developed by Stanford University, the University of California at Berkeley and at Santa Barbara, the San Diego Supercomputer Center, the California Digital Library, and others. The SDLIP protocol defines a layered, uniform interface to query and retrieve the results from each searchable collection through a common interface. SDLIP also supports an interface to access source metadata.

SDLIP is optimized for clients that know which source they wish to access. In contrast, the focus of STARTS (STAnford protocol proposal for Internet ReTrieval and Search) is on metasearching. A crucial component of STARTS is the definition of the specific metadata that a source should export to describe its contents. This metadata includes the vocabulary (i.e., list of words) in the source, plus vocabulary statistics (e.g., the number of documents containing each word). These summaries of the source contents are useful for the metasearchers' database selection task.

SDLIP and STARTS complement each other naturally. SDLIP has a flexible design that allows it to host different query languages and metadata specifications. The major parameter and return types of its methods are passed as XML, and the DTDs for this XML allow for extensions and instantiations of the protocol. Thus, SDLIP can easily host the main components of the STARTS protocol. The result of this combination is SDARTS, which can be regarded as an instantiation of the SDLIP protocol with a specific query language, and, more importantly, with the richer metadata interface from STARTS, which is useful for metasearching.

To simplify making document collections compliant with SDARTS, we have developed a software toolkit that is easily configurable. This toolkit includes software to index locally available collections of both plain-text and XML documents. Also, to be able to wrap external collections over which we do not have any control and which do not support SDARTS natively, the toolkit includes a reference wrapper implementation that can be augmented for new external collections with relatively small changes.

SDARTS and its associated software toolkit, which are the focus of this paper, provide the necessary infrastructure for metasearching and for incorporating collections into our digital libraries project with minimal effort. In Section 2 we describe in detail the metasearching tasks and the challenges involved. In Sections 3 and 4 we describe STARTS and SDLIP respectively, the two protocols on which we have based SDARTS. The resulting protocol is described in Section 5 and the toolkit and reference implementations are presented in Section 6. Finally, Section 7 provides further discussion of our overall experience.

2. METASEARCHING

As we briefly mentioned in the Introduction, metasearching mainly consists of three tasks: *database selection*, *query translation*, and *result merging*.

- **Database Selection:** A metasearcher might have hundreds or thousands of sources available for querying. An alternative to broadcasting queries to every source every time is to only send queries to "promising" sources. This alternative results not only in better efficiency but in better effectiveness as well. Selecting the best sources for a given query requires that the metasearcher have some information about the contents of the sources. Some database selection techniques rely on human-generated descriptions of the sources. More robust approaches rely on simple metadata about the contents of the sources like their vocabulary (i.e., list of words) together with simple associated statistics. Research in this area includes [10, 13, 21, 12, 18, 24].
- **Query Translation:** Metasearchers send queries to multiple different sources, which requires translating the queries to the local query language and capabilities supported at each source. Query translation is facilitated if sources export metadata on their query capabilities, on whether they support word stemming, on the attributes or fields available for searching (e.g., author and title), etc. Research in this area includes [6, 7, 8].
- **Result Merging:** Sources typically use undisclosed document ranking algorithms to answer user queries, which makes the combination of query results coming from multiple sources challenging. Furthermore, even two collections that use the same ranking algorithm might rate a common document quite differently depending on the other documents present in each collection. Result merging is facilitated if sources export metadata on their contents and query results. Research in this area includes [22, 5].

A metasearcher needs information about the underlying collections to perform the tasks above successfully. Consequently, there is a need for a layer on top of the information sources that will mask source heterogeneity and export the right metadata to metasearchers. One protocol that was designed to solve exactly this problem is STARTS, which we briefly review next.

3. STARTS: A PROTOCOL PROPOSAL TO FACILITATE METASEARCHING

STARTS [11] is a protocol proposal that defines the information that a source should export to facilitate metasearching. By standardizing on a common way of interacting with clients and by defining what information a document source should export, metasearching becomes a much easier task. Of course, the exported information alone is not a panacea and does not solve the metasearching problems, but at least it can make these problems more tractable. Specifically, STARTS defines the information that should be included in the queries to the sources, the format of the query results, and the metadata a source should export about its contents and capabilities [11].

For historical reasons, the original STARTS specification used Harvest's SOIF format [3] to encode queries, results, and metadata. Also, STARTS did not define explicitly how the information is transported. For our project, we defined an encoding of all the STARTS information in XML. All STARTS queries, result sets, and metadata objects are then represented as XML documents. This means that all STARTS elements can be easily manipulated by available XML technologies, such as XSL, SAX, etc. Additionally, the transformation from one XML dialect to another can be easily achieved using off-the-self tools, thus it is easy to support other query languages by just adding a thin layer

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE starts:scontent-summary SYSTEM "http://www.cs.columbia.edu/~dli2test/dtd/starts.dtd">
<starts:scontent-summary
        xmlns:starts="http://www.cs.columbia.edu/~dli2test/STARTS/"
        version="Starts 1.0"
        stemming="false"
        stopwords="false"
        case-sensitive="true"
        fields="false"
        numdocs="19997"
>
...
       <starts:field-freq-info>
              <starts:field type-set="basic1" name="body-of-text"/>
              <starts:term>
                     <starts:value>algorithm</starts:value>
              </starts:term>
              <starts:term-freq>75</starts:term-freq>
              <starts:doc-freq>34</starts:doc-freq>
...
```

Figure 1: A small fraction of a STARTS metadata object for a document collection.

that translates the query format of another query language to STARTS. Another strong reason to transform STARTS to XML format is that SDLIP, the protocol that we review next, uses XML to encode information, so this transformation makes combining the two protocols easier. We refer to the "XMLized" STARTS version as *STARTS XML*.

Figure 1 shows part of the metadata object that summarizes the contents of the "20 Newsgroups" document collection [2], which is frequently used in the machine learning community. This collection consists of 19,997 articles posted to 20 newsgroups. 34 of these articles contain the term "algorithm" in their body, as the metadata record in the figure indicates. Also, the record shows that this term appears 75 times over these 34 articles. [1] A metasearcher can use these content summaries to select what sources to send a given query (i.e., for database selection). More specifically, a metasearcher can estimate the number of matches that a given query will produce at each source. As a simple example, given a query on "algorithm" a metasearcher might conclude that it does need to contact a source with very few or no documents containing that word, as reported in the corresponding STARTS content summary for the source. (See [12, 18, 24] for research on how to exploit this type of information for database selection.)

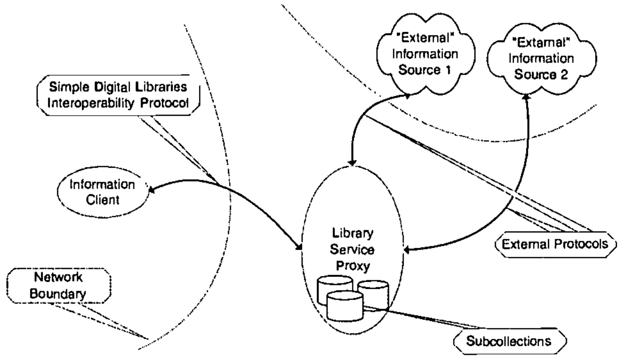

"External" Information Source 1
"Extarnal" Information Source 2
Simple Digital Libraries Interoperability Protocol
Information Client
Library Service Proxy
External Protocols
Network Boundary
Subcollections

Figure 2: The role of SDLIP in a digital library architecture with autonomous sources and wrappers (taken from [20]).

4. SDLIP: SIMPLE DIGITAL LIBRARY INTEROPERABILITY PROTOCOL

SDLIP [20] is a layered protocol that defines simple interfaces for interoperability between information sources. Its designers define it as "search middleware," lighter and easier to use for web-related applications than standard middleware protocols like Z39.50 [1].

The main purpose of SDLIP is to provide uniform interfaces to information sources for querying and retrieving results, and taking care of the transport of the data across the network. SDLIP defines the interfaces that a source or a wrapper of a source should implement so that it can be accessed by an SDLIP-enabled client. An implementation of the SDLIP interfaces is called a *Library Service Proxy (LSP)*, and functions as a lower-level wrapper around one or more underlying collections. Figure 2, borrowed from [20], depicts the role of SDLIP in a digital library architecture. LSPs mask access differences among the information sources in the digital library, and present these sources to clients through a uniform interface. Additionally, SDLIP places one more layer above the LSP, namely network transport, via HTTP/DASL or CORBA, to define exactly how the servers and the clients should communicate with each other. The SDLIP protocol supports both a stateful and a stateless mode of operation. Just as STARTS is a stateless protocol, we decided to use only the stateless version of SDLIP in our project. Often web search over distributed collections does not justify the added complexity of supporting stateful protocols.

In a typical deployment of SDLIP, the LSP is a wrapper that knows how to interact with the underlying collections and exports a uniform SDLIP interface. The LSP interface is divided into three parts: the *search* interface, the *results*

[1] Although XML is quite verbose to describe these statistics, we should note that these XML objects can be compressed effectively [16], and this compression can take place before a potential transmission to a client.

interface, and the *metadata* interface. The search interface defines the operations for submitting a search request to an LSP. The result interface allows clients to access the result of a search. Finally, the metadata interface allows clients to question a library service proxy about its capabilities and contents. For each interface, SDLIP supports, by design, a limited set of operations. This design decision is based on the observation that the SDLIP interfaces can be enhanced as needed through inheritance.

While inheritance is a useful tool for extending the SDLIP interfaces, there is another mechanism for extending the protocol's capabilities. The methods of the three interfaces described above are designed to pass XML documents as their major parameter and return types. The conditions for an SDLIP search, for example, would be specified not by primitive or object-based parameters, but rather as an XML "property sheet" of search parameters. SDLIP's grammar for this XML is extensible, making it easy to host additional metasearching-related features within SDLIP simply as some dialect of XML.

As mentioned above, SDLIP is optimized for clients that know which source they wish to access. In contrast, STARTS specifies, among other things, the metadata that sources should export for metasearching. This makes merging these two complementary protocols desirable. More specifically, we describe next how we instantiate the SDLIP interfaces with STARTS XML to obtain an expressive protocol for effective metasearching.

5. THE SDARTS PROTOCOL

In this section, we describe how we combined SDLIP and STARTS XML into the SDARTS protocol. Section 5.1 outlines our rationale, while Section 5.2 gives an overview of SDARTS.

5.1 SDARTS Design Rationale

SDARTS combines SDLIP and STARTS XML into an expressive protocol for distributed search. We now discuss our choice of these two existing protocols as the basis for SDARTS.

Our decision to adopt SDLIP was influenced by the fact that SDLIP is already in use by other DLI2 projects around the United States, and the ability to interoperate with resources made available by other digital libraries efforts is of course attractive. Additionally, the fact that there are already implementations that allow SDLIP clients to access the contents of sources supporting alternative protocols like Z39.50 [1] (but not vice versa) was another important factor in our choice of SDLIP as our "middleware" architecture. (Other emerging related efforts, notably the "Open Archives Initiative" [19], were still under development and not in stable form when we developed SDARTS.) Finally, we believe that the agreement on common interoperability protocols is a major factor in the success of efforts like the DLI2 projects.

At the same time, STARTS specifically describes the information that should be exchanged between sources and a metasearcher. STARTS includes support to plug in other attribute sets for searching like the Dublin Core [23] and Z39.50's Bib-1 [1] attribute sets. In fact, STARTS built on these efforts and already supports some of their most useful features. Additionally, STARTS specifies the metadata that sources should export for effective metasearching. Other protocols that define related metadata are the GILS Z39.50 profile [9] and Z39.50's Exp-1 [1] attribute sets. Again, STARTS built on these protocols. In particular, STARTS defines that the sources should export vocabulary frequency information (Section 3). SDLIP does not specify this kind of metadata, which is useful for metasearching. However, SDLIP is designed in such a way that it can easily incorporate additional capabilities, which can be exploited by clients that are aware of them. Thus, it is natural to enrich SDLIP with the STARTS components.

We standardized on STARTS XML as the XML "dialect" for SDLIP to exchange the extra information not included in the original version of SDLIP. SDARTS follows all of SDLIP's original DTDs, which include placeholders that can be exploited to include the necessary STARTS XML objects (e.g., the `getPropertyInfo()` method in SDLIP's metadata interface returns the `<propList>` element that can be used for this purpose). Since the vocabulary statistics for a source might be large, and given that a client other than a metasearcher might not need them, SDARTS returns only the URL where the content summary resides as part of the metadata for a source. Then, a metasearcher can download the summary outside of SDARTS with the given URL.

By standardizing on one XML format for SDLIP, we have created an architecture where an LSP is divided into two pieces: a standardized front-end that never needs to change, because it only has to deal with one dialect of XML, and an abstract back-end, which is implemented for specific underlying collection types. Thus, a programmer would need to write only this back-end implementation. Furthermore, this standardization and predictability of the back-end LSP makes it easier for us to create configurable reference implementations of the wrappers for frequently occurring collection types, at the expense, of course, of reduced flexibility.

Another factor that influenced our decisions was our intention to design a software design kit (SDK) for developers to create SDARTS-compliant wrappers. The main design goal was ease of implementation. To be sure, we wanted any programmer to be able to implement our standard wrapper interface from scratch, and create custom wrappers fine-tuned to their underlying collections. But in addition, we wanted to create reference implementations for some common collection types. Moreover, we wanted these implementations to be adjustable *without any additional writing of code* whenever possible. Thus, SDARTS includes a toolkit with reference implementations of wrappers for frequent types of collections. These wrappers are all configurable with text files. These files tell our wrappers things like how to index the documents in a locally available collection, how to transform queries into forms that external collections can understand, and how to translate results passed from collections. Once again, we chose XML as the format for these files. Later, in Section 6 we focus on the various reference implementations, and the pros and cons of our configuration-driven approach.

5.2 Overall SDARTS Architecture

Figure 3 shows the final SDARTS architecture, as it would be used over the Internet to make one wrapped collection available. The client layer is on the left side of the diagram, and the wrapped collection is on the right side.

In the diagram, we see two possible clients: a standard SDLIP client, since SDARTS uses the SDLIP interfaces and transport layer, and the `SDARTSBean`, a client component that we developed that simplifies access to SDARTS and en-

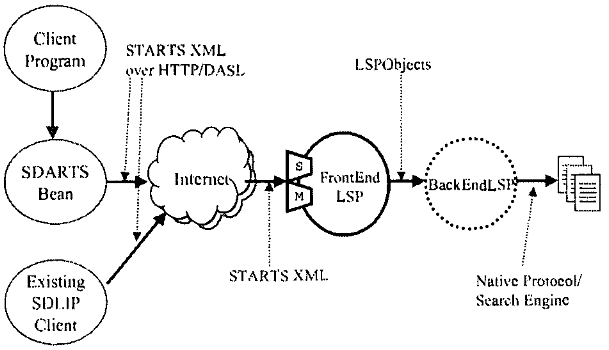


Figure 3: The use of the architectural components of SDARTS to query an SDLIP-compatible database over the Internet. (The "S" and "M" stand for SDLIP's "Search" and "Source Metadata" interfaces, respectively.)

hances SDLIP with the metadata information that can be used by metasearchers. For simplicity, we chose to use the HTTP/DASL protocol as transport layer and not to implement at this point the support for CORBA. In the diagram we show both prospective clients using the HTTP/DASL protocol for the communication. All messages passing between the clients and the top level of SDARTS are formatted in STARTS XML. As we discussed earlier, the benefits of this are:

- XML is an easily parsed and transformed language standard, and is ideal for client use; and
- Lower layers of SDARTS need only be implemented once, as the incoming XML protocol is already known.

The top layer of the server side is a Java component called the FrontEndLSP. This component implements the major SDLIP interfaces, and is for all intents and purposes an SDLIP LSP. The FrontEndLSP parses the incoming STARTS XML requests using the Simple API for XML (SAX), and generates various search objects that implement the common LSPObject interface.

These LSPObjects are just data containers; basically, they are objectified STARTS XML documents, and are passed to the lower layer. They represent the requests and responses that the system fulfills and produces. By standardizing on this object representation of the various SDARTS operations, we simplify the task for the prospective wrapper implementors. They will not need to parse any incoming XML or generate any outgoing XML, but instead they can just examine and create LSPObjects, whose simple interfaces make it easy to understand queries and generate results. Thus, the wrapper programmer needs only to implement the abstract BackEndLSP interface, which accepts and returns LSPObjects. This implementation reads LSPObjects that come from the FrontEndLSP, and generates other LSPObjects that it returns. Each LSPObject already knows how to represent itself as STARTS XML, so it is easy for the front-end to transform them into responses to a SDARTS client. In short, back-end developers do not really need to know anything about STARTS, SDLIP, or even XML in general. All they need to understand is what the back-end interface supports, what the LSPObjects are and how to read them, and, of course, how to handle the collection they are wrapping.

An alternative design choice for SDARTS could have used the Document Object Model (DOM), which provides an existing framework and implementation for converting XML documents into Java objects. While DOM does simplify the task of objectifying XML documents, it does so at the expense of specificity and performance. DOM objects have very generic interfaces. For example, if the BackEndLSP accepted DOM objects instead of LSPObjects, then wrapper programmers would have to remember what to expect when calling each of the generic getter methods on the DOM objects. They might have to make redundant calls in order to ascertain what data was actually available, and would certainly have to memorize what data types to expect. There would be much casting overhead during each interaction. In short, they would have to perform significant extra work to extract the necessary query data. As mentioned above, our overriding design goal was that it be relatively simple to implement the wrappers. As such, we opted to specify our own object representation of the XML documents, LSPObjects, with well-known method names and well-understood types.

The decisions made during the development of SDARTS resulted in a protocol that made the implementation of wrappers a relative easy task, even for a moderately experienced programmer. However, our goal was to facilitate even further the development of wrappers for existing collections. For this reason we have created reference implementations of wrappers for common collection types that can be easily configured, as described next.

6. CONFIGURABLE REFERENCE IMPLEMENTATIONS

In this section, we describe the reference wrapper implementations that we developed for common collection types. To wrap a collection, we can use the closest of these implementations and adapt it by defining simple XML-based files. Currently, we have three reference wrapper implementations:

- A wrapper for unindexed text documents residing in a local file system (Section 6.1);
- A wrapper for unindexed XML documents residing in a local file system (Section 6.2); and
- A wrapper for an external indexed collection fronted by a form-based WWW/CGI interface (Section 6.3).

6.1 TextBackEndLSP: Unindexed Text Document Collections

The first wrapper we created is for locally available collections of documents with no index over them. For this, we leveraged an open-source Java search engine known as Lucene [17] to index and search such collections. In its internal architecture, our wrapper hides the Lucene engine behind a more abstract interface, so in practice other search engines could be used with this wrapper.

Figure 4 shows the basic structure of the wrapper, known as TextBackEndLSP. The back-end administrator needs only write one file: doc_config.xml. This tells the wrapper where

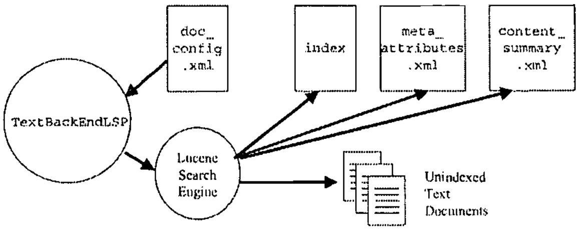


Figure 4: Structure of the TextBackEndLSP wrapper.

```
<doc-config re-index="true">
          <path>/home/dli2test/collections/doc1/20groups </path>
          <linkage-prefix>http://localhost/20groups</linkage-prefix>
          <stop-words>
                    <word>the</word>
                    <word>a</word>
          </stop-words>
          <field-descriptor name="author">
                    <start>
                              <regexp>^From: </regexp>
                    </start>
                    <end>
                              <regexp>$</regexp>
                    </end>
          </field-descriptor> . . . . . . . . .
</doc-config>
```

Figure 5: Portion of the contents of doc_config.xml.

the documents are, what fields should be indexed, and where to find the fields in the documents. TextBackEndLSP uses the file to index the documents offline, so that when the SDARTS server is available, the collection is fully searchable. During indexing, the wrapper also generates the two metadata files, meta_attributes.xml and content_summary.xml; these contain the standard STARTS metadata, and can be returned in response to requests from the front-end. The file doc_config.xml is itself in a very simple XML format (see Figure 5). An important point is the use of regular expressions to tell the wrapper how to extract the values associated with each searchable field like author or title. In our sample file, the author field of a document is found on a line that starts with the string "From:".

This format is simple enough for a non-programmer to edit by hand, and a GUI administration tool can certainly easily generate it. It has some limitations; it assumes that all documents in the collection are in the same format. In addition, field data must be contiguous within a document. This is especially a problem with XML documents, so we created a separate wrapper, described below, to deal with them.

6.2 XMLBackEndLSP: Unindexed XML Document Collections

This wrapper for documents formatted in XML is quite similar to its plain-text counterpart; the key difference is that to index collections of XML documents a slightly more complicated configuration file is needed. Here, the configuration file doc_config.xml only tells the wrapper where to find the documents. The administrator must then write a second file, doc_style.xsl, which is an XSL Transformations (XSLT) stylesheet. It is this second file that tells the wrapper how to find the fields in the documents during the indexing process.

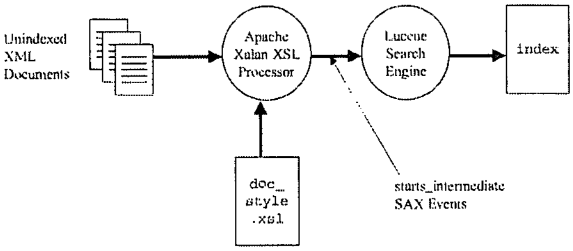


Figure 6: Indexing XML documents for the Lucene search engine, using an XSLT stylesheet to locate the fields in each document.

Figure 6 illustrates the indexing process in this wrapper. The doc_style.xsl file should be written to transform an XML document from the collection into a new XML format we devised called starts_intermediate. This is a simple subset of STARTS XML that describes STARTS documents. The transformed document is never materialized; rather, the transformation emits SAX events that ultimately tell the search engine indexer how to index the document. Like the text document wrapper, this process assumes that all documents in the collection are structured similarly.

When we set out to devise this and the final wrapper, we believed that writing XSLT stylesheets was easier than designing and implementing Java objects; while this is true, XSLT is still non-trivial. In Section 7, we assess XSLT's suitability for making wrappers easy to implement and configure.

6.3 WWWBackEndLSP: WWW/CGI Collections

WWWBackEndLSP is the most complex of the three wrappers. It is intended for autonomous, remote collections fronted by HTML forms and CGI scripts. Such collections include search engines such as Google, AltaVista, and thousands of other web-based searchable collections.

There are two major issues in creating a generic, configurable wrapper for such collections:

- How to convert a query into the proper CGI-BIN invocation onto a search engine; and
- How to interpret the HTML results returned by such an engine.

For metasearching, there is also the question of how to extract metadata from such engines, since most search engines do not provide any such metadata. Our current implementation relies on the wrapper administrator to write a metadata file (the meta_attributes.xml file) with the information specified by STARTS XML. In the future, we could automatically generate at least an approximation of the content summaries by using the results of research on metadata extraction from "uncooperative" sources [4, 15].

We decided that the best way to make this wrapper configurable without additional Java coding was through the use of XSLT stylesheets and the starts_intermediate format. We extended it to be able to describe CGI invocations

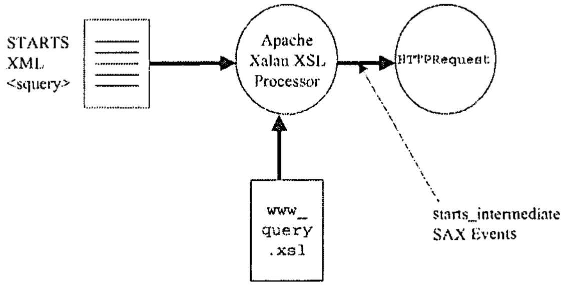


Figure 7: Converting a STARTS query to a CGI request using XSLT.

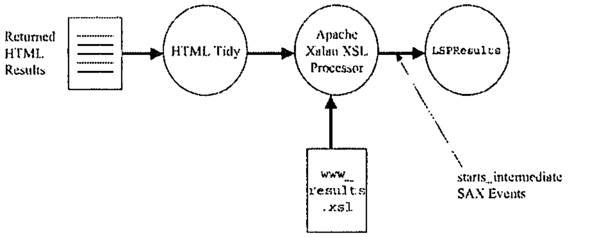


Figure 8: Converting HTML search engine results to a SDARTS object with XSLT.

using an element called `<script>`, which consists of a URL, a method (GET or POST), and a set of name/value pairs that are the script's parameters.

Figure 7 shows how the query is assembled. The administrator first writes an XSLT stylesheet called `www_query.xsl`. This stylesheet is used to map the transformation between a a STARTS `<squery>` and `starts_intermediate <script>`. The result of this transformation is output as SAX events, which ultimately build up a Java embodiment of the CGI invocation called `HTTPRequest`.

Figure 8 shows how the HTML results page returned by the wrapped search engine is parsed. While HTML is a close relative of XML, it is not as well-structured as XML. Hence, it cannot be truly classified as XML, and thus cannot be processed by XSLT. Thus we must first pass it through the HTML Tidy utility [14], which converts the page into well-formed XML (similar but not identical to XHTML). This XML is then transformed using an XSLT stylesheet called `www_results.xsl`, once again written by the administrator. The stylesheet maps the transformation between the XML-formatted results and a `starts_intermediate` result set. Once again, the transformation is output as SAX events, which ultimately build up an `LSPResults` object, which is the standard `LSPObject` subclass returned by a `BackEndLSP` in response to a search query.

Our wrapper is designed to work with search engines that return HTML pages of result records that may have a "more" or "next" button located on them, designed to retrieve further results. This is typical of most, if not all, web-based search engines. The stylesheet can also be written to detect such a button on the search results page, and convert it into a `starts_intermediate <script>` element. The wrapper will then understand that there are more result pages, and will invoke the `<script>` as an additional CGI call. This allows the wrapper to automatically page through a complete set of results from a search engine query.

7. FURTHER DISCUSSION AND CONCLUSIONS

In this paper we have described SDARTS, a protocol that is an instantiation of the SDLIP protocol with metasearch-specific elements from the STARTS protocol. We also reported on a toolkit with wrappers for a variety of heterogeneous collections. All the details (including the source code and the complete documentation of the various SDARTS wrappers that we have implemented) are publicly available at `http://www.cs.columbia.edu/~dli2test/`.

The reference SDARTS wrappers that we have implemented so far are meant to be customized for new collections. To build a wrapper for a new local text collection residing in a local file system, it is straightforward to write the required `doc_config.xml` file (Figure 5). In contrast, we have found that building wrappers for web-based collections and for local XML document collections by writing XSLT stylesheets can sometimes be more involved. XSLT turns out to be quite difficult to master. It is declarative, template-driven, and rule-based, which makes it quite different from the procedural and object-oriented languages most programmers are used to. In addition, writing these stylesheets requires a wide sampling of possible input documents. Many test searches on a web-based collection, for example, should be performed and saved before a wrapper administrator can write a `www_results.xsl` file for it.

The challenge of writing any configuration-driven system is to make it configurable but not so overly complex that writing the necessary configuration files is equivalent to writing a new piece of software. We think that using XSLT in our wrappers meets this challenge, although this is still not a perfect solution. In the end, in our opinion it is still easier than writing and re-writing Java code for parsing HTML or XML, extracting results, and formatting them. Furthermore, it would be easier to create a tool that could generate XSLT stylesheets than it would be to create one that could generate the Java code embodying the same transformation logic. XSLT is an area of active research and development, with a new generation of automatic stylesheet generators, many of them open source, already on the horizon. Therefore, we hope that future versions of our system might include GUI-based wrapper development tools that could simplify the generation of the configuration files.

8. ACKNOWLEDGEMENTS

We thank Sergey Melnik and Andreas Paepcke, from the Stanford Digital Libraries project, for their invaluable help and comments. We also thank Sergey Sigelman and Mike Medric for their contributions to the development and debugging of SDARTS code. Panagiotis Ipeirotis is partially supported by Empeirikeio Foundation and he thanks the Trustees of Empeirikeio Foundation for their support.

This material is based upon work supported by the National Science Foundation under Grant No. IIS-98-17434. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

9. REFERENCES

[1] International Standard Maintenance Agency. *Z39.50 Maintenance Agency Page*. Accessible at http://www.loc.gov/z3950/agency/. ISMA, 2000.

[2] C. Blake and C. Merz. University of California at Irvine repository of machine learning databases. Accessible at http://kdd.ics.uci.edu/.

[3] C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber, and M. F. Schwartz. Harvest: A scalable, customizable discovery and access system. Technical Report CU-CS-732-94, Department of Computer Science, University of Colorado-Boulder, Aug. 1994.

[4] J. P. Callan, M. Connell, and A. Du. Automatic discovery of language models for text databases. In *SIGMOD 1999, Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 479–490. ACM Press, 1999.

[5] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28. ACM Press, 1995.

[6] J. Calmet, S. Jekutsch, and J. Schü. A generic query-translation framework for a mediator architecture. In *Proceedings of the Thirteenth International Conference on Data Engineering*, pages 434–443. IEEE Computer Society, 1997.

[7] C.-C. K. Chang and H. Garcia-Molina. Mind your vocabulary: Query mapping across heterogeneous information sources. In *SIGMOD 1999, Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 335–346. ACM Press, 1999.

[8] C.-C. K. Chang and H. Garcia-Molina. Approximate query translation across heterogeneous information sources. In *VLDB 2000, Proceedings of the 26th International Conference on Very Large Data Bases*, pages 566–577. Morgan Kaufmann, 2000.

[9] E. Christian. Application profile for the government information locator service GILS, Version 2, Aug. 1997. Accessible at http://www.usgs.gov/gils/prof_v2.html.

[10] N. Craswell, P. Bailey, and D. Hawking. Server selection on the World Wide Web. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pages 37–46. ACM, 2000.

[11] L. Gravano, C.-C. K. Chang, H. García-Molina, and A. Paepcke. *STARTS*: Stanford Proposal for Internet Meta-Searching. In *SIGMOD 1997, Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 207–218. ACM Press, 1997.

[12] L. Gravano, H. García-Molina, and A. Tomasic. *GlOSS*: Text-source discovery over the Internet. *ACM Transactions on Database Systems*, 24(2):229–264, June 1999.

[13] D. Hawking and P. B. Thistlewaite. Methods for information server selection. *ACM Transactions on Information Systems*, 17(1):40–76, Jan. 1999.

[14] HTML Tidy. Accessible at http://www.w3.org/People/Raggett/tidy/, 2000.

[15] P. G. Ipeirotis, L. Gravano, and M. Sahami. Probe, count, and classify: Categorizing hidden-web databases. In *SIGMOD 2001, Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2001.

[16] H. Liefke and D. Suciu. XMILL: An efficient compressor for XML data. In *SIGMOD 2000, Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 153–164. ACM, 2000.

[17] The Lucene Search Engine. Accessible at http://www.lucene.com/, 2000.

[18] W. Meng, K.-L. Liu, C. T. Yu, X. Wang, Y. Chang, and N. Rishe. Determining text databases to search in the Internet. In *VLDB'98, Proceedings of the 24th International Conference on Very Large Data Bases*, pages 14–25. Morgan Kaufmann, 1998.

[19] Open Archives Initiative. Accessible at http://www.openarchives.org/, 2000.

[20] A. Paepcke, R. Brandriff, G. Janee, R. Larson, B. Ludaescher, S. Melnik, and S. Raghavan. Search middleware and the Simple Digital Library Interoperability Protocol. *D-Lib Magazine*, 6(3), 2000.

[21] A. Sugiura and O. Etzioni. Query routing for web search engines: Architecture and experiments. In *Proceedings of the Ninth International World-Wide Web Conference*. Foretec Seminars, Inc., 2000.

[22] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird. The collection fusion problem. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 95–104. Department of Commerce, National Institute of Standards and Technology, Mar. 1995.

[23] S. Weibel, J. Godby, E. Miller, and R. Daniel Jr. OCLC/NCSA metadata workshop report, 1995. Accessible at http://www.oclc.org:5047/oclc/-research/publications/weibel/metadata/-dublin_core_report.html.

[24] J. Xu and J. P. Callan. Effective retrieval with distributed collections. In *SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 112–120. ACM Press, 1998.

Database Selection for Processing k Nearest Neighbors Queries in Distributed Environments*

Clement Yu[1], Prasoon Sharma[1], Weiyi Meng[2], Yan Qin[1]
[1]Dept. of CS, U. of Illinois at Chicago, Chicago, IL 60607, yu@eecs.uic.edu
[2]Dept. of CS, SUNY at Binghamton, Binghamton, NY 13902, meng@cs.binghamton.edu

ABSTRACT

We consider the processing of digital library queries, consisting of a text component and a structured component in distributed environments. The text component can be processed using techniques given in previous papers such as [7, 8, 11]. In this paper, we concentrate on the processing of the structured component of a distributed query. Histograms are constructed and algorithms are given to provide estimates of the desirabilities of the databases with respect to the given query. Databases are selected in descending order of desirability. An algorithm is also given to select tuples from the selected databases. Experimental results are given to show that the techniques provided here are effective and efficient.

Keywords

k nearest neighbors, database selection, distributed databases, query processing.

1. INTRODUCTION

As pointed out in [3, 4], it is of great interest to find the k nearest neighbors, i.e., the k tuples in a database table which best match a given user query. If the table contains records that describe library items such as books, magazines and papers, then the problem becomes finding the k best matching items of a given query. Current commercial relational databases systems do not support the processing of such queries. Techniques for processing such queries in a centralized environment have recently been proposed by [3, 4]. In this paper, we examine the processing of these queries in large-scale distributed relational databases or distributed digital libraries, where hundreds or thousands of databases exist. Specifically, given a query which requests the k nearest neighbors from many databases, we propose a method to determine the databases which are likely to contain the desired results. This is of special interest in the Internet environment where there are numerous sites providing data about the same topic.

A straightforward way to process a nearest neighborhood query in a distributed environment is to send the query to all databases. At the site of each database, the k local nearest neighbors are returned. The returned results from the different sites are then merged at a common site to produce the overall k nearest neighbors for the query. This strategy is not efficient if the number of databases is large because most of them probably won't contain any of the desired k tuples. For example, if $k = 20$ and the number of databases is 200, then at least 180 of the databases won't be useful for this query. This straightforward method incurs unnecessary cost to send the query to the useless sites and unnecessary cost to process the query in these sites. Furthermore, when these sites return their retrieval results to the common site, there is further waste of communication and local processing resources.

In this paper, we propose a method to identify the databases which are likely to be useful for processing any given query and to determine the tuples from each useful site which are necessary for answering the query. In this way, both the communication cost and the local processing costs are saved. One common characteristic of these k nearest neighbors queries is that it is not necessary to obtain all the k nearest neighbors; it is often sufficient to get most of the k neighbors. Experimental results are provided to demonstrate that most of the k nearest neighbors (85% to 100%) are obtained using our approach. An average accuracy rate of 94.7% is achieved when the 20 closest neighbors are desired. To the best of our knowledge, this is the first paper on the processing of nearest neighbors queries in distributed relational databases.

EXAMPLE 1. Suppose each Computer Science Department maintains an online technical report database (see [6, 13] for more information about CS technical report libraries). Each database has a table that contains information about each report published by the department. Suppose each report has a title and a publication date, among other possible information fields. Consider the query Q "Find the 10 most similar reports published around 1998 on the topic 'digital library' ". This query has two components. The first component is about the topic "digital library" and is treated as a textual query instead of a character string condition (as a textual query, it can match "libraries of digital video" and even "library" to a less extent). This component

*This work is supported in part by the following NSF grants: IIS-9902792, IIS-9902872, EIA-9911099, CCR-9816633 and CCR-9803974.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

can be matched against the title field with each title being treated as a document. Methods for selecting potentially useful databases for textual queries in distributed environments are given in various papers such as [7, 8, 11]. In [11], an intermediate result of processing such a text query is a ranking of the databases in descending order of similarity. Each database is associated with a similarity (which is an inverse of distance) with respect to the query. The second component of the query is "around 1998" and can be considered as a structured condition. Reports published in 1998 satisfy this portion of the query exactly, with distance = 0. Reports which are published other than 1998 are at some distance away from the query condition; the further the year is away from 1998, the bigger the distance is.

In this paper, we provide techniques for processing this type of query conditions (i.e., structured conditions). Again, databases are ranked in descending order of distance with a distance associated with each database. By combining the techniques used for the above two types of query conditions (i.e., the textual condition and the structured condition), it is possible to process query Q, by ranking databases in ascending order of distance (or descending order of similarity). ■

Note that query Q in the above example can be considered as a typical digital library query, containing conditions against both textual and structured data. Thus, queries used in large-scale distributed digital libraries can be processed efficiently using our method.

The rest of the paper is organized as follows. In Section 2, several examples of the k nearest neighbors queries in relational database systems are provided. In Section 3, methods for determining which databases to search for a given query are provided. Experimental results are given to demonstrate the efficiency and the effectiveness of the methods in Section 4. Conclusion and related works are provided in Section 5.

2. MOTIVATING EXAMPLES OF K-NEAREST NEIGHBORS

In this section, we provide a few examples to illustrate the use of the k nearest neighbors queries. While such queries are widely used in text databases, they are rather unusual in relational databases but are gaining importance. Their interpretations are by no means standard and are application dependent. Due to their different interpretations and applicabilities, we classify these queries into three different types. We also provide "distance" functions which may be suitable for the three different situations.

(a) Standard k-nearest neighbors queries

Given a relation $R(A_1, ..., A_m)$, where the A's are the attributes of the relation and a query $Q(q_1, ..., q_m)$, where q_i is a condition on attribute A_i, $i = 1, ..., m$, a *distance function* d such as the *Euclidean distance* or the *Manhattan distance* can be defined such that a distance $d(Q, t)$ can be computed, where $t = (t_1, ..., t_m)$ is a tuple in R. The distance is a measure on how well the tuple t satisfies the query Q.

EXAMPLE 2. Consider a relation about used cars. Some of the attributes in this relation are price, p, and number of miles driven, m_d. Suppose a user is interested in finding a used car satisfying his requirements on these two attributes. He/she can then submit the following SQL-like query.

Select C.id, C.p, C.m_d
From Used-car C
Where C.p $\approx$ 2000 and C.m_d $\approx$ 100000

There may not be a tuple satisfying the given conditions or there can be too many tuples satisfying the conditions but they are not ordered in a way that the user can easily choose the suitable ones. In the former case, the user's desire to find a suitable car is not satisfied. In the latter case, the user is overwhelmed with too many tuples. The remedy is to have a distance function f such that a distance can be computed between the query, $Q(p \approx 2000, m_d \approx 100000)$ and each tuple t. Then the tuples are ordered in ascending order of distance. Finally, the k tuples having the smallest distances are presented to the user, where k is an integer specified by the user. This may be indicated as follows.

Select C.id, C.p, C.m_d (10)
From User-car C
Where C.p $\approx$ 2000 and C.m_d $\approx$ 100000
Order by Distance f

Here, the 10 nearest neighbors are to be given to the user, where the distance function is f. ■

Although the above example is reasonable, there are rooms for better interpretations. First, if a car has an additional 10,000 miles, it may still fit the user's need. But, if the car costs an additional $10,000, it will definitely not be suitable to the user. This can be remedied by normalizing the attribute value of each tuple by the corresponding query attribute value for each attribute. For example, if a tuple has price and number of miles driven given by $(2.3k, 110,000)$, then the percentage differences in the two attributes are $(0.3k/2k, 10,000/100,000) = (15\%, 10\%)$. For the rest of this paper, we shall use the percentage difference instead of the absolute difference. Another aspect which may better conform to the user's intention is that each attribute may affect the user differently. For example, price may be twice as important to the user than the number of miles driven. Such a desire can be expressed by modifying the "Where condition" in the above query to be

C.p $\approx$ 2000 (I1)
and C.m_d $\approx$ 100000 (I2)

where I1 and I2 are two numbers indicating the relative importance of the two attributes to the user. If I1 = 2 and I2 = 1, then price is twice as important to the user than the number of miles driven.

If the Manhattan distance is used, the "distance" due to the ith attribute, d_i, is given by $|t_i - q_i|/q_i * I_i$ and the overall distance due to multiple attributes is $\sum_i |t_i - q_i|/q_i * I_i$. If the Euclidean distance function is used, then $d_i = ((t_i - q_i)/q_i)^2 * I_i$ and the overall distance is $\sqrt{\sum_i ((t_i - q_i)/q_i)^2 * I_i}$.

(b) **Generalized "distance" k-nearest neighbors queries**

In the above case, a car costing \$1.5k is at the same distance as another car costing \$2.5k relative to the query condition of \$2k, though the former car with the same mileage as the latter is likely to be more desirable to the user. To achieve this effect, the "distance" function is adjusted to permit negative values. For example, the value of \$1.5k has a negative "distance" of $(1.5k - 2.0k)/2.0k = -25\%$; while the value of \$2k has a distance of 0% with respect to the query condition of \$2k. In deciding the nearest neighbors of a query, the tuples are sorted in ascending order of distance, with negative distances before positive distances and highly negative distances before slightly negative distances.

If the Manhattan distance function is used, $d_i = |t_i - q_i|/q_i * I_i$, if $t_i \geq q_i$; otherwise, $d_i = -|t_i - q_i|/q_i * I_i$. If the Euclidean distance is used, then $d_i = ((t_i - q_i)/q_i)^2 * I_i$, if $t_i \geq q_i$; otherwise, $d_i = -((t_i - q_i)/q_i)^2 * I_i$. Again, if the Euclidean distance is used and when all attributes are considered, the distances due to the individual attributes are summed and then the square root is taken. When a "distance" is negative, then the absolute value is taken before the square root is performed and then the negative sign is added back in.

(c) **Two sided generalized "distance" k-nearest neighbors queries**

Suppose we are interested in seeking an airplane ticket from Chicago to New York City. The attributes of interest could be the price and the time of departure. For the time of departure, we may specify a range of time which is acceptable, for example from 3pm to 6pm. Any time within the range will incur a distance of 0, but a time outside the range incurs a positive distance. As indicated before, we are interested in the percentage difference in each attribute. Thus, the denominator of the distance function for normalization due to time is set to be the mid-point, i.e., 4.5pm in the above example. This is to ensure that one hour deviation from either side outside the range incurs the same distance. For example, a 2pm departure time incurs a percentage distance of 1/4.5. Recall that for each attribute there is an importance factor. This can be set to eliminate the effect of where the mid-point lies. For example, the importance factor can be set to be $I_i * m_i$, where m_i is the mid-point of the interval (equal to 4.5 in the above example), and I_i is the relative importance of the ith attribute. With these parameter values, the mid-point m_i will be cancelled out from both the numerator and the denominator.

Let a range (l_i, u_i) be specified for the ith-attribute. Let m_i be the mid-point in the range, i.e., $(u_i - l_i)/2$. If the Manhattan distance is used,

$$d_i = \begin{cases} |t_i - u_i|/m_i * I_i, & if\ t_i > u_i \\ |l_i - t_i|/m_i * I_i, & if\ l_i > t_i \\ 0, & if\ t_i\ is\ in\ (l_i, u_i) \end{cases}$$

If the Euclidean distance is used,

$$d_i = \begin{cases} ((t_i - u_i)/m_i)^2 * I_i, & if\ t_i > u_i \\ ((l_i - t_i)/m_i)^2 * I_i, & if\ l_i > t_i \\ 0, & if\ t_i\ is\ in\ (l_i, u_i) \end{cases}$$

In the remaining part of this paper, we shall concentrate on these three types of nearest neighbors queries. For each type, we shall utilize the "Euclidean distance" function and the "Manhattan distance" function (in the case of the generalized distance neighbors queries, negative "distance" may arise.)

3. DECIDING DATABASES TO SEARCH

Our aim is to retrieve the k nearest neighbors for a given query. This is equivalent to find the k tuples which have the smallest distance from the query. This needs to be done in a distributed environment with many databases. In this paper, we assume that there is one relation in each database and each database is located at a different site. For the remaining part of this paper, "relation" and "database" will be used interchangeably.

To facilitate the identification of the k nearest neighbors for a given query, we store for each relation a *representative* which consists of a histogram for each attribute. These representatives are stored at a central site where user queries are answered (if processing of user queries is desired at every site, then these representatives need to be replicated and stored at every site). When a user query is received, the attributes specified in the query are identified. Based on the histogram on each such attribute, an estimate is made on the desirability of each database with respect to the query. The databases are then ranked with respect to their desirability. Then, they will be searched in descending order of desirability. Tuples from the selected databases are retrieved in such a way that if the databases are ranked *optimally*, then all the k nearest neighbors will be retrieved.

In Section 3.1, the histograms are discussed. In Section 3.2, the criterion for ranking databases optimally with respect to a given query is provided. The criterion is simply that for each database, the distance of the nearest neighbor to the query in the database is obtained and databases are ranked in ascending order of the distances of the nearest neighbors in all databases. This guarantees optimal ranking of databases. In Section 3.3, a generating function is introduced to provide an estimate of the distance of the nearest neighbor for each database. In Section 3.4, an algorithm to determine which tuples from each selected database to be returned to the user is provided.

3.1 Histogram Construction

Two methods for constructing a histogram for each attribute are sketched below. They are the *Simple Interval Construction* method and the *Greedy Merge* method.

(a) **Simple Interval Construction Method**

For each attribute, the range of values is partitioned into subranges of equal width. For each subrange, a frequency count which gives the number of tuples which have values within the subrange is kept. For example, a histogram for the mileages of cars can be as shown in Table 1. If a subrange has too many tuples, then it can be divided into smaller subranges of equal width. For example, the subrange [40k, 50k) may be partitioned into smaller subranges [40k, 45k) and [45k, 50k) and within each smaller subrange, the number of tuples is kept.

Miles	#tuples
[0,10k)	200
[10k,20k)	150
[20k,30k)	170
[30k,40k)	500
[40k,50k)	1000
......	......

Table 1: A Histogram for Car Mileages

From the histogram in Table 1, it can be seen that if there are 10,000 tuples in this relation, then the probability that a tuple with mileage in the range [0,10k) is 200/10000 = 0.02.

(b) Greedy Merge Method

This method was proposed in [10]. Initially, the range of values is partitioned into a large number of subranges of equal width. As in the simple interval construction method, the number of tuples which have values within each subrange is counted. Associated with each subrange, an error of estimation can be computed. For a subrange $[b, e)$, the error of estimation is given by $E = \sum_{i=b}^{e-1}(c_i - ac)^2$, where c_i is the count (the number of tuples) at attribute value i and ac is the average count per attribute value in the subrange. In the Greedy Merge method, the counts within a subrange are approximated by a linear function of the form $s(i) = a_0 + a_1 * i$. It can be shown that the error using the linear function, $E' = (1 - r^2)E$, where r is in $[-1, 1]$ and is the linear correlation between the counts and the attribute values within the subrange. This implies that using a linear approximation function yields a smaller estimation error than using the mean.

The Greedy Merge method merges 2 adjacent subranges with the smallest estimation error. This is repeated until a certain number of subranges is reached. At that point, the counts for the different subranges are kept. If proper statistics are kept for each subrange, then determining which adjacent subranges to be merged can be carried out efficiently. The details can be found in [10].

3.2 Criterion for Selecting Databases Optimally

DEFINITION 1. *Suppose a user is interested in retrieving the k nearest neighbors to a submitted query Q. Databases* $\{D_i, 1 \leq i \leq n\}$ *are optimally ranked in the order* $D_1, D_2 ..., D_n$, *if for every k, there exists a t such that* $D_1, D_2, ..., D_t$ *collectively contain all the k nearest neighbors of Q and each* $D_i, 1 \leq i \leq t$, *contains at least one of the k nearest neighbors.*

The criterion to rank databases is "for each database, obtain the distance of the closest neighbor in the database to the query; then, databases are ranked in ascending order of the distance of the closest neighbor."

EXAMPLE 3. Suppose there are 5 databases D_1, D_2, D_3, D_4 and D_5. Suppose that the distances of the nearest neighbors in these databases to query Q are 0.8, 0.6, 0.3, 0.7 and 0.5, respectively. Then, for query Q, the databases should be ranked in the order D_3, D_5, D_2, D_4, D_1. ■

This guarantees that databases are ranked optimally, as shown in the following proposition.

PROPOSITION 1. *For a given query Q, if databases are ranked in ascending order of the distance of the nearest neighbor to Q, then they are ranked optimally with respect to Q.*

Proof: In [14], a proposition was proved for ranking text databases (i.e., search engines) optimally in the context of retrieving the k most similar documents to a given text query across multiple text databases. The proposition in [14] states that if databases are ranked in descending order of the similarity of the most similar document in each database, then the databases are optimally ranked. The proof of the new proposition can essentially follow that given for the proposition in [14]. The only major difference is that in [14], similarities are used while in this paper, distances are used. Since similarity and distance are inverses, the result holds. ■

3.3 Generating Function to Provide the Estimate

For each attribute, say attribute A_i, specified in the user query, a polynomial is constructed for the attribute of each relation in the distributed database. This polynomial essentially gives the probabilities that tuples in this database are at various "distances" from the query condition specified on attribute A_i. Specifically, let $Q = (q_1, ..., q_i, ..., q_m)$ be the query where q_i is the value of attribute A_i. Let I_i be the importance factor of the attribute in the relation. Then the following polynomial is constructed:

$$g_i = p_1 X^{e_1} + p_2 X^{e_2} + ... + p_s X^{e_s} \qquad (1)$$

where s is the number of subranges of the ith attribute, X is a dummy variable, p_j is the probability that a tuple in the database has a value within the subrange whose mid-point is at weighted distance e_j (given by $d_j * I_i$) away from query condition q_i, where d_j is the "distance" of the mid-point of the subrange from the query condition q_j in the ith attribute. Let the subrange be (l_{ji}, u_{ji}). Then, in the computation of d_j, it is assumed that each tuple in that subrange takes on the mid-point value, i.e., $(u_{ji} + l_{ji})/2$. For example, in the car example, for the first subrange, the mid-point is 5k miles. If the query condition is 3k, then the "generalized distance" due to this attribute is $(5k - 3k)/3k * I_i$, if the Manhattan distance is used. The e's are in ascending order.

It should be noted that the distance d_j from the query condition q_j, which is associated with the subrange, is by no means unique nor most reasonable. Instead of using the mid-point of each subrange, the *mean* can be utilized. In the Greedy Merge method, a linear function is used to estimate the distribution of the attribute values within each subrange. From the linear function, it is possible to estimate the *mean* of the attribute values within the subrange. However, this requires storing the coefficients of the linear function. For simplicity, we use the mid-point of each subrange.

To summarize, for each attribute, say the ith attribute, g_i gives the distribution of the tuples of the relation which are in increasing weighted distances from the query specification on the ith attribute (and due to the ith attribute only). Each tuple is assumed to take on the mid-point value of the subrange where it resides. In the polynomial, for each term involving X, the coefficient of X is the probability that a tuple in the relation is at a weighted distance given by the exponent of X away from the query condition q_i. Suppose that the query has specifications on a set of attributes S. Then, these polynomials are multiplied together to yield $\prod_i g_i$, where the product is over all i in S. This product polynomial, after arranging the terms in ascending order of the exponent of X, gives the distribution of the tuples of the relation in ascending order of weighted distance from the query, taking into consideration all attributes specified by the query. Again, in the case of Euclidean distance, the actual distances should be the square root of the exponents of X. When a "distance" is negative, then the absolute value is taken before the square root is performed and then the negative sign is added back in.

PROPOSITION 2. *If the values of the tuples of a relation R are distributed independently in the attributes specified by the query and each value takes on the mid-point value of the subrange where it resides, then $\prod_i g_i$ gives the probability distribution of the tuples of the relation R in increasing distances from the query, after the product is arranged in ascending order of the exponent of X.*

Proof: Recall that $p_j * X^{e_j}$ in g_i gives the probability of a tuple which is at "weighted distance" e_j from the query condition q_i, due to the ith attribute only. For another attribute, say the kth attribute, $p_t * X^{e_t}$ gives the probability of a tuple which is at "weighted distance" e_t from q_k, due to the kth attribute only. By the independence assumption of the attributes, the probability that a tuple is at "distance" $e_j + e_t$ from the query based on the ith and kth attributes only is $p_j * p_t$. The corresponding term in the product of g_i and g_k is $p_j * p_t * X^{e_j+e_t}$. Thus, after all polynomials associated with the query are multiplied, a term of the form $a * X^e$ gives the probability, a, that a tuple in the relation is at distance e away from the query. (In the case of Euclidean distance, the actual distance is the square root of e.) All terms with the same exponent are added together. After the terms are arranged in ascending order of the exponent, the distribution of the tuples is given in ascending order of distance from the query, since the exponents are the distances. ■

Observations:

1. Usually attribute values are not distributed independently. In our experimental results in Section 4, the attributes are dependent but very reasonable results are obtained.
2. The assumption that each attribute value takes on the mid-point of a subrange is not realistic. However, it does not seem to affect our experimental results significantly.
3. The "distance" function that we use, say the "Euclidean distance" (actually the square of Euclidean distance) or the "Manhattan distance" functions, assumes that the distance is the sum of the weighted "distances" due to the individual attributes specified in the query. This allows us to add the exponents of X to arrive at the total distance.
4. The complexity of the algorithm (to multiply the polynomials) is exponential to the number of attributes which are specified in the query and are needed in the distance computation. Usually, the number of attributes involved in a query is very small. For example, in searching for a car having restrictions on price and mileage, only two attributes are involved. Other specification such as the make or the model of the car are exact conditions and are not involved in distance computations.

EXAMPLE 4. Consider the query Q(miles $\leq$ 10k, price $\leq$ 15k) using the generalized Manhattan distance. Suppose the histogram for mileage is that given in Table 1. Then, the mid-points of the subranges are $5k, 15k, 25k$, etc. The generalized distances of 10k from the mid-points of the subranges are $-5k, 5k, 15k$, etc. After the normalization by 10k, the percentage differences are $-0.5, 0.5, 1.5$, etc. Assuming that there are 10,000 tuples, the probabilities of the subranges are $0.02, 0.015, 0.017$, etc. As a result, the polynomial associated with the mileage attribute is $g_{mileage} = 0.02X^{-0.5} + 0.015X^{0.5} + 0.017X^{1.5} + ...$

Suppose the histogram for price is given by Table 2.

Price	#tuples
[0,1k)	30
[1k,2k)	80
[2k,3k)	50
[3k,4k)	35
[4k,5k)	100
......	......

Table 2: A Histogram for Price

The generalized distances of 15k from the mid-points of the subranges are $-14.5k, -13.5k, -12.5k$, etc. After normalization by 15k, the percentage differences are -0.967, -0.9, -0.833, etc. The probabilities of the subranges are 0.003, 0.008, 0.005, etc. Thus, the polynomial associated with the price attribute is $g_{price} = 0.003X^{-0.967} + 0.008X^{-0.9} + 0.005X^{-0.833} + ...$

The polynomial representing both the mileage and the price attributes is the product of the above two polynomials and is given by

$$0.00006X^{-1.467} + 0.00016X^{-1.4} + ...$$

■

The result of the product polynomial will now be used to estimate the distance of the nearest neighbor in a given relation. Suppose the product polynomial is

$$c_1 * X^{d_1} + c_2 * X^{d_2} + ... + c_k * X^{d_k} \quad (2)$$

where the exponents of X are in ascending order. Let N be the number of tuples of the relation. Then, $N * c_1$ is the expected number of tuples which are at distance d_1 away from the query. If $N * c_1 \geq 1$, then the distance of the nearest neighbor in this relation is estimated to be d_1; else, we compare $N * (c_1 + c_2)$ with 1. If the former is as least as large as 1, then the distance is estimated to be d_2. In general, if t is the smallest integer such that $N*(c_1+c_2+...+c_t) \geq 1$, then the distance is estimated to be d_t.

EXAMPLE 5. Continue on the Example 4. The expected number of tuples having generalized distance -1.467 from the user query is $10000 * 0.00006 = 0.6$. Since it is less than 1, we consider the next term. The expected number of tuples having generalized distance -1.4 from the user query is 1.6. Since $0.6 + 1.6 \geq 1$, the generalized distance of the nearest neighbor is estimated to be -1.4. ■

3.4 Algorithm to Select Tuples from Ranked Databases

Let databases be ranked in the order $[D_1, D_2, ..., D_m]$. Let k be the number of tuples the user wants to see. In order to improve the accuracy, the algorithm retrieves an additional s tuples (i.e. retrieve $k + s$ tuples), but return the k closest neighbors to the user. The databases are accessed in the order in which the databases are ranked, one at a time. (In practice, the first few highest ranked databases may be accessed in parallel, since it is expected that they will contain the desired tuples.) From the first and the second accessed databases, we obtain the nearest neighbor from each of these databases. Let the distances of these tuples be d_1 and d_2, respectively. Let $d = \max\{d_1, d_2\}$. Tuples from these two databases with distances $\leq d$ are gathered. If the number of such tuples is greater than or equal to $(k + s)$, then the k nearest neighbors are returned to the user; otherwise, the next ranked database is accessed. In general, suppose the first t databases have been accessed and d is the maximum value of the distances of the nearest neighbors, one from each of these t databases. If $(k + s)$ tuples have been retrieved, then the k retrieved nearest neighbors are returned to the user; else the next database is accessed. Let d_{t+1} be the distance of the nearest neighbor in database D_{t+1}. If $d_{t+1} > d$, then retrieve the nearest neighbor from database D_{t+1} and tuples which have not been retrieved but with distance $\leq d_{t+1}$ from the first t databases else retrieve from database D_{t+1} tuples with distance $\leq d$. In either case, d is updated to be $\max\{d, d_{t+1}\}$. If the total number of retrieved documents retrieved is $(k + s)$ or more, then return the k nearest neighbors to the user and terminate.

This algorithm guarantees that if the databases are ranked optimally, then all the desired k nearest neighbors of the query will be retrieved. For a proof, see [14], where similarities which are the inverses of distances are used.

4. EXPERIMENTS

In Section 4.1, we describe the data and query collection used in the experiments. In Section 4.2, two measures of retrieval, one reflecting the quality (i.e., accuracy) and the other reflecting the efficiency are provided. Experimental results are provided in Section 4.3.

4.1 Data Collection and Query Collection

Used car data were collected from Excite's Classification 2000 website with the following conditions: Make = "any", Model = "all models", Year = "1900 to 2000", Price = "$500 to $27,000". There are more than 50,000 tuples. The tuples are arbitrarily assigned to 29 databases, without any duplication of tuples. The queries are two attribute queries involving price and mileage. The values associated with the two attributes are chosen to reflect reality. Specifically, if the mileage is high, the price is low; if the mileage is low, the price is high. The relative degrees of importance associated with the two attributes are arbitrarily chosen. 35 queries are generated. For each query, the three interpretations as discussed in Section 2 are applied and they are: the standard k nearest neighbors queries; the generalized k nearest neighbors queries and the two sided-generalized k nearest neighbors queries. For each interpretation, the two "distance functions", namely, the Euclidean distance and the Manhattan distance functions are used.

4.2 Criterion of Performance

Performances are measured in two ways: the quality of the retrieved tuples and the efficiency of retrieval. The former is measured by the number of the tuples which are retrieved and are among the actual k nearest neighbors divided by k. If the quantity is 100%, then all of the k nearest neighbors are retrieved. The higher the percentage, the higher the quality of retrieval is achieved. The second quantity measures the efficiency and is the ratio of the number of databases searched to the actual number of databases containing the k nearest neighbors. If the percentage is 100%, then the number of databases accessed is the same as the number of databases containing the k nearest neighbors although not necessarily the same set of databases is accessed. A value exceeding 100% indicates inefficiency. The lower the percentage, the higher the efficiency is achieved. However, any value in this measure below 100% indicates the quality of retrieval is also below 100%. Ideal retrieval would have both measures equal to 100%.

4.3 Experimental Results

The two methods of constructing histograms, i.e., the Simple Interval Construction method and the Greedy Merge method are employed. For each method, the parameter $s = 20\%k$, i.e., whenever the user wants to obtain the k closest neighbors, an additional 20% of the tuples are retrieved by our algorithm but only k tuples are returned to the user.

# tuples desired	Accuracy	Efficiency
10	0.86	0.94
20	0.90	0.88
30	0.90	0.90

Table 3: Results based on Standard Manhattan distance, Simple Interval

# tuples desired	Accuracy	Efficiency
10	0.86	0.93
20	0.92	0.92
30	0.93	0.94

Table 4: Results based on Standard Manhattan distance, Greedy Merge

Tables 3-14 give the experimental results of the three types of query interpretations, each with the Euclidean distance and the Manhattan distance functions. A summary of the results is given as follows.

1. For the Simple Interval Construction Method, the accuracy ranges from 85% to 98%. For the Greedy Merge

# tuples desired	Accuracy	Efficiency
10	0.85	0.93
20	0.88	0.90
30	0.90	0.90

Table 5: Results based on Standard Euclidean distance, Simple Interval

# tuples desired	Accuracy	Efficiency
10	0.86	0.94
20	0.90	0.92
30	0.91	0.92

Table 6: Results based on Standard Euclidean distance, Greedy Merge

# tuples desired	Accuracy	Efficiency
10	0.93	1.12
20	0.93	1.24
30	0.96	1.06

Table 7: Results based on Generalized Manhattan distance, Simple Interval

# tuples desired	Accuracy	Efficiency
10	1.00	1.00
20	1.00	1.14
30	0.99	0.98

Table 8: Results based on Generalized Manhattan distance, Greedy Merge

# tuples desired	Accuracy	Efficiency
10	0.95	1.05
20	0.95	1.23
30	0.97	1.14

Table 9: Results based on Generalized Euclidean distance, Simple Interval

# tuples desired	Accuracy	Efficiency
10	1.00	1.00
20	1.00	1.21
30	0.99	1.05

Table 10: Results based on Generalized Euclidean distance, Greedy Merge

# tuples desired	Accuracy	Efficiency
10	0.97	1.11
20	0.98	1.12
30	0.97	1.09

Table 11: Results based on Two-sided Manhattan distance, Simple Interval

# tuples desired	Accuracy	Efficiency
10	0.94	1.06
20	0.93	1.04
30	0.92	1.07

Table 12: Results based on Two-sided Manhattan distance, Greedy Merge

# tuples desired	Accuracy	Efficiency
10	0.96	1.08
20	0.97	1.12
30	0.97	1.10

Table 13: Results based on Two-sided Euclidean distance, Simple Interval

Method, it ranges from 86% to 100%. The average accuracy rate for retrieving the top 20 tuples by the Simple Interval method, averaged over the 6 interpretations of the "distance" functions, is 93.5%. The corresponding accuracy rate for the Greedy Merge method is 94.7%. Thus, the Greedy Merge method yields slightly higher accuracy than the Simple Interval Construction Method. However, there are slight deteriorations for the two-sided Manhattan distance and the two-sided Euclidean distance. Clearly, if both the histograms constructed by the Simple Interval Construction Method and the histograms constructed by the Greedy Merge Method are kept, then we can apply the former histogram for the two-sided Manhattan distance queries and for the two-sided Euclidean distance queries and the latter histogram for the other 4 types of queries to yield the best results.

# tuples desired	Accuracy	Efficiency
10	0.94	1.03
20	0.93	1.08
30	0.92	1.08

Table 14: Results based on Two-sided Euclidean distance, Greedy Merge

2. For the generalized and the two-sided queries, the accuracy rates are over 90% and there is not much room for improvement; for the standard queries with the Greedy Merge Method, the accuracy rates range from 86% to slightly above 90%. Thus, it may be desirable to have a 5% improvement.

3. For efficiency, the average worst case is 1.24, meaning that in the worst case an additional 24% of databases need to be accessed. For most situations, the efficiency rates are between 90% and 100%. Thus, there is not much room for improvement.

5. CONCLUSION AND RELATED WORKS

The work reported here are extensions from the following works:

1. It extends the processing of the top-k queries from centralized relational databases [3, 4] to distributed relational databases. We also utilize "distance" functions which are suitable for different applications.

2. It modifies the technique of processing text queries in distributed document databases to be applicable to distributed relational databases. In document processing environment, the number of keywords or terms is very large and usually exact matching of terms is required. In relational databases, the number of attributes in a relation is usually rather small. Two distinct values of the same attribute are separated by a "distance"; the further the separation of the two values the larger the distance. Due to these differences, the "generating function" technique in [12] is modified to be applicable in this environment.

The histograms that we utilize to select databases (sites) to search for a given query are rather primitive. But it has the advantage of being simplistic and space efficient. It may be possible to have slightly higher accuracies by utilizing the linear estimation function within each subrange to estimate the mean of attribute values within the subrange [10] and then use the mean to estimate the distance of a tuple in the subrange from the query condition.

The experimental results provided here show that the methods we employ in retrieving the k nearest neighbors for a given query in a distributed database environment are effective and are efficient. We also sketch how the technique given here and our earlier technique [11] can be combined to process digital library queries involving both text and structured data. Issues regarding the determination of attributes which are semantically the same or related for the purpose of interoperability across databases have been addresses in the literature, see for example [15].

6. REFERENCES

[1] M. Carey and D. Kossmann. *On Saying "Enough Already!" in SQL*, Proc. of ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, May 1997, pp. 219-230

[2] M. Carey and D. Kossmann. *Reducing the Braking Distance of an SQL Query Engine*, Proc. of 24th International Conference on Very Large Data Bases, New York City, August 1998, pp. 158-169.

[3] S. Chaudhuri and L. Gravano. *Evaluating Top-k Selection queries*, Proc. of 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, September 1999, pp. 397-410.

[4] D. Donjerkovic and R. Ramakrishnan. *Probabilistic Optimization of Top N Queries*, Proc. of 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, September 1999, pp. 411-422.

[5] R. Fagin. *Combining fuzzy Information from Multiple Systems*, Proc. of ACM Symposium on Principles of Database Systems, Montreal, Quebec, 1996, pp. 216-2226.

[6] J. French, E. Fox, K. Maly, and A. Selman. *Wide Area Technical Report Service: Technical Report Online.* Communications of the ACM, 38, 4, April 1995, pp. 45-46.

[7] S. Gauch, G. Wang, and M. Gomez. *Profusion: Intelligent Fusion from Multiple, Distributed Search Engines*, Journal of Universal Computer Science, 2, 9, 1996, pp. 637-649.

[8] L. Gravano and H. Garcia-Molina. *Generalizing GlOSS to Vector-Space databases and Broker Hierarchies*, Proc. of 21st International Conferences on Very Large Data Bases, Zurich, Switzerland, September 1995, pp. 78-89.

[9] Y. Ioannidis and V. Poosala. *Histogram-based Approximation of Set-valued Query Answers*, Proc. of 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, September 1999, pp. 174-185.

[10] A. Konig and G. Weikum. *Combining Histograms and Parametric Curve Fitting for Feedback-Driven Query Result-Size Estimation.* Proc. of 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, September 1999, pp. 423-434.

[11] K. Liu, C. Yu, W. Meng, W. Wu and N. Rishe, *A Statistical Method for Estimating the Usefulness of Text databases*, IEEE Transactions on Knowledge and Data Engineering, (to appear).

[12] W. Meng, K. Liu, C. Yu, X. Wang, Y. Chang, N. Rishe. *Determine Text Databases to Search in the Internet.* Proc. of 24th International Conference on Very Large Data Bases, New York City, August 1998, pp. 14-25.

[13] Networked Computer Science Technical Reference Library (http://cs-tr.cs.cornell.edu/).

[14] C. Yu, K. Liu, W. Meng, Z. Wu, and N. Rishe. *A Methodology to Retrieve Text Documents from Multiple Databases.* IEEE Transactions on Knowledge and Data Engineering (to appear).

[15] C. Yu, W. Sun, S. Dao, and D. Keirsey. *Determining relationships among attributes for Interoperability of Multidatabase Systems.* Proc. of the 1st International Workshop on Interoperability in Multidatabase Systems, Kyoto, Japan, April 1991.

Building Searchable Collections of Enterprise Speech Data

James W. Cooper, Mahesh Viswanathan
IBM T J Watson Research Center
PO Box 704, 208
Yorktown Heights, NY 10598
001 914-784-7285, 001 914-945-1754
jwcnmr, mashehv@watson.ibm.com

Donna Byron
University of Rochester
Rochester, NY
dbyron@cs.rochester.edu

Margaret Chan
Columbia University
New York, NY
Maggie@scientist.com

ABSTRACT

We have applied speech recognition and text-mining technologies to a set of recorded outbound marketing calls and analyzed the results. Since speaker-independent speech recognition technology results in a significantly lower recognition rate than that found when the recognizer is trained for a particular speaker, we applied a number of post-processing algorithms to the output of the recognizer to render it suitable for the Textract text mining system.

We indexed the call transcripts using a search engine and used Textract and associated Java technologies to place the relevant terms for each document in a relational database. Following a search query, we generated a thumbnail display of the results of each call with the salient terms highlighted. We illustrate these results and discuss their utility. We took the results of these experiments and continued this analysis on a set of talks and presentations.

We describe a distinct document genre based on the note-taking concept of document content, and propose a significant new method for measuring speech recognition accuracy. This procedure is generally relevant to the problem of capturing meetings and talks and providing a searchable index of these presentations on the web.

Categories and Subject Descriptors

Sound information, Information repositories, Speech information, Evaluation methods, Human-computer interaction, Interface design, Visualization, Concept representation, Document genres, Markup schemes, Metadata, Information retrieval, Multimedia retrieval, Text retrieval.

General Terms

Algorithms, Management, Measurement, Design, Experimentation, Human Factors.

Keywords

Speech analysis, Speech retrieval, Text mining, Search, Document display.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

1. INTRODUCTION

The problem of finding important and relevant documents in an online document collection becomes increasingly difficult as documents proliferate. Our group has previously described the technique of Prompted Query Refinement [7, 8] to assist users in focusing or directing their queries more effectively. However, even after a query has been refined, the problem of having to read too many documents still remains.

We have also previously reported the details of the "Avocado" summarization system we developed for producing rapid displays of the most salient sentences in a document. [13]

Users would prefer to read or browse through only those documents returned by a search engine that are important to the area they are investigating. We have previously described document retrieval systems that can utilize a set of relatively easily derivable numerical parameters to predict which documents will be of most interest to the user. [5]

We now report applying these techniques to data from a speech recognition engine. Since this technology assumes that input is well-edited text, such as articles or news stories, performing this mining on the output of the speech engine represents a new and somewhat complex challenge.

We first describe how we obtained the initial dataset we studied, and then describe the processing necessary to obtain transcripts of these data. Then, we describe post-processing of the speech transcripts and how text mining was carried out on the transcripts. We describe two client server systems and user interfaces we developed for representing these data, and discuss the advantages and limitations of the speech mining techniques. We propose a simple technique for evaluating the accuracy of speech transcripts of these informal conversations.

Finally, we describe two experiments in indexing and summarizing consultant reports and technical talks, and compare our results with those found by human listeners. We describe some of the most promising applications of this system.

2. BACKGROUND

Finding documents in a collection is a well-known problem and has been addressed by any number of commercial search engine products, including Verity, IBM Intelligent Miner for Text and Alta-Vista.

There have been a number of approaches to solving document retrieval problems in recent years. For example, Fowler [10] has described a multi-window document interface where you can drag terms into search windows and see relationships between terms in a graphical environment. Relevance feedback was utilized by Buckley [3] and Xu and Croft, [19] who also utilized local context analysis using the most frequent 50 terms and 10 two-word phrases from the top ranked documents to perform query expansion. Schatz *et al* [16] describe a multi-window interface that offers users access to a variety of published thesauri and term co-occurrence data.

2.1 The Talent Toolkit

Our group at IBM has developed a number of technologies to address these problems. In thisproject, we utilized the suite of text analysis tools collectively known as Talent (Text Analysis and Language Engineering Tools) for analyzing all the documents in the collection. The portions of TALENT relevant to these experiments are described in the following sections.

2.2 Textract

The primary tool for analyzing this collection is **Textract**, itself a chain of tools for recognizing multi-word terms and proper names. Textract reduces related forms of a term to a single *canonical form* that it can then use in computing term occurrence statistics more accurately. In addition, it recognizes abbreviations and finds the canonical forms of the words they stand for and aggregates these terms into a vocabulary for the entire collection, and for each document, keeping both document and collection-level statistics on these terms.

Each term is given a collection-level importance ranking called the IQ or Information Quotient [5, 7]. IQ is effectively a measure of the document selectivity of a particular term: a term that appears in "clumps" in only a few documents is highly selective and has a high IQ. On the other hand, a term that is evenly distributed through many documents is far less selective and has a low IQ. IQ is measured on a scale of 0 to 100, where a value of X means that X% of the vocabulary items in the collection have a lower IQ. Two of the major outputs of Textract are the IQ and collection statistics for each of these canonical terms, and tables of the terms found in each document.

2.3 Context Thesaurus

We have previously described the context thesaurus [7, 8] It is computed from a concordance of all the sentences and occurrences of major terms in those sentences. It is an information retrieval (IR) index of the full text of the sentences surrounding these terms and thus provides a convenient way for a free text query to return terms that commonly co-occur with the query phrase. It also provides an entry point into the collection of terms that actually have been found in the collection, rather than terms that might be predicted *a priori* or using standard dictionaries and thesauri. It is similar to and was inspired by the Phrase-finder [19].

2.4 Named and Unnamed Relations

The Textract system also produces tables of discovered named and unnamed relations. Unnamed relations are strong bi-directional relations between terms which not only co-occur but occur together frequently in the collection. These terms are recognized from the document and term statistics gathered by Textract and by the relative locations of the terms in the document. Named relations [2] are derived by a shallow parsing of the sentences in each document, recognizing over 20 common English patterns which show a named relation between two terms.

3. THE TECHNICAL CHALLENGE

In phase one of this project we worked with a customer in a financial services organization to investigate the effectiveness of analyzing outbound marketing telephone calls with several objectives in mind. It was suggested in preliminary discussions that it might be possible to mine additional information about customers that would be useful in identifying additional financial products that might be of value to them. For example, one could imagine key concepts like "college-age" or "retirement" being used to help assess a customer's financial strategy. Additionally, initial discussion suggested that we might be able to profile the performance of a successful sales call or of a successful salesman by comparing these transcripts with sales success data.

While we began this particular study because of connections to customers, we feel that the overall approach is broadly applicable to any number of libraries of talks and other speech data, such as phone messages, conference calls, and meetings.

Over the course of the project we modified these objectives to ones that while more modest, provided significant benefits to the customer in analyzing the calls.

Data were captured by analog recording of outbound marketing calls for 16 salesmen over a 4-week period. The data were then manually transcribed using a transcription service and used as a model to evaluate the accuracy of speech recognition when applied to these recordings.

The speech recognition system we used was the Large Vocabulary Telephone Transcription system (LVTT), developed from IBM's large vocabulary continuous speech recognition system (LVCSR) [4]. Use of this system in indexing broadcast news has been discussed by Dharanipragada and Roukos [9] and by Viswanathan *et al* [17].

The LVTT system is a large vocabulary speaker-independent system for recognizing and transcribing speech from telephone calls. For this project, the vocabulary was enhanced using terms found on the financial vendor's web site.

Initial results using analog recordings were not promising, and the recordings were repeated using a digital recording system. This system recorded 16 marketing personnel during a single month. To avoid excessive disruption to the customer's business, eight salesmen were recorded for two weeks and the other eight during the remaining two weeks.

This resulted in about 11,000 logical phone call units, including random misfires of the recording system, "he's not here" calls and some interspersed business and personal calls. All of these were processed using the LVTT system and provided as text files for further analysis.

Of these 11,000 calls, we excluded all calls with less than 3K of text as not containing any useful information. This reduced the actual number of calls we analyzed to 523.

4. ANALYSIS OF SPEECH RECOGNIZED DATA

Speech recognized data from telephone calls presents some unique challenges compared to, for example, the high quality PC-based dictation systems (such as IBM's ViaVoice) now available. Not only must the speech recognition be speaker-independent, but it must also deal with a wide variety of accents for both the marketing people and the customers, and significantly reduced audio quality. In this case, the callers and most of the customers called had a wide variety of difficult regional accents, in addition to any number of foreign accents.

Finally, and most significant, telephone conversation is informal speech, consisting of phrases, fragments, interruptions and slang expressions not normally found in formal writing. Thus, the predictive model that speech recognition engines use to recognize which words are likely to come next is much more likely to fail.

Speech recognition systems are built on two models: a language model and an acoustic model. The acoustic model for telephone transcription can help mitigate the reduced frequency spread in the resulting recording. The language model is built on some millions of words found in general writing. It can be enhanced by including domain terms for the area being discussed, in this case, for the financial industry.

However, even with this additional enhancement, the quality of speech recognition is at best 50% of the words in the telephone conversations, and in problematic cases significantly worse. It is these data that we then sought to process and mine into information useful to our research and eventually to our commercial partner.

4.1 Analysis of Raw Recognition Results

As we indicated earlier, the word error rate in these samples of speech recognized data was no more than 50%. Further, the transcripts were not divided either by speaker or even by sentence. A typically problematic transcript fragment is shown below:

that has sweat what you have a minus for the one year before that you you look have all along are right you feel that has performed for you right-now one term I would say average before if there's I would still say to go over average top ten we what you what his you consider that man yeah time you want my name and then I'm-sorry a blue-chip fund number's one because the middle bond fund over ten year period has returned an IRA over a five year period of five point eight are

While there is clearly information in this fragment, we can see that it is going to be extremely difficult for text mining technologies to pull out useful concepts from such data.

Accordingly we developed a number of algorithms to process these data further before submitting them to the Textract text mining and search engine indexing processes. Text mining assumes well-edited text, such as news articles or technical reports, rather than informal conversation, inaccurately recorded.

Much of the post processing analysis we performed on these call transcripts was driven by Textract's requirements of well-edited text in sentences and paragraphs. Textract uses these boundaries to decide whether it can form a multiword term between adjacent word tokens and how high the level of mutual information should be in determining co-occurring terms.

4.2 Timing Information

We first used the timing information in the raw speech data to insert periods and paragraph breaks in the text stream. While the speech recognition engine provided estimates of these points, we were able to fine-tune this process by applying our own empirically derived parameters. In this suite of calls, we replaced pauses of between 0.80 seconds and 1.19 seconds with a sentence break. Specifically, we added a period, two blanks and capitalized the following word.

We replaced pauses of 1.2 seconds or more with a new paragraph, by adding a period, two blank lines and a capital letter to the next word. Paragraph boundaries were important in this analysis because speaker separation information was not available in this research version of the voice recognition engine, and in mining text for related terms, paragraph boundaries provide a break between sentences that reduces the strength of the computed relationships between terms.

The speech engine provided silence information as a series of "silence tokens," where each one was assigned a duration. Frequently, there would be several sequential silence tokens, presumably separated by non-speech sounds. When this occurred, we summed the silence tokens to a single token that we used to determine whether to insert punctuation.

4.3 Word Certainty

The LVTT speech engine provided us with estimates of the certainty it had recognized a word correctly. It also provided a series of alternate choices that it had considered less likely.

Our initial theory was that if we examined the words and alternates as a continuous matrix, we might be able to recognize multi-word terms among the words and alternate choices to improve the quality of recognition. This is similar to the procedure suggested in TREC-8 by Johnson [11]. However, our analysis showed that considering these alternate choices and looking for phrases among them did not result in any improvement at all. In fact, we did not discover any cases where choosing word alternates would improve recognition. This is probably not surprising, because the predictive speech models used in recognition already take these facts into account.

We also investigated document expansion using a parallel corpus in a manner similar to that suggested by Woodland [18]. We indexed a number of well-formed business documents describing the customer's business and products. Then we queried the index using each of the transcribed calls and augmented the transcribed calls with the few documents returned from this query. The purpose of this exercise was to overcome word recognition errors to try to improve recall. We found a very small improvement and decided that it was not sufficient to warrant the large amount of extra computation to achieve it.

We did find the certainty figures useful in the call analysis in another significant way. If the speech engine indicated that a word was recognized with low certainty we considered whether eliminating the word would provide a more useful transcript. This became important because the speech engine tended to insert proper nouns for words it recognized with low certainty and these nouns were frequently incorrect. When the Textract text-mining system is run on such text, these proper nouns are recognized as salient when they in fact should not have been found at all.

Our initial impulse was to remove these low confidence terms from the transcripts entirely, but this would lead to the text-mining system forming multi-word terms across these boundaries when it should not have been able to do so. So instead, we replaced each occurrence of a low confidence terms with the letter "z." These non-word tokens prevented the formation of spurious multiwords without significantly reducing the clarity of the transcript.

For example, in one call we found the sentence fragment:

Over a five year period has returned an IRA...

where the final word "IRA" was of low certainty. Our algorithm converted that to

Over a five year period has returned an z...

which removes the incorrectly recognized token "IRA" that was not in fact part of the originally spoken sentence, but would lead to a wildly misleading summary of the document. (In fact the correct phrase was "returned a 6 and one-half percent...")

The speech engine also produced tokens for non-word utterances such as "uh," "um" and <smack> which we removed entirely. Removing these was actually quite necessary, since they often interfered with the formation of multi-word terms, so that *bond <uh> funds* was reduced to *bond funds*.

5. USING LINGUISTIC CUES

In addition to our reanalysis of the data provided by the speech engine, we also realized that there are some English language cues we can use to improve our recognition of sentence boundaries. There are a number of common English words and phrases which are used exclusively or primarily to start sentences, such as *Yes, OK, Well, Incidentally, Finally* and so forth. In consultation with other linguists in our group, we tabulated a list of these words and phrases and then used them to further process our transcripts. Whenever we found such words or phrases, we inserted a period, two spaces and capitalized the token we found. In these two-way conversations, we found that in most cases, we could insert a paragraph boundary as well. Applying all of these techniques to the initial text we showed above, gives the somewhat more coherent transcript below.

Yeah, that has sweat what.
You have a minus for the one year before that you look have all along are.
Right, you feel that has performed for you right now one Term I would say average before if there's I would still say to go over average top ten we what you what his you consider that man.
Yeah, time you want my name and then I'm sorry a blue chip fund number's one because the middle bond fund over ten year period has returned an z. Over a five year period of five point eight.

6. TRANSCRIPT ANALYSIS

Once we performed all these analyses and post-processing techniques, we still had very poor and confusing transcripts to deal with. Considering the technological barriers we had to overcome, this is not entirely a surprise. However, the question then arose as to whether there was any value at all in such transcripts. It turns out that there definitely is a great deal of information in these documents once we overcome the idea that we are producing transcripts of conversations.

If instead we consider the idea that we are actually producing notes on a meeting to help in summarizing what transpired, we find that there is real value to be extracted even from these noisy data. For example, even from the intentionally vague selection quoted above, we find that we can extract the concepts

- Minus for the one year
- Average top ten
- Blue chip fund
- Middle bond fund
- Ten year period
- Five year period

From those concepts we can begin to outline the nature of the conversation, even without accurate speech recognition.

6.1 Indexing the Transcript Files

Once the transcripts were processed as we described above, we analyzed them using Textract and indexed them using a standard search engine indexing system.

We found that in these unstructured conversations, Textract was not able to form any significant named relations and only a very few unnamed relations, so we did not use this information in constructing our retrieval system.

As we have described previously [13], we used a set of Java programs to analyze the Textract output and load it into a database. From this database, we can easily ask for the most salient terms in any document, and can even restrict the query to the most salient multiword terms in the document.

For example, for the complete conversation we excerpted above, the most salient terms are shown in Table 1. Textract categorizes terms into eight types. In this table and in similar queries in this research, we excluded terms assigned to the Unknown Name (Uname) and Unknown Term (Uterm) categories, which produced a large quantity of fairly uninformative terms like "room."

Table 1- Terms discovered in a single conversation using Textract

Term	IQ
Middle bond fund	85
Chemical bond fund	85
Negative point	85
Strategic income	50
Ginny Mae	47
Percent return	40
Year period	24
Core bond	16
Tax exempt	6
Capital gain	3

7. A CLIENT-SERVER CALL QUERY SYSTEM

After constructing these indexes, we were able to construct a Java-based client-server search system.

The server consists of a search engine index, and a document and terms database. Here the search engine was initially IBM's TSE search engine, later replaced with IBM's GTR search engine, and the database was DB2. The system is driven using a Java RMI server and communicates with a Java RMI client running in a web browser, using the Java plug-in. This system is illustrated in Figure 1.

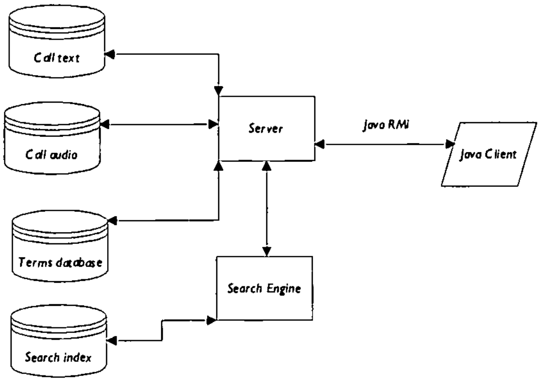
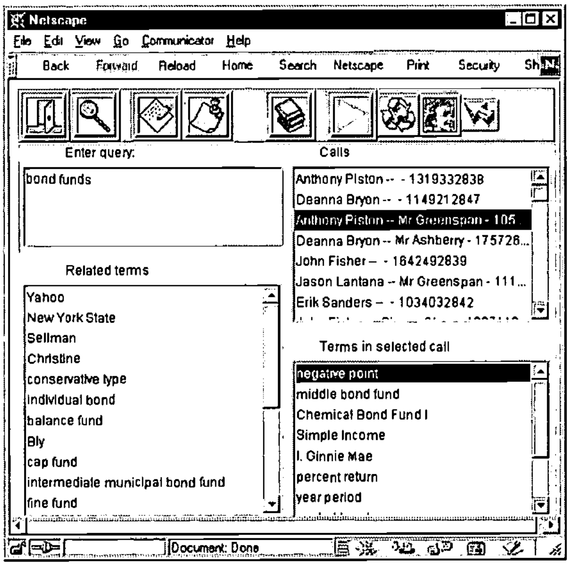


Figure 1- The Java client-server search system.

Figure 2 shows the Java client in action. After the user types in a query in the upper left entry field, the client sends the query to the server. The server returns a list of related terms from the Context thesaurus index, and a list of call titles from the document search index. Clicking on a particular call brings up a list of the terms in that call in the lower right list box.


Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Sh
Document: Done

Enter query:

bond funds

Related terms

Yahoo
New York State
Sellman
Christine
conservative type
individual bond
balance fund
Bly
cap fund
intermediate municipal bond fund
fine fund

Calls

Anthony Piston -- - 1319332838
Deanna Bryon -- - 1149212847
Anthony Piston -- Mr Greenspan - 105
Deanna Bryon -- Mr Ashberry - 175728...
John Fisher -- - 1642492839
Jason Lantana -- Mr Greenspan - 111...
Erik Sanders -- - 1034032842

Terms in selected call

negative point
middle bond fund
Chemical Bond Fund I
Simple Income
I. Ginnie Mae
percent return
year period

Figure 2 – The Java search system, showing the query, the related terms, the call list and the terms in that call. The names and a few of the terms have been modified to preserve confidentiality.

7.1 Displaying the Call Documents

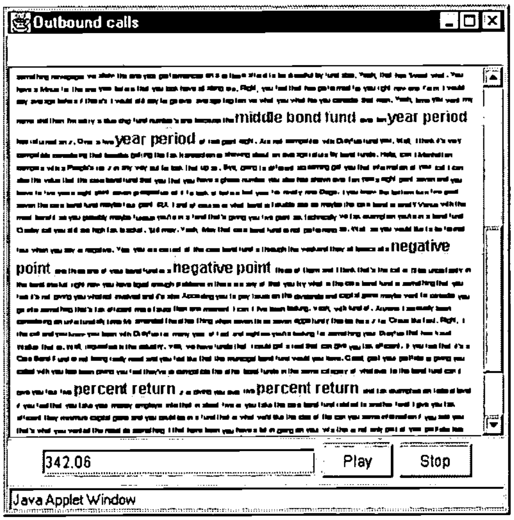
Once a user selects a specific call to investigate, it is important to present the call in the most meaningful way possible. Remember that most of the words in the call are inaccurate. However, we are fairly confident that the multi-word terms that Textract recognizes are likely to be correct. Thus, we constructed a system in which the server looked up the most salient terms in the transcript and marked them as word-objects in the data stream sent to the client. The client could then display these to best benefit the user.

We selected the terms in three stages until a large enough number of terms were found. This depended on the length and conversational density of the call transcription.

1. Select all multiword terms having an IQ >50.
2. If there are less than 10, select all single and multiword terms with an IQ >50.
3. If there are still fewer than 10, reduce the IQ limit to 30 and select all terms above that level.
4. Then, convert the word stream into individual token objects, and look for case insensitive matches within one character (to allow for plurals) of each successive word in the term.
5. Combine each located multiword term into a single token and mark it as salient.
6. Send the entire object stream to the client for display.

After some experimentation, we arrived at a client display in which the font of the terms not recognized as salient by Textract are shown as small as possible, so that they cannot easily be read, but only indicate the sequential position of words in the call.

We display the words that were found to be salient in a larger font, with a contrasting highlight color as shown in Figure 3. This client display was written in Java using the Java Swing JTextPane component, along with the DefaultStyledDocument, Highlighter and Highlight-Painter objects. We have described how to program these somewhat obscure components in another article [Cooper, 1999].



342.06

Figure 3- The document display and playback system

The display shown in Figure 3 illustrates a new way of looking at documents. Rather than just presenting a list of the salient terms in the document, this display shows the logical progress of the phone call and the points at which the important terms are discussed. These highlighted terms represent a kind of note taking on the actual phone call, and in a display represent a new kind of document genre, in which only the major terms are displayed to the users, but in which the viewer can deduce a general time line and grasp the progress of the call.

8. CALL PLAYBACK

Each of these calls was provided to us as digital files in PCM audio format. We converted these to the Sun au format using the copyaudio utility available from McGill University [McGill, 1999]. It was then possible to play back these files from the Java client server system. The Java client could request the call audio file from the server and receive it for playback on demand.

Since the JTextPane component can respond to clicks anywhere on its surface, and since we can calculate which word is clicked on, it is possible to construct a playback system in which the position of the word in the text stream can be recognized. Since word timing information was provided to us by the original speech engine output, we constructed each salient word object in the data stream to include this timing information. Then, the click on any word can be converted to its time offset in the call audio file. The time of a selected term is shown in the text field at the bottom of the window in Figure 3.

In order to play back the audio data, we used the Java Media Framework (JMF) to provide playback from a given time offset. The Java client JMF playback control requests the audio file from the web server and plays it when it arrives. While this is not "streaming audio," the time to download an entire audio file is short enough that the pause before the audio playback begins (even from the middle of a call) is quite acceptable when attached to a relatively high-speed network. Audio playback from a dialup connection was not successful.

The playback data window display is shown in Figure 4. It also contains the same JTextPane as well as the JMF audio player component.

8.1 Using JavaServer Pages for a Light Weight Display

While the Java client we described above is extremely powerful and flexible, it did not meet everyone's needs when restrictions of various browser types and Java RMI issues through firewalls were considered. Accordingly, we developed a second, lighter-weight client interface controlled by modified server code configured as a Java Bean that operated in conjunction with a JavaServer page [Pekowsky, 2000].

The JSP page makes Java calls to the Bean classes to obtain the context thesaurus and call information and fills two list boxes as shown in Figure 5.

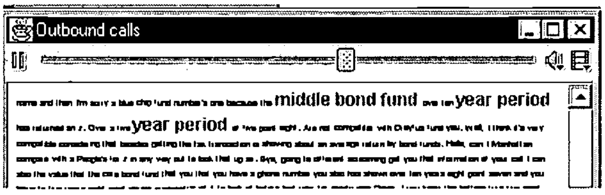
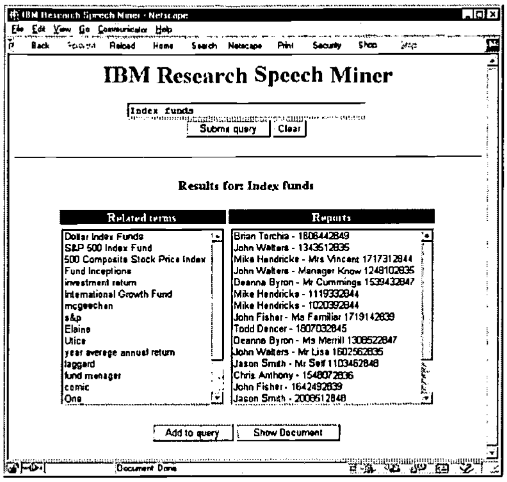


Figure 4 – The JMF Playback Component combined with the JTextPane document display window.

It was also necessary to write a lightweight playback window that could work in this environment. Using DHTML and style sheets for highlighting we were able to devise a playback client which played the audio files using a Java 1.1 audio player. Highlighting of spans in DHTML was accomplished by statements like

```
<span class="yellow">
<a href="javascript:void 0" onClick=
"playSound(230.32)"> stock </a>
</span>
```

Clicking on a highlighted area starts the Java audio player at that point in the sound file. The playback client is shown in Figure 6.


IBM Research Speech Miner - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Submit query
Clear
Add to query
Show Document
Document Done

IBM Research Speech Miner

Index funds

Results for: Index funds

Related terms	Reports
Dollar Index Funds	Brian Torchia - 1806442849
S&P 500 Index Fund	John Walters - 1343512835
500 Composite Stock Price Index	Mike Hendricks - Mrs Vincent 1717312844
Fund Inceptions	John Walters - Manager Know 1248102835
investment return	Deanna Byron - Mr Cummings 1539432847
International Growth Fund	Mike Hendricks - 1119332844
mcgeechen	Mike Hendricks - 1020392844
s&p	John Fisher - Ms Familiar 1719142839
Elaine	Todd Dencer - 1807032845
Utica	Deanna Byron - Ms Merrill 1308522847
year average annual return	John Walters - Mr Lisa 1602562835
laggard	Jason Smith - Mr Self 1103462848
fund manager	Chris Anthony - 1548072836
comic	John Fisher - 1642492839
One	Jason Smith - 2008512848

Figure 5 – A Lightweight client JavaServer page. Again, some names and terms have been slightly modified.

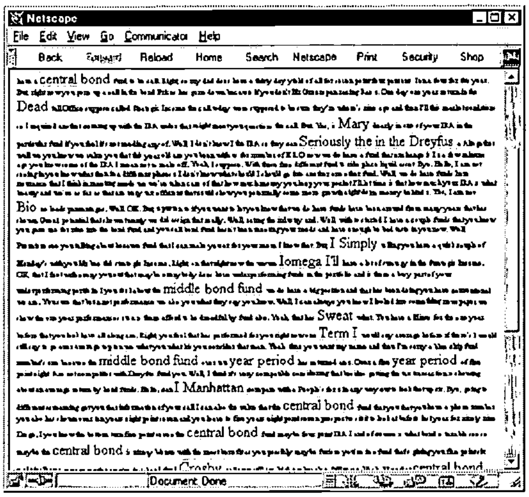


Figure 6 - A Lightweight playback client using DHTML. Some terms were modified as before.

9. TALKS AND REPORTS

Once we developed this process for indexing and a method of displaying the results that we found adequate, we applied the procedure to two new domains of discourse - consultant "notes from the field" and meetings. We accumulated data from two sources: dictation into portable recorders, and talks at a digitally recorded conference on knowledge management technology . Our two objectives: were to find out if this system could be used to provide a way for e-Business consultants to make reports on customer engagements, and to find out if meetings could be summarized. automatically.

9.1 Consultant Reports

We obtained several pocket digital recorders, (Olympus DS-150) and provided them to a series of IBM marketing and consulting people to record their reports in an experimental fashion. The idea behind this experiment was that consultants feel that their primary responsibility is to interact with customers and do not feel they are rewarded for writing reports on these interactions. Thus, it was felt, important knowledge that might be of benefit to consultants in related engagements was lost.

The proposed scenario was that consultants would record these reports by speaking into the portable voice recorder after leaving the customer, and upload them to a web site we provided where we could produce a searchable database of these reports for other consultants.

Our technical findings were encouraging. The Olympus recorders are shipped with IBM ViaVoice and software for training the recognition system on your digitally recorded voice. Consultants, even with foreign accents, were readily recognized once they read a 100-sentence training script into the recorder. We generated an model for each consultant and processed their data. Following processing we created web pages much like those shown in Figure 6, and created a searchable index of the reports.

A segment of a typical report is shown below. (We note that "jowl one" actually refers to JavaOne.)

Activity report for July of the technology is marketing. One jowl one to tell that from June third to June ninth was made successful brief highlights an estimated 27 thousand people saw IBM sessions and DOS and jowl one. Of folks download trebled during that period over 7 hundred expanded IBM hospitality went. Over 5 z z taken from winnable PC initiative. 20 + articles

However, it developed that while this solution was technically quite feasible, the social aspects of this system were not at all what the sponsoring managers hoped. Even though both managers and consultants were quite enthusiastic about this system in theory, we received only a handful of reports over several months of the trial. We concluded that even though the consultants would like to have been able to browse a database of such reports, they had no impetus to help create this database, for the same reason they did not feel they had time to create written reports: it simply was not an important part of their job assignment.

So, even though we were easily able to create such a system, we found that it was quite difficult to interest the relevant practitioners in using it. There is a very important message here: technical systems will only be adopted by users if they are perceived not to impede their work flow, and if they are in some way rewarded for using them. In this case the rewards were too nebulous to justify the consultants' participation.

9.2 Conference Analysis

In our most recent experiment, we undertook the analysis of an internal marketing conference of knowledge management products. The conference was recorded on digital videotape. Each participant was given a lavaliere microphone to wear when they were the primary speaker.

We segmented the audio portion of he video tapes into individual presentations, converted the data into wave file format and used ViaVoice with a speaker-independent acoustic model to produce a transcript. For most speakers, the results were quite helpful. Figure 7 shows the results of one such presentation.

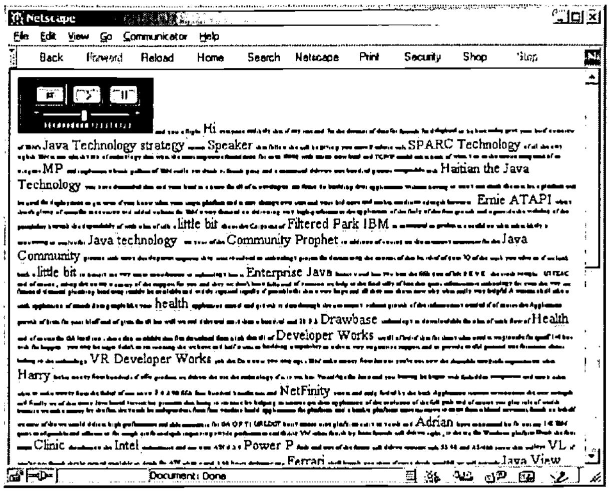


Figure 7 – A speech-recognized document from an internal knowledge management conference.

10. MEASURING ACCURACY USING SALIENT TERMS

If you look at the absolute word recognition accuracy of these transcriptions, you would find it to be as low as 20-30% in some cases, which at first seems a depressingly low number. This low accuracy is a result of the informal speech, incomplete sentences, regional accents and careless use of grammar and even pronunciation that frequently occur in telephone calls.

We selected the particularly problematic document described above for further analysis, and carefully transcribed the actual conversation, omitting stalling sounds such as "um" and "uh," and corrected for or omitted any overlapping dialog. We then ran this document through Textract as part of the collection of the remaining 522 documents, so that the same canonical forms could be developed in term recognition for this document.

After loading the results of Textract into our DB2 database, we compared the terms found in the recognized and manually transcribed documents. The results are shown in Table 2.

Table 2 –Recognition of Multiwords in a manually transcribed and automatically recognized document.

	manual	Automatic	%
Multiwords found in document	22	12	55
Multiwords less spurious finds	19	12	63
Multiwords where singles found	19	13.5	71

Of the several hundred terms found in both documents, we tabulated the multiword terms found in both. For the most part, these multiwords represent the high IQ salient concepts that are the highlighted terms in our display, and serve as a thumbnail outline of the conversation. There were 22 such terms in the manually transcribed document, where the opportunities for forming words correctly without intervening noise were greatest. Of those, 12 were found in the speech-recognized document.

However, once we eliminate multiwords caused by Textract "misfires" and the customer's name, which the recognition engine could not be expected to get, the correct rate rises to 12 out of 19.

Finally, if you give half-scores if all words of the multiword are found individually, we get 13.5 out of 19 or 71% accuracy in finding the salient concepts in the document.

Thus, we find, that even when very difficult conversations are subjected to speech recognition, the speech engine is quite capable of finding the preponderance of the salient terms in a document, and text mining systems like Textract are very capable of extracting these concepts and using them to provide note-like summaries of the major concepts in those conversations. In fact, even with extremely problematic recognition, caused by careless speakers and difficult regional accents, the ability of the speech engines to provide valuable document summaries remains very strong.

10.1 How Relevant Are the Terms?

Our final experiment deals with the question of the quality of the terms we are able to mine from speech transcripts. As outlined above, we start with raw speech data, process the output to detect sentence boundaries, and run the Textract text mining engine to find the main multi-word terms in the collection, and in each document.

The question we still needed to answer was whether the digital "notes" these documents represented were anything like the terms human subjects would find if asked to take notes on the same speech. If there is substantial overlap between the machine recognized multi-word terms and those found by the "note-takers," we can consider that the system is useful.

To test this hypothesis, we asked 10 subjects to watch an 18-minute segment of video from the meeting and take notes of the important concepts. We specifically asked for a note-taking style of lists of terms, to make sure that they used an approach that was similar to our automated system and similar to each other. In this experiment, we used the commercial version of IBM ViaVoice Millennium Edition, and an readily available TCL/TK toolkit for extracting the timing information from the ViaVoice data.

The results were quite encouraging. When we compared their results with the multi-words found by Textract on the voice-recognized transcript of the same 18-minute call, 17 keywords were found by all 10 human note-takers. These results are summarized in Table 3.

Table 3 – Terms found automatically and by 5 or more human note-takers in an 18-minute segment of video.

Phrase recognized by	Number of multi-words
Textract + 10 subjects	17
Textract + 9 subjects	25
Textract + 8 subjects	86
Textract + 7 subjects	91
Textract + 6 subjects	97
Textract + 5 subjects	102

We note that traditional measures of precision and recall are difficult to correlate with these data, because there is no agreed-upon way of determining the complete correct list of terms: it is subjective both from the human subject and the computational point of view.

From these preliminary experiments, we conclude that these keyword-highlighted summary representations of speech-recognition output created by our system are sufficiently accurate to represent a useful set of meeting notes that can be searched and displayed for use in playback of such presentations.

11. RESULTS AND DISCUSSION

In these experiments, we discovered that while it is not yet possible to produce word-accurate transcripts of informal conversations. reports and meetings, it is still possible to provide extremely useful information regarding the content of this voice data. Rather than regarding the recognized text as a transcript, it is more useful to consider it a beginning of a set of "notes" on the event such as a party might take down to remind themselves later of what was said.

We found that by recognizing the salient terms in the conversation and providing a search system for searching the call or report archive, we can provide supervisors and other consultants with a way of looking into what information was discussed, and a way of playing back the interesting portions of conversations returned from such a search.

Finally, we found that while the word recognition accuracy of these transcripts was in many cases fairly low, the salient term accuracy was quite high and made these searchable summaries extremely useful.

We note here that while the Textract tool we used here is an internal research prototype, there is a product version available and that commercial products from other vendors may also be used. In general, we have found that the Textract tools is more efficient in aggregating variant forms of terms than most of the current commercial systems, but for this particular application, they might well be considered roughly equivalent.

12. ACKNOWLEDGMENTS

We'd like to thank Zunaid Kazi for his work in developing the unnamed relations and context thesaurus code, Eric Brown for recording the conference on digital videotape, Branimir Boguraev for helpful suggestions which led to the development of the client display, Yael Ravin for modifications to Textract to make name detection more reliable in this environment, Paul Kaye for writing the media player interface, Bob Mack for helpful general suggestions, Salim Roukos for helpful conversations, Harry Kolar for thoughtful analysis of the results, Michael Hehenberger for his work in developing the project, and Roy Byrd and Alan Marwick for their technical support.

13. BIBLIOGRAPHY

[1] Brandow, Ron, Karl Mitze, and Lisa Rau, 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31, No. 5, 675-685.

[2] Byrd, R.J. and Ravin, Y. Identifying and Extracting Relations in Text. *Proceeedings of NLDB 99*, Klagenfurt, Austria.

[3] Buckley, C., Singhal, A., Mira, M & Salton, G. (1996) "New Retrieval Approaches Using SMART:TREC4. In Harman, D, editor, Proceedings of the TREC 4 Conference, National Institute of Standards and Technology, Special Publication.

[4] Chen, Scott Shaobing, M.J.F. Gales, P.S. Gopalakrishnan,

R.A. Gopinath, H. Printz, D. Kanevsky, P. Olsen, L. Polymenakos, IBM's LVCSR System for Transcription of Broadcast News Used in the 1997 Hub-4 English Evaluation in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998

[5] Cooper, J.W. and Prager, John M, *Anti-Serendipity: Finding Useless Documents and Similar Documents,* Proceedings of HICSS-33, Maui, HI, January, 2000.

[6] Cooper, J.W., "Colorful Language," JavaPro 4(1), 44, 2000.

[7] Cooper, James W. and Byrd, Roy J. "Lexical Navigation: Visually Prompted Query Expansion and Refinement." Proceedings of DIGLIB97, Philadelphia, PA, July, 1997.

[8] Cooper, James W. and Byrd, Roy J., OBIWAN – "A Visual Interface for Prompted Query Refinement," Proceedings of HICSS-31, Kona, Hawaii, 1998.

[9] Dharanipragada, S. and S. Roukos, Experimental Results in Audio Indexing in Proceedings of the DARPA Speech Recognition Workshop, 1997

[10] Fowler, Richard H., Wilson, Bradley A., and Fowler, Wendy A.L. "Information Navigator: An information system using networks for display and retrieval." Report NAG9-551, No.92-1. Department of Computer Science, University of Texas, Pan American, Edinburg, TX.

[11] Johnson, S.E., Jourlin, P., Spark-Jones, K. and Woodland, P.C. *Spoken Document Retrieval for TREC-8 at Cambridge University,* Proceedings of TREC-8, 1999.

[12] McGill University FTP Archive. See ftp.TSP.ECE.McGill.CA/pub/Afsp

[13] Neff, Mary S. and Cooper, James W. 1999a. Document Summarization for Active Markup, *in Proceedings of the 32nd Hawaii International Conference on System Sciences,* Wailea, HI, January, 1999.

[14] Pekowsky, Larne. *JavaServer Pages*, Addison-Wesley, Boston, MA, 2000.

[15] Prager, John M., Linguini: Recognition of Language in Digital Documents, *in Proceedings of the 32nd Hawaii International Conference on System Sciences,* Wailea, HI, January, 1999.

[16] Schatz, Bruce R, Johnson, Eric H, Cochrane, Pauline A and Chen, Hsinchun, "Interactive Term Suggestion for Users of Digital Libraries." *ACM Digital Library Conference, 1996.*

[17] Viswanathan, M, H.S.M. Beigi, S. Dharanipragada, and A. Tritschler, "Retrieval from Spoken Documents Using Content And Speaker Information," Proceedings, International Conference on Document Analysis and Retrieval (ICDAR99), Bangalore, India, 1999, pp. 567--572.

[18] Woodland, P.C., Johnson, S.E., Jourlin, P. and Sparck-Jones, K. *Effects of Out of Vocabulary Words in Spoken Document Retrieval.* Proceedings of SIGIR 2000, Athens, Greece, July, 2000.

[19] Xu, Jinxi and Croft, W. Bruce. "Query Expansion Using Local and Global Document Analysis," *Proceedings of the 19th Annual ACM-SIGIR Conference*, 1996, pp. 4-11

Transcript-Free Search of Audio Archives for the National Gallery of the Spoken Word

[Extended Abstract]

John H.L. Hansen
University of Colorado
Cntr. for Spoken Lang. Res.
Boulder, CO 80309 USA
jhlh@cslr.colorado.edu

J.R. Deller, Jr.
Michigan State University
Dept. Elec. & Comp. Engr.
E. Lansing, MI 48824 USA
deller@msu.edu

Michael S. Seadle
Michigan State University
E308 Main Library
E. Lansing, MI 48824 USA
seadle@msu.edu

ABSTRACT

The National Gallery of the Spoken Word (NGSW) project is creating a carefully organized on-line repository of spoken-word collections spanning the 20th century. Unprecedented technical challenges are inherent in the development of an archive of such extensive scale and diversity. This paper describes research on the development of text-free search-engine technology used to locate requested content in the audio records. A companion paper in these proceedings addresses watermarking technologies for copyright protection.

Categories and Subject Descriptors

H.3.7 [**Info. Storage & Retrieval**]: Digital libraries—*Sys. issues*; E.5 [**Data**]: Files—*Sorting & searching*

1. THE NGSW

The National Gallery of the Spoken Word (NGSW) project is sponsored by the Digital Libraries II Initiative. The goal of the NGSW project is to create a carefully organized on-line repository of spoken word collections, based largely upon the renowned Vincent Voice Library at Michigan State U. (MSU). The collaborative project among specialists in the library sciences, humanities, engineering, and education, will provide the first large-scale repository of its kind. MSU is creating the NGSW in partnership with several universities and agencies, the U. Colorado–Boulder (CU) representing the key collaborator in the engineering developments described here. Further information is found on the NGSW web site at URL www.ngsw.org.

This paper is one of two papers appearing in these proceedings whose purpose is to describe the key technical research issues that are being investigated by academic engineers and allied colleagues in connection with the NGSW. The use of the term "academic engineers" is deliberate in an effort to distinguish the work described here from the significant amount of engineering and technical development (principally creation and maintenance of hardware and software infrastructure) essential to the existence of such a vast resource. In particular, this paper describes research focused on the construction of integrated search mechanisms for locating audio resources in the NGSW collection. This work is being led by researchers at the Center for Spoken Language Research (CSLR) at CU in collaboration with the MSU's Speech Processing Laboratory and MSU Libraries. The companion paper describes research centered at MSU on the development of "watermarking" technologies for secure, efficient, delivery of aural material [6].

2. A CHALLENGING SEARCH PROBLEM

A long-term goal of the NGSW project is to be able to automatically "mine" audio resources for material that is responsive to gallery-users' queries. This feature will be of particular utility to researchers at all levels - from primary school students through professional scholars, journalists, and authors. This aspect of the engineering research is fraught with daunting challenges and uncertainties, as the NGSW represents an audio database whose large scale and content diversity have never been remotely approached.

Although the problem of audio stream search is relatively new, it is related to a number of previous research problems. Many systems developed for audio search, however, assume the existence of associated text or a clean audio stream [5]. Direct information retrieval via audio mining generally focuses on relatively noise-free, single-speaker recordings.[1] Alternative methods have included ways to time-compress or modify speech to allow listeners the ability to skim more quickly through recorded audio data [1]. While a keyword spotting system can generally be used for topic, gisting, or phrase search applications, the system must be able to recover from errors in both a user's text query and in rank-ordered phrase hypotheses in the stream. Phrase search focuses more on locating a single requested occurrence, whereas keyword/topic spotting systems assume a number of possible outcomes. Great strides have also been made in large-vocabulary, continuous speech recognition, suggesting the use of forced transcripts of the NGSW material. While this

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

[1] speechbot.research.compaq.com and www.dragonsys.com.

may be a manageable task for even the larger databases used for speech research (e.g., the Broadcast News Database of 100 hours [9]), the initial offering for NGSW will be 5000 hours (with a potential of 40,000+ total hours based on the existing collection alone), and it is not feasible to achieve accurate forced transcription, even if the text data were available. Further discussion of the NGSW database complexity is found in [7].

3. SYSTEM CONCEPT

A diagram of the current search-engine concept is shown in Fig. 1. The system employs a multifaceted approach to solve this complex search problem. Component modules include[2]

Natural language parsing (NLP): In response to a user query, an N-best parser will be used to rank order audio streams for search, and to correct ill-formed requests. Related research involves the development of adaptive assessment of search success likelihood based on query content.

Environment characterization: An environment processor will be used in conjunction with meta-tag information for the input stream under test. This processor will identify distortion type: acoustic background or recording media noise, restricted channel, reverberation, multiple speakers, etc.

Three adaptation modules: •*Noise adaptation* measures will be based on the results of the environment characterization. The current focus is on rapid methods for parallel model combination [8]. • For repeated searching of material from the same speaker, *hidden-Markov model* (HMM) *adaptation* [4] is performed. Limited data for adaptation are available. Methods being considered include selective training [3] and decision-bias correction [2].

Restricted channel adaptation: The bandwidth of audio streams from Edison cylinder disks, for example, is very small (~ 2 kHz). We are investigating methods to normalize feature sets for models trained with 8 and 16kHz speech.

Speech enhancement: A set of speech enhancement algorithms will be available for user for quality improvement, and audio feature enhancement prior to stream search.

HMM recognition search (e.g., [4]): Following adaptation, an HMM phrase search will be performed. If the environment classifier determines that front-end enhancement could be effective, the input parameter set will be appropriately modified. Detected phrase sets will be rank-ordered using confidence measures and the NLP processor. Essential higher-level knowledge to be incorporated into the recognition search will include extensive metadata records that are being developed by the MSU Libraries using Encoded Archival Description (EAD). This blending of information sources – the stochastic engineering framework with the deterministic archival information – poses one of the interesting problem areas arising out of this digital library research.

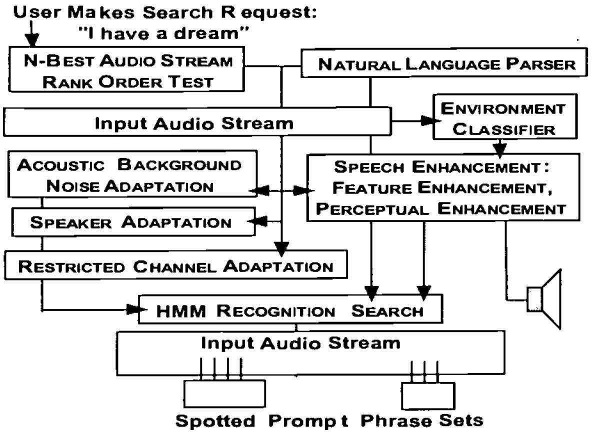

User Makes Search Request:
"I have a dream"
N-Best Audio Stream Rank Order Test
Natural Language Parser
Input Audio Stream
Environment Classifier
Acoustic Background Noise Adaptation
Speech Enhancement: Feature Enhancement, Perceptual Enhancement
Speaker Adaptation
Restricted Channel Adaptation
HMM Recognition Search
Input Audio Stream
Spotted Prompt Phrase Sets

Figure 1: Flow diagram of the audio-stream search engine under development for the NGSW.

Some preliminary results on the application of these search modules to NGSW speech data are found in [7].

Acknowledgments

This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. IBIS-9817485. Opinions, findings, or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF.

4. REFERENCES

[1] B. Arons. A system for interactively skimming recorded speech. *ACM Trans. Computer-Human Interaction*, 4:3–38, 1997.

[2] L. Arslan and J. Hansen. Likelihood decision boundary estimation between HMM pairs. *IEEE Trans. Speech & Audio Process.*, 6:410–414, 1998.

[3] L. Arslan and J. Hansen. Selective training in HMM recognition. *IEEE Trans. Speech & Audio Process.*, 7:46–54, 1999.

[4] J. Deller, Jr., J. Hansen, and J. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, 2d edition, 2000.

[5] J. Foote *et al.* Talker-independent keyword spotting for information retrieval. *Proc. Eurospeech*, volume 3, pp. 2145–2149, 1995.

[6] J. Deller, Jr., A. Gurijala, and M. Seadle. Audio watermarking techniques for the National Gallery of the Spoken Word. Elsewhere in these proceedings.

[7] J. Hansen *et al.* Stream phrase recognition for the NGSW: One small step. *Proc. Int. Conf. Spoken Lang. Process.*, vol. 3, pp. 1089–1092, Beijing, 2000.

[8] R. Sarikaya and J. Hansen. PCA-PMC: A novel use of *a priori* knowledge for fast parallel model combination. *Proc. IEEE ICASSP*, 2000.

[9] P. Woodland *et al.* Experiments in broadcast news transcription. *Proc. IEEE ICASSP*, pp. 909–912, 1998.

[2] For readers from other disciplines, we offer a few general remarks to clarify the need for each of the system modules. Roughly speaking, "NLP" refers to the use of preprogrammed rules of syntax and grammar of a language to expedite speech search and recognition. The "environment" of an utterance refers to any quantifiable characterization of the utterance (e.g., background noise, accents) that would not be evident (e.g., words), or inherent (e.g., grammar), in a written transcript of the message. "Restricted channel" refers to the different frequency ranges that characterize different recording media – a factor that must be compensated for in comparing them to template "features" derived from full audio band data. Speech enhancement algorithms are used to improve the quality of digitized speech which is degraded by various processing measures. The "HMM" is the principal tool used in modern speech processing to capture the statistical structures among sounds and word orderings in a language. For further information, see, for example, [4].

Audio Watermarking Techniques for the National Gallery of the Spoken Word

[Extended Abstract]

J.R. Deller, Jr.
Michigan State University
Dept. Elec. & Comp. Engr.
E. Lansing, MI 48824 USA
deller@msu.edu

Aparna Gurijala
Michigan State University
Dept. Elec. & Comp. Engr.
E. Lansing, MI 48824 USA
gurijala@msu.edu

Michael S. Seadle
Michigan State University
E308 Main Library
E. Lansing, MI 48824 USA
seadle@msu.edu

ABSTRACT

This is one of two companion papers describing technical challenges faced in the development of the National Gallery of the Spoken Word (NGSW). The present paper describes watermarking technologies for intellectual property protection. Following an introduction to data watermarking, the paper focuses on a new algorithm called *transform encryption coding* (TEC) and its application to watermarking the NGSW archives. TEC has a number of flexible features that make it amenable to the NGSW development.

Categories and Subject Descriptors

H.3.7 [**Info. Storage & Retrieval**]: Digital libraries; E.3 [**Data**]: Data encryption; E.3 [**Data**]: Coding & Info. Theory—*Data compression*

1. INTRODUCTION

The National Gallery of the Spoken Word (NGSW) project is a Digital Libraries II sponsored effort whose goal is the creation of a carefully organized on-line repository of spoken word collections, based largely upon the renowned Vincent Voice Library at Michigan State University (MSU). A brief introduction to the NGSW is found in these proceedings in [3], and further information is available at www.ngsw.org. This is one of two papers describing the challenging technical issues being investigated by engineers and colleagues in developing the NGSW. The companion paper describes research on the construction of integrated search mechanisms for locating audio resources in the NGSW [3]. The present paper describes research on digital "watermarking" technologies for secure, efficient delivery of aural material.

Watermarking research is integral to the project because of its importance for copyright protection. The internet continues to create unprecedented legal, ethical, and economic issues surrounding intellectual property rights. The fundamental problem facing developers of repositories like the NGSW is typified by a simple scenario: A copyright-conscientious library gets permission to put a valuable sound file on its web site with language restricting the use to educational purposes. A collector copies it and loads it onto a CD which he then sells. Since most sound materials have copyright protection, this copying is illegal, and, since it is being sold, it almost certainly would not meet the strict "fair use" test in Title 17, Section 107, of the U.S. Code.

Modern technologies make it possible for high-quality copies of legally-protected materials to be created easily with virtual impunity. Such copying happens all the time across the spectrum of materials and media. Sales of pirated materials are significantly less common, but much feared by those who own the rights to items with commercial value. Many rights-holders refuse to grant permissions to libraries and other entities which would provide public access to the protected materials. This refusal creates an awkward situation for libraries in particular. In some cases libraries can exercise preservation reformatting rights (18 USC 108), but the access restrictions in that portion of the law undercut most of the value of digitization. The library might make a risk assessment about making illegal copies available, depending on the flexibility of institutional rules. A more desirable approach is to take measures that make rights-holders more confident that their property is secure. In the cases of audio, image, and video data, watermarking the digital records is an attractive tool, because it allows identification when a stolen file has been used for unauthorized purposes. Watermarking does not prevent copying, but it makes the sale of the copies unprofitable by enabling legal redress. This modest level of deterrence can suffice to sway copyright-holders to grant a reasonable and affordable permission.

2. WATERMARKING

Digital watermarking refers to the process of embedding an imperceptible signal (the *watermark*) into a copyrighted host signal (the *coversignal*). The result is called a *stegosignal.* The unmarked coversignal is never released to the public, and the means for separating the watermark from it are known only to the copyright-holder. When copyright questions arise, the watermark is recovered from the stegosignal as evidence of title. A watermarking scheme generally derives its security from secret codes or patterns, called *keys*, that are used to embed the watermark. Public knowledge of a watermarking technology should not lessen its security.

Robustness against "attacks" is an important requirement of a watermarking technique [6]. An *attack* is an operation

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

that reduces a watermark's value as authentication, while causing minimal damage to the stegosignal. Attacks may be an unintended consequence of signal operations like compression, but such processing can also be used deliberately with malicious intent. Other attacks employ more overt content manipulations like data "cropping" [2] which may be imperceptible, but desynchronize watermark recovery.

3. TRANSFORM ENCRYPTION CODING

Most watermarking research has focused on digital images. In fact, the *transform encryption coding* (TEC) algorithm featured here was originally conceived by Kuo *et al.* at MSU as an image compression algorithm for efficient, robust transmission and storage [4]. However, algorithmic steps that result in compact data representations render, in effect, highly secure encryption. Encryption keys known as *quasi m-arrays* and *gold-code arrays* [5] are used in the TEC process to achieve a high level of unpredictability in the processed signal. TEC is being deployed in a flexible watermarking strategy for speech data in the NGSW. Similar methods can be used for other audio (e.g., music) signals.

Figure 1 illustrates TEC watermarking of a small speech record. The coversignal was obtained from the TIMIT speech database [1] and consists of a male utterance: "She had your dark suit in greasy wash water all year." The frequently-used "mandrill" image was embedded as a watermark with 127 × 127 quasi m-arrays used for encryption. TEC is used to encrypt both the coversignal and watermark, but with different keys used for the two operations. The encrypted watermark is subjected to a masking algorithm to ensure its perceptual transparency before embedding it in the coversignal. Application of the inverse TEC to decrypt the stegosignal, subjects the watermark to a second level of encryption. To recover the watermark, the inverse operations are applied. Two sets of keys are required for each watermark recovery. The keys and location(s) of the watermark(s) in the coversignal are known only to the copyright owner.

Features of TEC watermarking which make it well-suited to the NGSW application are: **(1)** the ability to vary the degree of protection across a spectrum of security requirements (with computation effort scaling linearly with increasing security). This is effected by the frequency, locations, and the strategies for encrypting the watermarks and the source material; **(2)** the feasibility of real-time, "on the fly" watermarking as content is requested by gallery users; **(3)** robustness [2] to many known forms of attack; **(4)** secure transmission; **(5)** available enhancements to effect signal compression [4].

Informal testing has shown that a coversignal-to-watermark energy ratio (CWR) of approximately 18.5 dB is sufficient to assure imperceptibility of the scrambled watermark in the stegosignal. This threshold CWR can almost certainly be reduced by masking schemes that exploit perceptual properties of human audition. From a technical point of view, a higher CWR implies that watermarks are more deeply embedded in a stegosignal, and are therefore more difficult to detect. On the other hand, robustness to certain forms of attack diminishes with increasing CWRs because the embedded marks can be damaged by minor alterations that may not affect perceptual quality (alterations of low-significance bits, low-level noise, etc.). At a baseline CWR of 18.5 dB, we have found that, for attack by addition of wideband noise, the mean recovered watermark of a series of several embedded in a stegosignal (typically four) is identifiable at a stegosignal-to-noise ratio SgoNR of approximately 16.5 dB. (This corresponds to a coversignal-to-noise ratio of 15.9 dB). At SgoNR = 20 dB, the individual watermarks in the series are also identifiable.

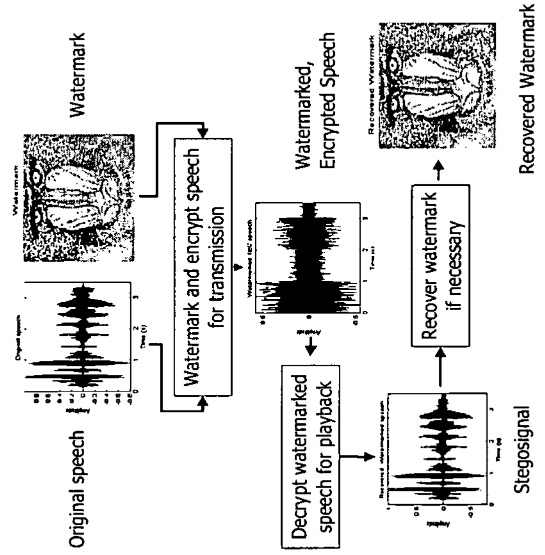

Watermark
Original speech
Watermark and encrypt speech for transmission
Watermarked, Encrypted Speech
Decrypt watermarked speech for playback
Stegosignal
Recover watermark if necessary
Recovered Watermark

Figure 1: Block diagram of TEC-based audio watermarking. A detailed explanation appears in the text. The audio component of the exhibited scheme is found at www.ngsw.org.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. IBIS-9817485. Opinions, findings, or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF.

4. REFERENCES

[1] W. Fisher *et al.* The DARPA speech recognition research database .... *Proc. DARPA Speech Recognition Workshop*, pp. 93–99, 1986.

[2] A. Gurijala and J.R. Deller, Jr. Robust algorithm for watermark recovery from cropped speech. *Proc. IEEE ICASSP*, Salt Lake City, 2001 (in press).

[3] J.H.L. Hansen, J.R. Deller, Jr., and M.S. Seadle. Transcript-free search of audio archives for the NGSW. Elsewhere in these proceedings.

[4] C.J. Kuo, J.R. Deller, Jr., and A.K. Jain. Pre/post filter for performance improvement of transform coding. *Image Comm.*, 8:229–239, 1996.

[5] C.J. Kuo and H.B. Rigas. 2-D quasi m-arrays and gold-code arrays. *IEEE Trans. Info. Theory*, 37:385–388, 1991.

[6] S. Voloshynovskiy *et al.* Attacks and benchmarking. *IEEE Comm. Mag.*, 2001. Spec. issue on watermarking.

Music-Notation Searching and Digital Libraries

Donald Byrd
CIIR, Department of Computer Science
University of Massachusetts
dbyrd@cs.umass.edu

ABSTRACT

Almost all work on music information retrieval to date has concentrated on music in the audio and event (normally MIDI) domains. However, music in the form of notation, especially Conventional Music Notation (CMN), is of much interest to musically-trained persons, both amateurs and professionals, and searching CMN has great value for digital music libraries. One obvious reason little has been done on music retrieval in CMN form is the overwhelming complexity of CMN, which requires a very substantial investment in programming before one can even begin studying music IR. This paper reports on work adding music-retrieval capabilities to Nightingale®, an existing professional-level music-notation editor.

1. INTRODUCTION

In recent years, interest in music information retrieval has been growing at a tremendous pace. The first meeting devoted exclusively to music IR was held late last year [14]; Byrd and Crawford [6] list much more evidence of the growth of interest in terms of grants and papers. There are three basic representations of music: audio, events (normally MIDI), and notations of various sorts. Almost all work on music IR to date has concentrated on the first two domains. However, music in the form of notation, especially the Conventional Music Notation (CMN) of Western society, is of much interest to musically-trained persons, both amateurs and professionals, so searching CMN has great importance for digital music libraries. Of the total music holdings of the Library of Congress, estimated at well over 10,000,000 items, there are believed to be over 6,000,000 pieces of sheet music and tens of thousands, perhaps hundreds of thousands, of scores of operas and other major works [15]. The sheet music and scores are all, of course, in some form of music notation, and the vast majority are undoubtedly in CMN. It is obvious that mechanical assistance could be invaluable in searching a collection of such magnitude.

It seems clear that a major reason little has been done on music retrieval in CMN form is the overwhelming complexity of CMN, which requires a very substantial investment in programming before one can even begin studying music IR. As evidence of its complexity, the source code for Nightingale®, an existing professional-level music-notation editor, amounts to some 160,000 lines of C. We will have more to say about the complexity of CMN.

Another likely reason for the dearth of music-retrieval work on CMN is a lack of collections with which to experiment. The practical availability of what CMN exists in machine-readable form is seriously hampered by the fact that, nothwithstanding several attempts at a standardized format for CMN representations of music [7], no effective standard exists. But the lack of CMN collections is likely to change soon, especially in view of work like the Levy sheet-music project at Johns Hopkins University [8], which is applying Optical Music Recognition on a large scale to create a CMN collection.

This paper reports on work adding music-retrieval capabilities to Nightingale.

2. BACKGROUND

2.1. Basic Representations of Music and Audio

The material in this section is an abridgement of the section of the same title in [6].

There are three basic representations of music and audio: the well-known *audio* and *music notation* at the extremes of minimum and maximum structure respectively, and the less-well-known *time-stamped events* form in the middle. Numerous variations exist on each representation. All three

This material is based on work supported in part by the Digital Libraries Initiative, Phase 2, under NSF grant IIS-9905842, by the U.K. Joint Information Systems Committee under project code JCDE/NSFKCL, and by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623.

Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsors.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

are shown schematically in Figure 1, and described in Figure 2.

The "Average relative storage" figures in the table are for uncompressed material and are our own estimates. A great deal of variation is possible based on type of material, mono vs. stereo, etc., and—for audio—especially with such sophisticated forms as MP3, which compresses audio typically by a factor of 10 or so by removing perceptually unimportant features.

"Convert to left" and "Convert to right" refer to the difficulty of converting fully automatically to the form in the column to left or right. Reducing structure with reasonable quality (convert to left) is much easier than enhancing it (convert to right).

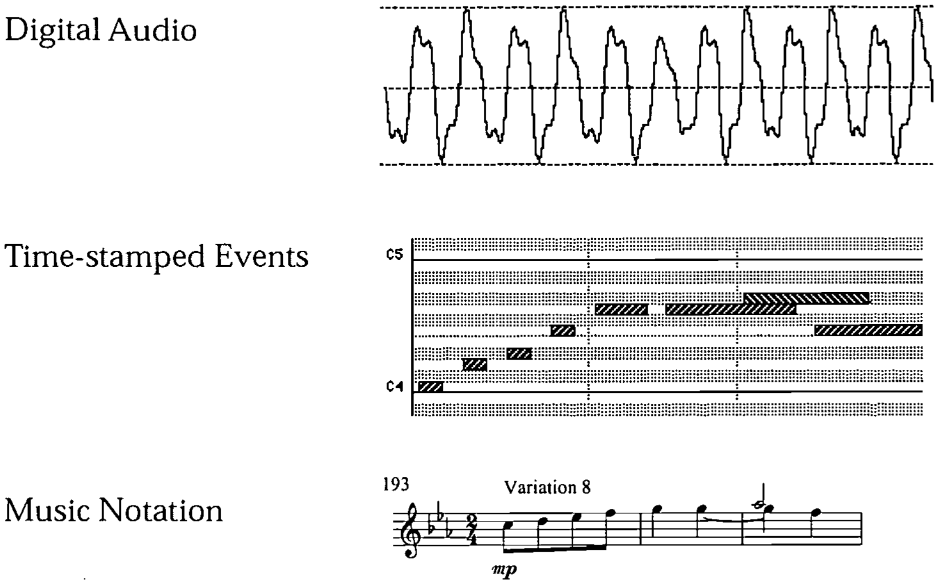


Fig. 1. Basic representations of music

Representation	*Audio*	*Time-stamped Events*	*Music Notation*
Common examples	CD, MP3 file	Standard MIDI File	sheet music
Unit	sample	event	note, clef, lyric, etc.
Explicit structure	**none**	**little (partial voicing information)**	**much (complete voicing information)**
Avg. rel. storage	2000	1	10
Convert to left	-	easy	OK job: easy
Convert to right	1 note/time: pretty easy; 2 notes/time: hard; other: very hard	OK job: fairly hard	-
Ideal for	music bird/animal sounds sound effects speech	music	music

Fig. 2. Basic representations of music

2.1.1. *Music Notation*

There is little doubt that CMN is among both the most elaborate and the most successful graphic communication schemes ever invented. Its complexity places great demands on developers of music-notation software: we have already mentioned the amount of code Nightingale requires. For details of CMN, see standard texts such as those by Read [21] and Ross [22]. For a discussion of its complexity and the implications for software, see [4], especially Chapter 2, and [5].

The success of CMN is obvious from the facts that it has survived with relatively minor changes for over 300 years (see for example [20], pp. 15 ff.), and that it has withstood numerous attempts at major overhaul or complete replacement (see "Notation", Sec. III.4.v, in [23]). Nonetheless, there are other established notations for music, for example tablature (mostly for guitar, lute, and similar instruments: see [20], pp. 143–171), Braille (for blind musicians), and the notations of such other cultures as China, India, Indonesia, and Japan; these systems are beyond the scope of this paper.

2.1.2. *Multiple Representations in Music-IR Systems*

It is important to realize that, in a music-IR system, the internal representation and the external representation—the form used in all aspects of the user interface—may be different; in fact, a system might use a different form in the query and document-display interfaces. In particular, a system might deal with event-level databases, yet accept queries and/or display results in notation form. In an extreme case, it might accept queries in notation form, search an audio database, and display results in a graphic display of events in retrieved audio documents.

2.2. OMRAS and This Work

This work is part of the OMRAS (Online Music Recognition and Searching) project [19]. Among the major goals of OMRAS is to handle music in all three basic representations discussed above with as much flexibility as possible. We are working on searching databases of polyphonic music in all three basic representations, with a full GUI for complex music notation. But beyond this, we are attempting to maximize flexibility with a modular (plug-in) architecture, and exploiting that flexibility by developing and testing two systems with different representations, search methods, and user interfaces (my own NightingaleSearch, and Matthew Dovey's Java Musical Search (JMS) [11]. We feel that the three basic representations can be usefully combined in several ways. Most relevant here is that even when the database is in audio or MIDI form, for many people, CMN will still be useful for formulating queries and displaying retrieved documents. (Admittedly, this is not always practical. As we have said, converting MIDI to CMN for display purposes is not easy, and converting audio to CMN for display is a great deal harder.)

Other threads of the OMRAS project that should eventually interact with CMN-based retrieval work are research on recognition of music from polyphonic audio [3] and research on efficient algorithms for searching music [10].

2.3. Related Work

The research most closely related to this is probably Donncha O'Maidin's C.P.N.View [17, 18]. However, O'Maidin has concentrated on folk music, and his system appears to handle only simple monophonic music. McNab's MR system—part of the MELDEX project—maintains a database in notation form, and it can display both queries and melodies it retrieves in CMN [16, 2]. But again it can handle only simple monophonic music, in this case without tuplets, beams, etc. Furthermore, queries must be entered in audio form: there is no CMN entry or editing.

The well-known commercial music editor Finale has for years had a command for searching music in CMN form by content, but it can search only within a single score at a time [9]. Perhaps more important, Finale limits itself to what might be called "document-editor" style searching, i.e., finding the next match for Boolean criteria. This is as opposed to the "IR" style searching for all matches in a document or database that makes possible best-match IR and ranking.

In fact, work on music retrieval in CMN form is conspicuous thus far by its scarcity. The obvious reason is the huge investment in programming complex CMN demands before one can even begin studying music IR.

So-called "piano roll" notation is the graphic equivalent of music in the event representation. For complex music, piano roll is a great deal less demanding than CMN, and it can convey much of the same information; but it has not been used in music IR much, either. One system that does use piano roll, albeit in a simplified form indicating note onsets but not durations, is Dovey's, in his testbed framework Java Music Search (JMS) [11]. Dovey not only displays both queries and retrieved music in this form, he also uses it as an abstract model of music.

2.4. Music Information Needs and the Audience for Searching CMN

It seems obvious that—in the face of MIDI and, especially, audio as alternatives—CMN as a basis for a music-retrieval system will be of interest only to those with some knowledge of CMN. On the other hand, for Western music of the last few centuries, at least, CMN is arguably the best graphic representation ever developed: it has value purely as a user-interface device.

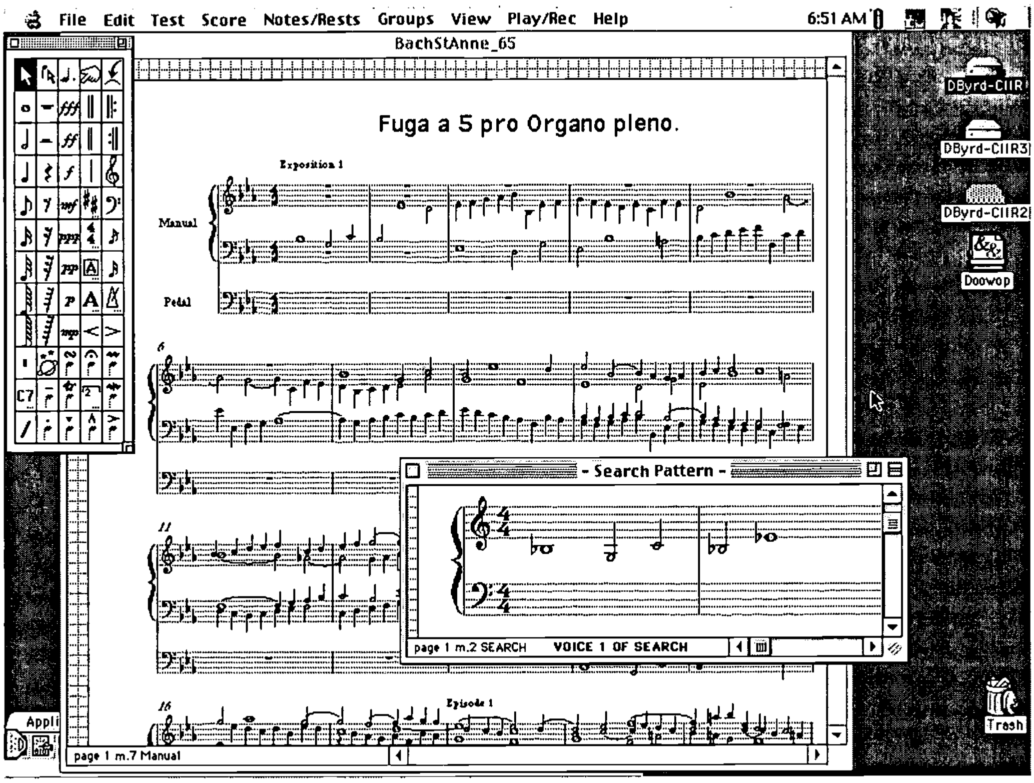


Fig. 3. Bach: "St. Anne's" Fugue, with Search Pattern

3. NIGHTINGALESEARCH

Nightingale® is a professional-level music-notation editor for the Macintosh computer, written in the C language; it has been marketed commercially for a number of years [1]. Since I led the team that developed Nightingale, I not only had access to the source code, I knew it well. I decided to use it as a platform for studying CMN-based music IR by adding several music-searching features and commands: the resulting program is "NightingaleSearch".

3.1. Overview

NightingaleSearch inherits all the normal functionality of Nightingale. It can display and edit any number of *scores*—CMN documents—at the same time, and it supports several ways of creating music, including recording from a MIDI device (usually a synthesizer keyboard), importing standard MIDI files, pasting from other scores, etc. The searching commands use the contents of a special score, the "Search Pattern", as the query. In nearly all respects, this is an ordinary Nightingale score, and music can be entered into it with any of Nightingale's facilities. See Figure 3.

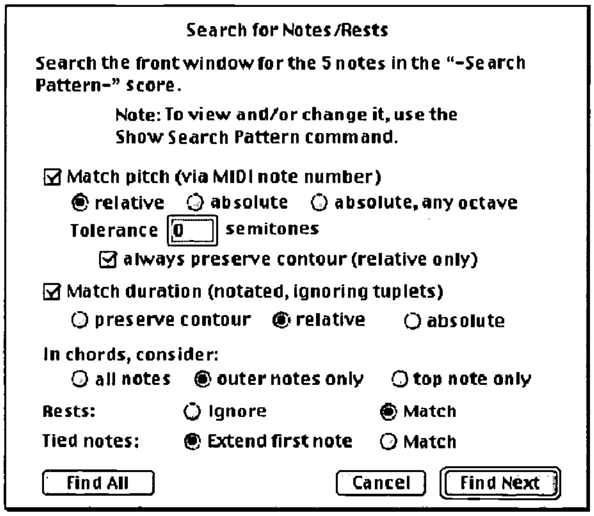
Menu commands to "Search for Notes/Rests" and "Search in Files" bring up the dialog in Figure 4. NightingaleSearch is a research prototype, and I show the dialog only to make clear what the program can do: there are far too many options for mortal users. To sum it up, matching can be based on pitch, duration, or both. In IR terms, matching is Boolean: there are no approximate matches, except for those allowed by Tolerance (for pitch) and preserve contour (for duration) as described below. The main options are:

- **Match pitch (via MIDI note number)**: if not checked, matching ignores the pitches of the notes. *Relative* matches any transposition of the entire pattern; *absolute* matches only the exact original pitches. Pitch options include:
    - Tolerance: each interval can be off from the corresponding interval in the pattern by the given number of semitones. However, for "relative" matches, if "always preserve contour" is checked, the match will still fail unless the upward, repeat, or downward motion of each interval in the pattern is preserved. This is very useful to avoid "false positives": without it, for example, a tolerance of 2 would allow an upward chromatic scale to match a downward one or a series of repeated notes.
- **Match duration (notated, ignoring tuplets)**: if not checked, matching ignores the durations of the notes and—if rests are included—,of the rests. *Relative* matches the original series of durations multiplied by

any factor: in musical terms, it recognizes augmentation and diminution. *Absolute* matches only the exact original series of durations. Duration options include:

- Preserve contour: this is analogous to the "always preserve contour" option for pitch in that it distinguishes just three relationships (in this case, longer, shorter, and the same), though it differs by being an alternative to relative or absolute rather than modifying relative.

- **In chords, consider**: all notes, outer notes only, or top note only. Notice that a chord in Nightingale is entirely within a voice, so these options do not apply, say, to a brass quintet where each instrument plays a single note: they are mostly for keyboard music. In any case, "all notes" will rarely be useful, since inner notes of chords nearly always serve just to enrich the harmony or texture.



Search for Notes/Rests

Search the front window for the 5 notes in the "-Search Pattern-" score.

Note: To view and/or change it, use the Show Search Pattern command.

☑ Match pitch (via MIDI note number)

● relative ○ absolute ○ absolute, any octave

Tolerance [0] semitones

☑ always preserve contour (relative only)

☑ Match duration (notated, ignoring tuplets)

○ preserve contour ● relative ○ absolute

In chords, consider:

○ all notes ● outer notes only ○ top note only

Rests: ○ Ignore ● Match

Tied notes: ● Extend first note ○ Match

Fig. 4. Search Dialog

Search for Notes/Rests just searches the score in the frontmost window. Search in Files is more interesting. It exists in a version that searches all Nightingale scores in a given folder, and a version that searches a "database". As of this writing, the database is simply a file that describes in order of occurrence all the notes in any number of Nightingale scores, with information identifying the original scores. Thus, it does not provide a way "to avoid the efficiency disaster of sequential searching". [6] Text IR gets around this problem by indexing, which can improve performance with a large database by thousands of times; research on indexing polyphonic music is underway or planned by several groups, including OMRAS.

3.1.1. Retrieval Levels and the Result List

NightingaleSearch does passage-level retrieval, i.e., it looks for and reports individual occurrences of matches for the search pattern. In contrast, most IR systems, for music as well as text, retrieve entire documents that match the pattern in one or more places. It could be argued that the "average" music document is much longer and more complex than the "average" text document, and therefore retrieval of passages is much more important with music. This is a strong argument, though of course it depends on the document collection: by any obvious measure, the average article in *The New Yorker* is longer than the average folksong.

Currently, the result list is displayed in a scrolling-text window; there is no link to let the user choose an entry in the list and view that "match" in CMN. MELDEX [2] lets the user listen to any entry in its result list as well as view it, and both options would be very helpful for NightingaleSearch.

3.2. NightingaleSearch in Action

Notation representations of music—CMN or other—are distinguished from audio and event representations mostly by the amount of explicit structure they contain. In particular, with minor exceptions, music in CMN contains complete voicing information, i.e., the voice membership of every note is evident from the notation. For example, the opening of Bach's "St. Anne's" Fugue is shown in Figure 3: the three staves contain five voices, as suggested by the stems going up and down for notes on the upper two staves. The first five notes of the piece are enough for a human musician to identify all 20 or so clearcut occurrences of the main subject (essentially, the theme), but searching for exact (except for transposition) matches of them finds only 5, all valid. This is 100% precision but only 25% recall. One problem is that some instances are so-called "tonal answers", resulting in pitch intervals. slightly different from the original. For example, the second occurrence of the subject, starting in m. 3, begins by going down 1 semitone rather than the original version's 3. Setting the tolerance in the search dialog to 2 results in finding 8 matches: again all are valid, but 12 valid "hits" were still not found, for a precision of 100% and recall of 40%. The result list appears in Figure 5. Notice that, for each match, NightingaleSearch displays a label for the section of the piece (the passage) as well as the measure number, plus the voice number and "instrument" (actually, "Manual" and "Pedal" are both parts of the single instrument this piece was written for, the organ). This much information is very rarely available in event representations, and never in audio.

```
Time 0.13 sec. 8 matches (in order of error):
 1: BachStAnne_65: m.1 (Exposition 1), voice 3 of Manual, err=p0 (100%)
 2: BachStAnne_65: m.7 (Exposition 1), voice 1 of Manual, err=p0 (100%)
 3: BachStAnne_65: m.14 (Exposition 1), voice 1 of Pedal, err=p0 (100%)
 4: BachStAnne_65: m.22 (Episode 1), voice 2 of Manual, err=p0 (100%)
 5: BachStAnne_65: m.31 (Episode 1), voice 1 of Pedal, err=p0 (100%)
 6: BachStAnne_65: m.26 (Episode 1), voice 1 of Manual, err=p2 (85%)
 7: BachStAnne_65: m.3 (Exposition 1), voice 2 of Manual, err=p6 (54%)
 8: BachStAnne_65: m.9 (Exposition 1), voice 4 of Manual, err=p6 (54%)
```

Figure 5. Result list for search of the "St. Anne's" Fugue

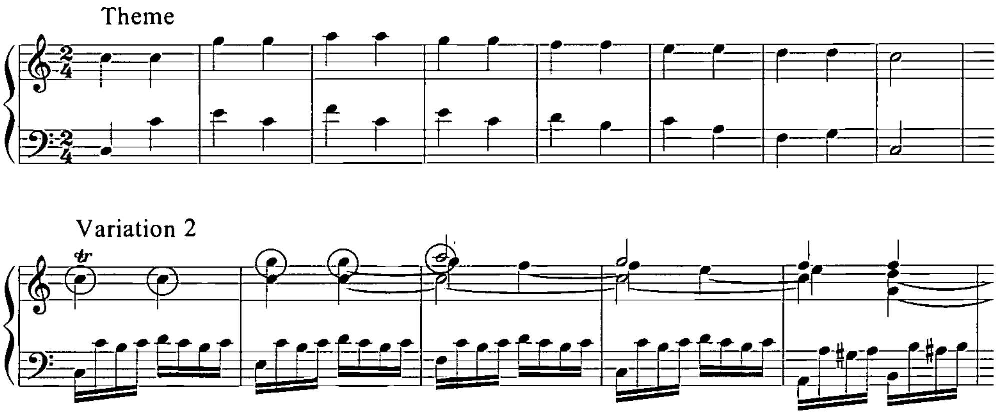


Fig. 6a (above) and b (below) (Mozart)

Using more of the fugue subject as the query naturally tends to increase precision at the expense of recall. However, with the first seven notes of the piece as query, tolerance of 2, and ignoring duration, it does well on both metrics: it finds 22 matches, of which 4 are false, for a precision of 82% and recall of 90%.

For another example, consider a user looking in a digital music library for the old children's song that is called in English-speaking countries by several names, but best known as "Twinkle, Twinkle, Little Star". This melody has been used in many ways, including music by (among others) Mozart, Dohnanyi, and the violin pedagogue Shinichi Suzuki. Mozart used it in his Variations for piano, K. 265, on "Ah, vous dirais-je, Maman"; the melody is shown in his version in Figure 6a. One difficulty this piece demonstrates is the effects of complete voicing on music IR. In Variations 2 (Figure 6b), 4, and 9, the melody starts in one voice, then, after four notes—not enough for a reliable match—moves to another. Of course, it is easy simply to ignore voice information, but doing so is likely to have catastrophic effects on precision [6].

In fact, this piece of Mozart's demonstrates several difficult problems for music IR. Some of the other variations employ tricks like distorting the melody or adding ornamental notes to it, but others discard the melody completely while retaining the harmony and bass line! But none of these subtleties really matters to our hypothetical digital-music-library user, who presumably simply needed their attention drawn to the Mozart piece: in other words, document-level retrieval is adequate in this case. Searching for the first four notes of the Twinkle theme in a very small database finds the matches shown in Figure 7.

```
Time 1.27 sec. 13 matches (in order found):
 1: BaaBaaBlackSheep: m.1, voice 1 of Unnamed
 2: BaaBaaBlackSheep: m.9, voice 1 of Unnamed
 3: Mozart-TwinkleVar_10: m.1 (Theme), voice 1 of Piano
 4: Mozart-TwinkleVar_10: m.84 (Variation 9), voice 2 of Piano
 5: Suzuki-TwinkleVar: m.16 (Variation D), voice 1 of Violin
 6: Suzuki-TwinkleVar: m.21 (Theme), voice 1 of Violin
 7: Suzuki-TwinkleVar: m.29 (Theme), voice 1 of Violin
 8: Twinkle-Hirsch2ndGraderVer: m.1, voice 1 of Unnamed
 9: Twinkle-Hirsch2ndGraderVer: m.9, voice 1 of Unnamed
 10: TwinkleHARMONETVar: m.1, voice 1 of Original
 11: TwinkleHARMONETVar: m.9, voice 1 of Original
 12: TwinkleMelody: m.1, voice 1 of Unnamed
 13: TwinkleMelody: m.9, voice 1 of Unnamed
```

Figure 7. Result list for search for the "Twinkle" theme

3.3. Intuition vs. Evaluation in Music IR

No formal evaluation has yet been done of NightingaleSearch. In fact, a great deal of work on music IR to date has been speculative, and what evaluation of systems has been done has generally not been at all rigorous. It is tempting to criticize researchers for their unscientific work, but, in the words of Byrd and Crawford [6] (citations omitted):

> To put things in perspective, music IR is still a very immature field... For example, to our knowledge, no survey of user needs has ever been done (the results of the European Union's HARMONICA project are of some interest, but they focused on general needs of music libraries). At least as serious, the single existing set of relevance judgements we know of is extremely limited; this means that evaluating music-IR systems according to the Cranfield model that is standard in the text-IR world...is impossible, and no one has even proposed a realistic alternative to the Cranfield approach for music. Finally, for efficiency reasons, some kind of indexing is as vital for music as it is for text; but the techniques required are quite different, and the first published research on indexing music dates back no further than five years. Overall, it is safe to say that music IR is decades behind text IR.

I would argue that the state of the art of music-IR evaluation is so primitive, there is little point in trying to evaluate music-IR systems and techniques rigorously. Instead, the field is best served by music-IR system developers relying on intuition and informal evaluation, while other researchers develop tools to make meaningful evaluation possible.

4. CONCLUSIONS

Other than Finale—which is limited to finding a single match at a time in a single file—NightingaleSearch is the only program I know of that allows searching complex music with a query in any type of music notation, and the only program that displays the results of such a search in notation form. NightingaleSearch has many shortcomings. Not the least is that any music to be searched must first be in a format it can use, but we are working on connectivity with other programs, for example, via a utility that converts music in the well-known Humdrum kern format [13]. Also, any evaluation of NightingaleSearch, even the most basic, remains to be done. In any case, there would not be much point to evaluating it with the primitive tools available now. But informal use to date strongly supports intuitions of the value of notation-based music retrieval. In the not-too-distant future, the ability to search music notation will surely be part of every digital music library that contains notation.

5. ACKNOWLEDGEMENTS

In addition to their other contributions to the OMRAS project, Tim Crawford, Matthew Dovey, and Jeremy Pickens made many helpful comments on this paper. Thanks also to Alan Belkin for his comments.

6. REFERENCES

[1] AMNS (2000). Nightingale. http://www.ngale.com .

[2] Bainbridge, D. (1998). MELDEX: A Web-based Melodic Index Service. In Hewlett & Selfridge-Field, *op. cit.*

[3] Bello, J.P., Monti, G., and Sandler, M. (2000). Techniques for Automatic Music Transcription. Read at the First International Symposium on Music Information Retrieval; available at http://ciir.cs.umass.edu/music2000 .

[4] Byrd, D. (1984). *Music Notation by Computer.* PhD dissertation, Indiana University (available from UMI, Ann Arbor, Michigan, U.S.A.).

[5] Byrd, D. (1994). Music Notation Software and Intelligence. *Computer Music Journal* 18(1), pp. 17–20.

[6] Byrd, D. and Crawford, T. (2001). Problems of Music Information Retrieval in the Real World. To appear in *Information Processing and Management.*

[7] Castan, G. (2001). Musical notation codes. http://www.s-line.de/homepages/gerd_castan/compmus/notationformats_e.html .

[8] Choudhury, G.S., Droetboom, M., DiLauro, T., Fujinaga, I., and Harrington, B. (2000). Optical Music Recognition System within a Large-Scale Digitization Project. Read at the First International Symposium on Music Information Retrieval; available at http://ciir.cs.umass.edu/music2000 .

[9] Coda (2000). *About Finale for Macintosh* (Coda Music Technology).

[10] Crawford, T., Iliopoulos, C.S., and Raman, R. (1998). String-Matching Techniques for Musical Similarity and Melodic Recognition. In Hewlett and Selfridge-Field, *op. cit.*

[11] Dovey, M. (1999). A matrix based algorithm for locating polyphonic phrases within a polyphonic musical piece. In *Proceedings of AISB '99 Symposium on Artificial Intelligence and Musical Creativity.* Edinburgh, Scotland: Society for the Study of Artificial Intelligence and Simulation of Behaviour.

[12] Hewlett, W., and Selfridge-Field, E., eds. (1998). *Melodic Similarity: Concepts, Procedures, and Applications (Computing in Musicology 11)* (MIT Press).

[13] Huron, D. (1998). Humdrum and Kern: Selective Feature Encoding. In Hewlett & Selfridge-Field, *op. cit.*

[14] ISMIR (2000). International Symposium on Music Information Retrieval. http://ciir.cs.umass.edu/music2000 .

[15] LaVine, K. (2000). Personal communication, May 2, 2000.

[16] McNab, R., Smith, S., Bainbridge, D., and Witten, I. (1997). The New Zealand Digital Library MELody InDEX. DLib Magazine, May 1997; at http://www.dlib.org .

[17] O'Maidin, D. (1995). "A Programmer's Environment for Music Analysis". Technical Report UL-CSIS-95-1, Department of Computer Science, University of Limerick, Ireland.

[18] O'Maidin, D. (1998). "A Geometrical Algorithm for Melodic Difference." In Hewlett & Selfridge-Field, *op. cit.*

[19] OMRAS (2000). Online Music Recognition and Searching. http://www.omras.org .

[20] Rastall, R. (1982). *The Notation of Western Music* (St. Martin's Press).

[21] Read, G. (1969). *Music Notation,* 2nd ed. (Crescendo).

[22] Ross, T. (1970). *The Art of Music Engraving and Processing* (Hansen).

[23] Sadie, S., ed. (1980). The New Grove Dictionary of Music and Musicians (Macmillan).

Feature Selection for Automatic Classification of Musical Instrument Sounds

Mingchun Liu and Chunru Wan
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore 639798
Tel: +65 790 6298
E-mail: {P147508078 I Ecrwan}@ntu.edu.sg

ABSTRACT

In this paper, we carry out a study on classification of musical instruments using a small set of features selected from a broad range of extracted ones by sequential forward feature selection method. Firstly, we extract 58 features for each record in the music database of 351 sound files. Then, the sequential forward selection method is adopted to choose the best feature set to achieve high classification accuracy. Three different classification techniques have been tested out and an accuracy of up to 93% can be achieved by using 19 features.

KEYWORDS: Sequential forward feature selection, classification, musical instrument, feature extraction.

INTRODUCTION

The collection of musical instrument sounds is an obligatory part of comprehensive music digital libraries. Automatic musical instrument classification can be very helpful for indexing the database as well as for annotation and transcription. In [2], four instruments, guitar, piano, marimba, and accordion, could be identified using an artificial neural network or nearest neighbor classifier. The results of this preliminary work achieved were encouraging although only temporal features were utilized. In [3], polyphonic music was separated into each monophonic one using comb filters and musical instruments were estimated by frequency analysis. More recently, a system for musical instrument recognition was presented that used a wide set of features to model the temporal and spectral characteristics of sounds [1].

Due to a large number of audio features available, how to choose or combine them to achieve higher classification accuracy is studied in this paper. Simply choosing all the features available often doesn't yield the best performance, because some features give poor separability among different classes and some are highly correlated. These bad features have a negative effect when added into the feature vector. Therefore, a sequential forward selection method is adopted to select the so-called best feature set. Experiments of classifying the musical instruments into the right families have been conducted using nearest neighbor (NN) classifier, modified k-nearest neighbor (k-NN) classifier, as well as Gaussian mixture model (GMM), based on the selected best features.

THE DATABASE

The musical instruments are commonly sorted into five families according to their vibration nature, which are string, brass, percussion, woodwind, and keyboard. Currently, there are 351 files in the musical instrument database. A brief description of the sound files is given in Table 1.

Table 1. The musical instrument collection

Classes	Instruments
Brass	Fanfare, French horn, trombone, trumpet, tuba
Keyboard	Piano
Percussion	Bell, bongo, chime, conga, cymbal, drum, gong, maraca, tambourine, triangle, timbales, tomtom, tympani
String	Guitar, violin
Woodwind	Oboe, saxophone

FEATURE EXTRACTION

The lengths of the sound files range from 0.1 second to around 10 seconds. Every audio file is divided into frames of 256 samples, with 50% overlap at the two adjacent frames. Each frame is hamming-windowed and 58 features are extracted for each frame. Means and standard deviations of the frame-based features are computed as the final features for each audio file. The 58 features from three categories are showed in Table 2. The features 1–8 are temporal features, 9–32 are spectral features, and 33–58 are coefficient features.

FEATURE SELECTION

The extracted features are normalized by their means and standard deviations. Then, a sequential forward selection (SFS) method is used to select the best feature subset. Firstly, the best single feature is selected based on classification accuracy it can provide.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

Table 2. Feature description

Feature number	Descriptions
1-2	Mean and standard deviation of volume root mean square
3	Volume dynamic ratio
4	Silence ratio
5-6	Mean and standard deviation of frame energy
7-8	Mean and standard deviation of zero crossing ratio
9-10	Mean and standard deviation of centroid
11-12	Mean and standard deviation of bandwidth
13-20	Means and standard deviations of four sub-band energy ratios
21-22	Mean and standard deviation of pitch ·
23-24	Mean and standard deviation of salience of pitch
25-28	Means and standard deviations of first two formant frequencies
29-32	Means and standard deviations of first two formant amplitudes
33-58	Means and standard deviations of first 13 Mel-frequency cepstral coefficients

Next, a new feature, in combination with the already selected features, is added in from the rest of features to minimize the classification error rate. This process proceeds until all the features are selected. The SFS method can quickly provide a sub-optimized set of features in comparison with the exhaustive searching approach which is not practical due to exorbitant computation time involved in the concerned applications.

EXPERIMENTS

The database is split into two equal parts: one for training, and the other for testing. Three classifiers, NN, modified k-NN, GMM, are used to classify the musical instruments. In the modified k-NN, we firstly find the k (k=3 in this paper) nearest neighbors from each class instead of whole training set, their means are calculated and sorted, then assign the testing feature vector with the class corresponding to the smallest mean. The classification accuracy versus feature dimension of the three classifiers using combination of temporal, spectral, and coefficient features, is showed in Figure 1. From the figure, we can see that the performance increases rapidly with the increase of features at the beginning. It remains more or less constant and even decreases after a particular number. Other experiments using single temporal, spectral, coefficient feature or any combination of them have the similar phenomenon. The classification accuracies of the best feature set and the corresponding feature numbers are listed in the Table 3. The best feature sets for different classifiers are different. Among all the experiments, the modified k-NN classifier using 19 features, in which 6 are temporal, 8 are spectral, and 5 are coefficients, achieves the highest accuracy of 93%.

Table 3. Classification performance

Features	NN	k-NN	GMM
Time	0.73(7)	0.73(6)	0.71(6)
Frequency	0.80(8)	0.82(10)	0.80(11)
Coefficient	0.86(17)	0.85(13)	0.80(17)
Time and Coefficient	0.90(14)	0.86(9)	0.84(14)
Frequency and coefficient	0.89(29)	0.91(15)	0.85(17)
Time and frequency	0.85(12)	0.87(7)	0.81(8)
Time, frequency, and coefficient	0.91(13)	0.93(19)	0.87(22)

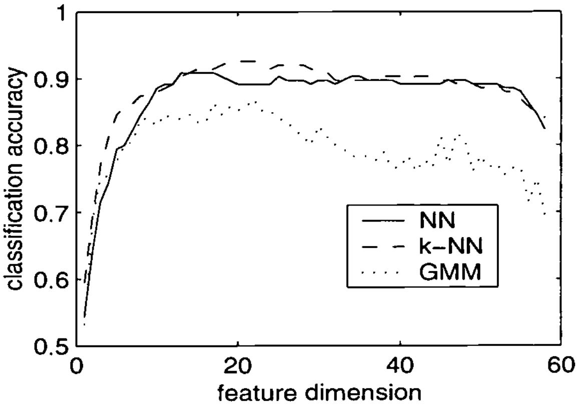


Figure 1. The Classification accuracy versus feature dimension for testing patterns

CONCLUSION

In this paper, we use a sequential forward feature selection scheme to pick up the best feature set in single or any combination of temporal, spectral, and coefficient space for classifying musical instruments into five families. Simple classifier using small set of features can achieve a satisfactory result. Since the number of features is reduced, less computation time is required for classifying the music instrument sounds. This will be beneficial to real-time applications such as sound retrieval from large databases.

REFERENCES

[1] Eronen, A.; Klapuri, A. Musical instrument recognition using cepstral coefficients and temporal features. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 63-71, 2000.

[2] Kaminsky, I.; Materka, A. Automatic source identification of monophonic musical instrument sounds. IEEE International Conference on Neural Networks, Vol. 1, pp. 189-194, 1995.

[3] Miiva, T.; Tadokoro, Y. Musical pitch estimation and discrimination of musical instruments using comb filters for transcription. 42nd Midwest Symposium on Circuits and Systems, Vol. 1, pp. 105-108, 1999.

Adding Content-Based Searching to a Traditional Music Library Catalogue Server

Matthew J. Dovey
Visiting Research Fellow, OMRAS Project
Dept. of Computer Science, Kings College, London
Tel: +44 1865 278272 E-mail: matthew.dovey@las.ox.ac.uk

ABSTRACT

Most online music library catalogues can only be searched by textual metadata. Whilst highly effective - since the rules for maintaining consistency have been refined over many years - this does not allow searching by musical content. Many music librarians are familiar with users humming their enquiries. Most systems providing a "query by humming" interface tend to run independently of music library catalogue systems and not offer similar textual metadata searching. This paper discusses the ongoing investigative work on integrating these two types of system conducted as part of the NSF/JISC funded OMRAS project (http://www.omras.org).

Categories and Subject Descriptors

H.3.7[Information Systems]: Information Storage and Retrieval – *Digital libraries, Standards, Systems issues.*

General Terms

Algorithms; Design

Keywords

Music Information Retrieval; Z39.50

1. INTRODUCTION

OMRAS (Online Music Retrieval And Searching) is a three-year collaborative project between Kings College London and the Center for Intelligent Information Retrieval, University of Massachusetts. One of its aims is to look at issues surrounding content-based searching of polyphonic music; current research in content-based music searching has tended to concentrate on monophonic music, i.e. music consisting of a single melodic line, ignoring the complexities in more complex music textures such as those found in, say, an orchestral symphony.

However, a key mission statement of the joint JISC/NSF International Digital Library Initiative, which is funding the OMRAS work, is to make existing digital collections more accessible. We intend that the work of OMRAS will achieve this for music collections by making content-based searching possible as well as standard metadata searching such as by composer or title. This paper outlines some collaborative work with JAFER, another JISC funded project in the UK, to provide a prototype applying the searching technologies developed in OMRAS to enhance an existing online music library catalogue.

2. The OMRAS Test Framework

The OMRAS project has developed a Java based framework to test the performance of various search algorithms and techniques. The framework is currently command-line driven and not aimed at novice users. We can load music files (in different formats), different algorithms and different user-interface components (for displaying and editing queries and results) into the system. We can then experiment with and compare different algorithms and representations for music.

The framework takes advantage of Java's object-oriented nature: all components such as file-format modules, user-interface elements and search algorithms are implemented as independent java objects. These are manipulated in the framework using a scripting interface such as BeanShell[1], but selected components can be used in other software. At the moment, we have modules for handling a small number of file formats (including MIDI) and are working on others. The user-interface components are based on piano-roll type displays, but we are planning to incorporate better music-score display software soon. With this framework we have devised a number of algorithms for searching musical content, described elsewhere [1][2].

3. INTEGRATING MUSIC SEARCHING WITH LIBRARY SYSTEMS

3.1 Z39.50 and Library Systems

Many music libraries will already have an electronic catalogue of their collections. Typically systems contain a database of records in a cataloguing format called MARC[2] (an established standard for library catalogue systems) and follow established cataloguing practices such as AACR2 [3]. Many systems allow querying using a protocol called Z39.50 (ISO 23950)[3]. This is fairly rich in functionality; it allows a third-party system to query an external database as show in Figure 1. The client, possibly a dedicated client such as EndNote or a web gateway, sends a query to the library system. The library server then returns the records.

A major strength of Z39.50 is that the query uses abstract search points such as "author" and "title", and requires no knowledge of the underlying database structure of the server. In 1998, I successfully proposed adding "musical phrase" to the standard as a potential search point[4]. This allows a music library catalogue to

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

[1] http://www.beanshell.org

[2] http://www.loc.gov/marc

[3] http://www.loc.gov/z3950/agency

[4] http://lcweb.loc.gov/z3950/agency/defns/bib1.html#music

respond to queries which include traditional textual search terms as well as a query by musical content. Most library systems at present cannot cope with queries which include music content. Although some systems do provide music content query, e.g. the MELDEX system[5], these typically do not have rich metadata or detailed holding/circulation information as in library systems.

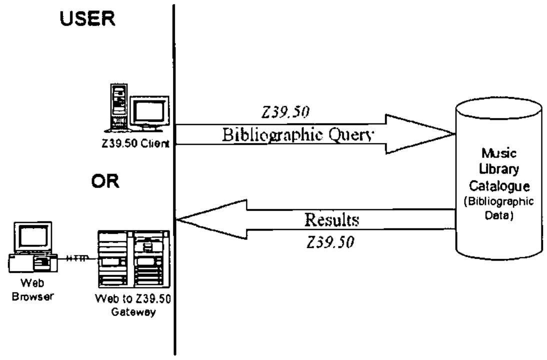


Figure 1

3.2 A Z39.50 Proxy Solution

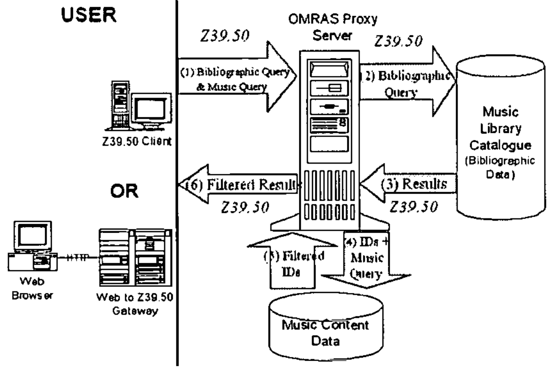


Figure 2

It is possible to put a Z39.50 proxy server in the system to do additional processing. Figure 2 illustrates how such a proxy could add music content searching to an existing library catalogue without modifying the existing system. The query consisting of textual metadata (e.g. composer and title details) and a musical phrase is sent to the proxy server. The proxy passes the textual part of the query to the music library catalogue and receives the appropriate results. The proxy then filters these results by performing a music content query on its own database of musical scores. Each record on the library catalogue has a unique identifier with which the corresponding score in the proxy's database can be tagged. These identifiers are used to determine the records to be returned to the user. The end-user sees a system that can take combined textual and music based queries.

There is a slight difference if the user only submits a music query as shown in figure 3. Here the music search is carried out first, giving the proxy a list of potential matches. The identifiers returned from this search are sent to the music library catalogue in order to retrieve full details of the item (including location and circulation details) for the user.

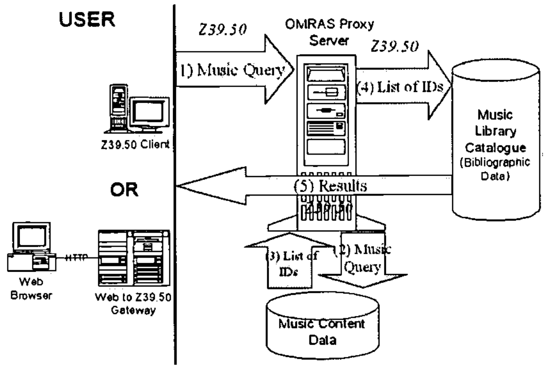


Figure 3

4. The OMRAS/JAFER Prototype

The JAFER project[6] is a JISC funded project based at Oxford University which is developing Java based Z39.50 tools. Part of the toolkit allows the building of Z39.50 proxies. The modular nature of the OMRAS test-bed framework means we could take the java objects for the search and indexing algorithms and plug them into the JAFER toolkit's Z39.50 proxy to produce a proxy as described above. This prototype is still very much proof of concept but it has been demonstrated against the library systems at Kings College London, Oxford University and University of Massachusetts at Amherst. The proxy also modifies the returned records to add a pointer to a MIDI file for the work.

5. FUTURE RESEARCH

We are continuing to develop our algorithms to cope with more complex representations of music. We are also looking at methods for ranking the result lists in order of relevance [4]. There are other questions concerning the user interface both in how users should enter musical queries and how the results should be displayed which are being addressed within OMRAS.

6. ACKNOWLEDGEMENTS

I would like to thank Tim Crawford of Kings College London and Don Byrd of the University of Massachusetts at Amherst for their comments on this paper, and Katherine Ferguson of Oxford University for additional proofreading. I also would like to acknowledge the NSF/JISC IDLI and JISC DNER programmes which are funding the OMRAS and JAFER projects upon which this work is based.

7. REFERENCES

[1] Dovey, M. J. An algorithm for locating polyphonic phrases within a polyphonic piece. AISB'99 Symposium on Musical Creativity (Edinburgh, April, 1999), 48-53.

[2] Dovey M. J. & Crawford T. Content-based music retrieval (in preparation). The 5th World Multi-Conference on Systemics, Cybernetics and Informatics (Orlando, 2001).

[3] Gorman, Michael. Anglo-American Cataloguing Rules 2. The American Library Association. (1998). ISBN: 0838932118

[4] Crawford, Tim and Dovey, M. J. Heuristic Models of Relevance Ranking in Musical Searching. Proceedings of the Fourth Diderot Mathematical Forum (Vienna, 1999).

[5] http://www.nzdl.org/musiclib

[6] http://www.lib.ox.ac.uk/jafer

Locating Question Difficulty through Explorations in Question Space

Terry Sullivan

Texas Center for Educational Technology
University of North Texas
Denton, TX 76203 USA
email: tsullivan@unt.edu

ABSTRACT

Three different search effectiveness measures were used to classify 50 question narratives as easy or hard. Each measure was then encoded onto a spatial representation of interquestion similarity. Discriminant analysis based on the resulting map was able to predict question difficulty with approximately 80% accuracy, robust across multiple measures. Implications for the design of digital document collections are discussed.

Keywords
Information visualization, question classification

INTRODUCTION

Recent years have seen a growing interest in the quantitative measurement of question difficulty. Some researchers have analyzed sets of questions, positing that increased variability among multiple question narratives corresponds to greater difficulty of the constituent questions [2]. Other studies have raised the enticing possibility that it may be possible to characterize the difficulty of specific questions, based on scaled visual representations of "Question Space" [3].

The present study is based on the premise that search effectiveness is a meaningful surrogate for question difficulty. That is, when measured and compared using common points-of-reference, superior search performance corresponds to easier questions, while diminished search performance corresponds to harder questions. Under these assumptions, traditional measures of information retrieval (IR) effectiveness can be used as meaningful indicators of question difficulty.

METHOD

Material

The questions analyzed in the present study were drawn from Volume 6 of the TREC IR test collection [4]. TREC questions are roughly paragraph-sized narratives describing a user's information need. In all, some 50 questions (TREC Questions 301 to 350) were analyzed. The text of each question was stoplisted, stemmed, and reduced to a binary term vector. Similarity among questions was quantified using a cosine similarity coefficient. (See [1] for a detailed discussion of document similarity measures.)

Procedure

The resulting question similarity matrix was processed using multidimensional scaling (MDS), a statistical technique that allows similarity among objects to be represented as proximity among points in n-dimensional space. The questions were scaled in two dimensions under Maximum Likelihood Estimation assumptions using SAS. Because MDS output consists of spatial coordinates, it can be used to create a map of Question Space, upon which one or more measures of search effectiveness can be imposed.

Search Performance and Question Difficulty

Depending on circumstance and information need, a user's desired answer may be conceived in terms of either timeliness or comprehensiveness. A user looking for a quick answer to a given question needs only a few relevant documents, corresponding to the traditional retrieval performance measure known as *precision* (the proportion of retrieved documents relevant). A different user researching the same question may be looking for a more comprehensive answer, consisting of nearly all available relevant documents, corresponding to the traditional retrieval performance measure known as *recall* (the proportion of relevant documents retrieved). Traditionally, precision and recall are presumed to be inversely related.

Accordingly, different measures of search effectiveness were used to represent these distinct types of information needs: precision at 10 retrieved documents, recall (adjusted for the total number of relevant documents), and a third, composite measure, originally proposed by Meadow. This composite measure (cited in [1], p. 196), compares obtained search results against the classical definition of a "perfect" search--corresponding to 100% precision and 100% recall.

The explicit goal of the TREC program is to advance the state-of-the-art in experimental text retrieval systems. It is reasonable to expect that a particular experimental search

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

engine might be especially well-suited to answering particular types of questions. To minimize the effect of any such system-specific variations, search effectiveness was operationalized using the average performance of six different search engines, all of which were among the best performers on TREC's automatic *ad hoc* test runs.

This average performance was computed individually for each of the three effectiveness measures (precision at 10 documents, adjusted recall, and overall effectiveness), across all 50 questions. Each output set was then ranked in descending order of performance. The top 25 values on a given measure were treated as indicative of "easy" questions, while the bottom 25 were treated as indicative of "hard" questions. Each point on the Question Map was then coded as "easy" or "hard," accordingly. Figure 1 shows the map of Question Space encoded with the results of Meadows' composite effectiveness measure.

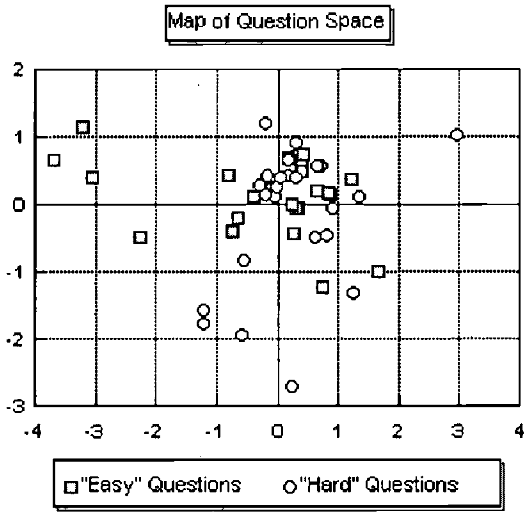


Figure 1. Search Success Displayed in Question Space

Each of the similarity maps thus encoded reveals a visually distinct "patchiness," in which small "islands," consisting mostly of easy (or hard) questions can be identified. This patchiness suggests that the difficulty of a specific question might be accurately predicted based on the difficulty of its nearest neighbors--that is, easy questions tend to be near other easy questions, and vice-versa. For each effectiveness measure, each question was classified as easy or hard, using a crossvalidated nonparametric discriminant analysis, based on its two nearest neighbors.

RESULTS

Classification accuracy was analyzed for each of the three search effectiveness measures, and the statistical significance of the correct classification rate was assessed via t-test. All obtained p values were < 0.01. As summarized in Table 1, the correct classification rates ranged from a low of 68% to a high of 92%. The overall correct classification rate was over 79%.

	Hard	Easy
Precision at 10 documents	72%	92%
Adjusted recall	76%	80%
Overall effectiveness	68%	88%

Table 1. Crossvalidated Correct Classification Rates

CONCLUSION

These results have profound implications for the design of digital document collections. They provide strong evidence for the feasibility and utility of interactive question classifier tools for use in digital collections. Whenever a (probably collection-specific) repository of feedback regarding search effectiveness is available, users could enter a narrative paragraph describing their information need, and optionally even a type of search desired (e.g., "quick" or "comprehensive"). The system could then return both a visual representation and statistical estimate of the likely outcome of the corresponding topical search, based on its analysis of similar questions. Armed with such information, users can make informed assessments regarding the relative costs and potential benefits of actually committing resources to the search itself.

More importantly, these results represent the successful application of quantitative classification techniques to the characterization and prediction of the difficulty of individual question narratives. The overall correct classification rate of nearly 80% was robust across multiple, and presumably inverse, performance measures. Classification accuracy may be subject to further improvement based on additional analytical refinements.

The present study needs to be extended in several important ways. Optimal question classification requires systematic comparative evaluation of similarity measures and scaling methods. Additional research may also reveal specific question features that contribute to search success.

REFERENCES

1. Boyce, B., Meadow, C., and Kraft, D. (1994). *Measurement in information science*. San Diego, CA: Academic Press.

2. Rorvig, M. (1999). Retrieval performance and visual dispersion of query sets. In Voorhees, E., and Harman, D. (Eds.) NIST Special Publication 500-246: *The Eighth Text REtrieval Conference (TREC-8)*.

3. Sullivan, T., Norris, C., and Pavur, R. (2000). Visual question maps as search aids. Presented at the *IEEE Symposium on Information Visualization 2000*.

4. Voorhees, E., and Harman, D. (Eds.) (1997). NIST Special Publication 500-240: *The Sixth Text REtrieval Conference (TREC-6)*.

Browsing by Phrases: Terminological Information in Interactive Multilingual Text Retrieval

Anselmo Peñas
Dpto. Lenguajes y Sistemas Informáticos
UNED,Spain
anselmo@lsi.uned.es

Julio Gonzalo
Dpto. Lenguajes y Sistemas Informáticos
UNED,Spain
julio@lsi.uned.es

Felisa Verdejo
Dpto. Lenguajes y Sistemas Informáticos
UNED,Spain
felisa@lsi.uned.es

ABSTRACT

This paper present an interactive search engine (*Website Term Browser*) which makes use of phrasal information to process queries and suggest relevant topics in a fully multilingual setting.

Categories and Subject Descriptors

Retrieval Issues: *Cross-lingual retrieval, Text Retrieval, Browsing.* Social Issues: *Multilingual access.*

Keywords

Multilingual Information Access, Interaction, Natural Language Processing, Terminology Extraction.

1. INTRODUCTION

In an interactive setting, phrasal information has been used to suggest the user ways of enhancing and refining queries or browsing/classifying search results:

- Handcraft hierarchies based on thesauri (e.g. ERIC) or topic hierarchies (e.g. Yahoo) to browse the document space.
- Automatic building of terminological hierarchies. For instance, automatic clustering of documents into nested classes [3] or subsumption relations between terms [7].
- Extraction of links between documents with similar keywords [4].
- Query expansion with phrases suggested by the system [1].

Most or all of this work has been done only for monolingual retrieval. It is, however, in a multilingual environment where phrasal information is most likely to enhance retrieval, as shown e.g. in [2]: the ambiguity produced by translating separately each term in the query can be greatly reduced by considering possible translations for larger indexing units.

This paper proposes a way of extracting and using phrasal information in an Interactive Multilingual Retrieval environment. The system, "Website Term Browser" (WTB[1]), applies NLP techniques to perform the following tasks:

1. Terminology Extraction and Indexing.
2. Query Processing and Translation
3. Browsing by phrases.

The next sections explain each part of the system in greater detail.

2. TERMINOLOGY-BASED INDEXING

The collection is processed to obtain a large list of terminological phrases. The detection of the phrases in the collection is based on syntactic patterns (figure 1) applied over the tagged documents. The selection of phrases is based on document frequency and term subsumption. Such processing is performed separately for each language (Spanish, English and Catalan in the current version).

1. N N	1. A N [N]
2. N A	2. N N [N]
3. N [A] Prep N [A]	3. A A N
4. N [A] Prep Art N [A]	4. N A N
5. N [A] Prep V [N [A]]	5. N Prep N

Figure 1. Syntactic patterns for Spanish.and English

Rather than relying on lexical dispersion, as in [1], we reuse, in a relaxed way, a terminology extraction procedure [6] originally meant to produce a terminological list to be used by documentalists in a thesaurus construction process. For our purposes, such a list is more useful than the final thesaurus items, which are more conceptual and less related to language usage.

3. QUERY PROCESSING

In query translation for Cross-Language Retrieval, term translation ambiguity can be drastically mitigated by restricting the translation of the components of a phrase into terms that are highly associated as phrases in the target language [2]. This process is generalized in the Website Term Browser as follows:

1. Lemmatized query terms are expanded with semantically related terms in the query language and all source languages using the EuroWordNet lexical database [8].
2. Phrases containing some of the expanded terms are extracted. The number of expansion terms is usually high, and the use of semantically related terms (such as synonyms or meronyms) produce a lot of noise terms. However, the ranking via phrasal information discards most inappropriate combinations, both in the source and in the target languages.
3. Unlike batch cross-language retrieval, where phrasal information is used only to select the best translation for individual terms according to their context, in this process all salient phrases are retained for the interactive selection process.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

4. In a first pass, documents are ranked primarily according to the frequency and salience of the relevant phrases that they contain.

4. BROWSE BY PHRASES INTERFACE

Figure 2 shows the WTB interface. The query process produces a ranking of documents and a ranking of phrasal expressions that are salient in the collection and relevant to the user's query. Both kinds of information are presented to the user, who may directly click on a document or browse the ranking of phrases.

Phrases in different languages are shown to users ranked and hierarchised, according to:

1. Number of expanded terms contained in the phrase. The higher the number of terms within the phrase, the higher the ranking. Original query terms are ranked higher than expanded terms.
2. Salience of the phrase according to their weight as terminological expressions. This weight is reduced to within-collection document frequency if there is no cross-domain corpus to compare with.
3. Subsumption of phrases. For presentation purposes, a group of phrases containing a sub-phrase are presented as subsumed by the most frequent sub-phrase in the collection. That helps browsing the space of phrases similarly to a topic hierarchy.

5. CONCLUSIONS

The Web Term Browser is, to our knowledge, the first interactive search engine that makes use of phrasal information to process queries and suggest relevant topics in a fully multilingual setting. This work should help bridging the gap between research in CLIR algorithms (that use phrasal information to restrict the set of candidate translations) and interactive CLIR, where the focus has been on interactive selection of translation terms and foreign-language document selection [5].

6. REFERENCES

[1] Anick, P. G. and Tipirneni S. The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking. Proceedings of ACM. 1999.

[2] Ballesteros, L. and Croft W. B. Resolving Ambiguity for Cross-Language Information Retrieval. Proceedings of ACM SIGIR'98. 1998.

[3] Hearst, M. A. and Pedersen J. O. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. Proceedings of ACM SIGIR'96. 1996.

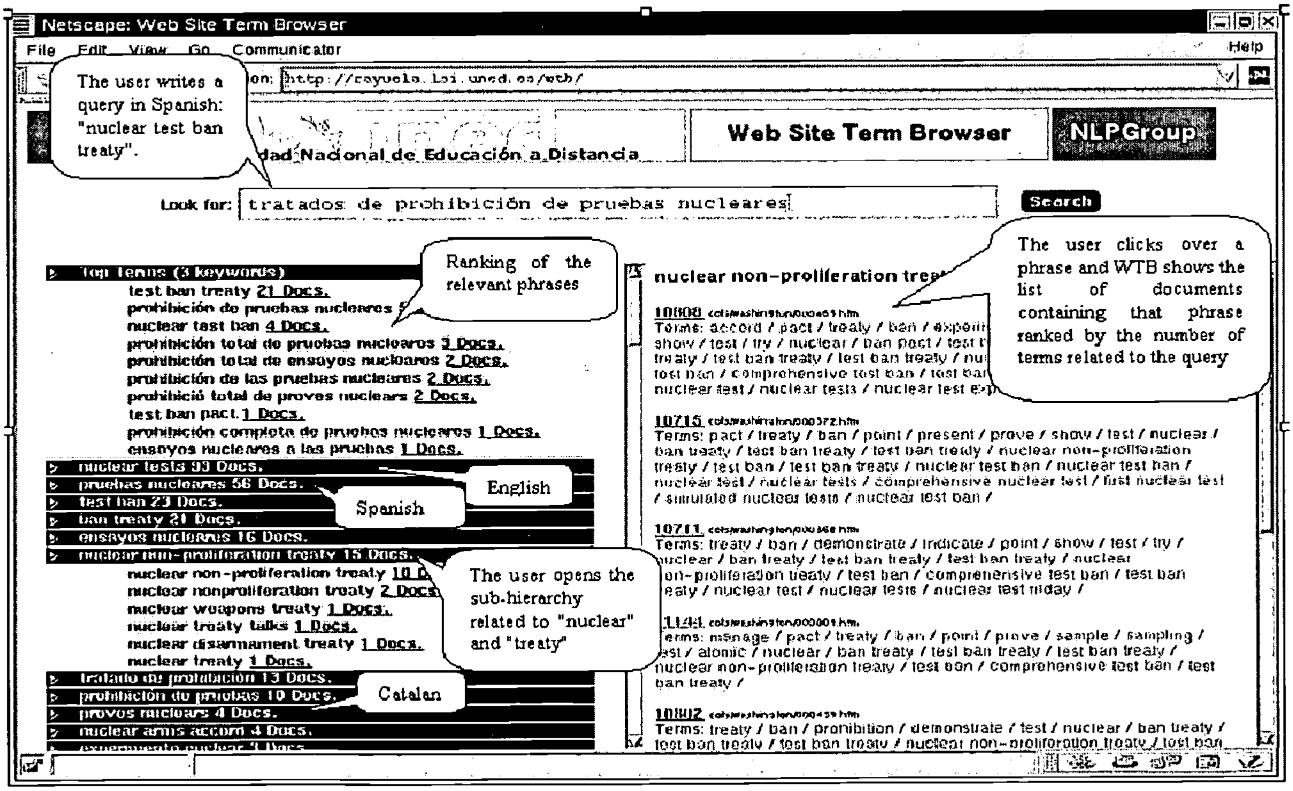
[4] Jones, S. and Staveley M. S. Phrasier: a System for Interactive Document Retrieval Using Keyphrases. Proceedings of ACM SIGIR'99. 1999.

[5] Oard, D. Evaluating Cross-Language IR: Document selection. Proc. CLEF'2000: Springer-Verlag; 2001.

[6] Peñas, A., Verdejo, F. and Gonzalo, J. 2001. Corpus-based Terminology Extraction applied to Information Access. In Proceedings of Corpus Linguistics 2001, Lancaster University, UK.

[7] Sanderson, M. and Croft B. Deriving concept hierarchies from text. Proceedings of ACM-SIGIR'99. 1999; 206-212.

[8] Vossen, P. Introduction to EuroWordNet. Computers and the Humanities, Special Issue on EuroWordNet. 1998.



Netscape: Web Site Term Browser
File Edit View Go Communicator Help

Location: http://rayuela.lsi.uned.es/wtb/

dad Nacional de Educación a Distancia

Web Site Term Browser

NLPGroup

Look for: tratados de prohibición de pruebas nucleares

Search

The user writes a query in Spanish: "nuclear test ban treaty".

Ranking of the relevant phrases

▷ Top Terms (3 keywords)
test ban treaty 21 Docs.
prohibición de pruebas nucleares 5
nuclear test ban 4 Docs.
prohibición total de pruebas nucleares 3 Docs.
prohibición total de ensayos nucleares 2 Docs.
prohibición de las pruebas nucleares 2 Docs.
prohibició total de proves nuclears 2 Docs.
test ban pact. 1 Docs.
prohibición completa de pruebas nucleares 1 Docs.
ensayos nucleares a las pruebas 1 Docs.
▷ nuclear tests 93 Docs.
▷ pruebas nucleares 56 Docs.
▷ test ban 23 Docs.
▷ ban treaty 21 Docs.
▷ ensayos nucleares 16 Docs.
▷ nuclear non-proliferation treaty 15 Docs.
nuclear non-proliferation treaty 10 D
nuclear nonproliferation treaty 2 Docs
nuclear weapons treaty 1 Docs.
nuclear treaty talks 1 Docs.
nuclear disarmament treaty 1 Docs.
nuclear treaty 1 Docs.
▷ tratado de prohibición 13 Docs.
▷ prohibición de pruebas 10 Docs.
▷ proves nuclears 4 Docs.
▷ nuclear arms accord 4 Docs.
▷ experimento nuclear 3 Docs.

English

Spanish

The user opens the sub-hierarchy related to "nuclear" and "treaty"

Catalan

nuclear non-proliferation trea

10808 cols/washington/000465.htm
Terms: accord / pact / treaty / ban / experi
show / test / try / nuclear / ban pact / test t
treaty / test ban treaty / test ban treaty / nu
test ban / comprehensive test ban / test ba
nuclear test / nuclear tests / nuclear test exp

The user clicks over a phrase and WTB shows the list of documents containing that phrase ranked by the number of terms related to the query

10715 cols/washington/000372.htm
Terms: pact / treaty / ban / point / present / prove / show / test / nuclear / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban / test ban treaty / nuclear test ban / nuclear test ban / nuclear test / nuclear tests / comprehensive nuclear test / first nuclear test / simulated nuclear tests / nuclear test ban /

10711 cols/washington/000368.htm
Terms: treaty / ban / demonstrate / indicate / point / show / test / try / nuclear / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban / comprehensive test ban / test ban treaty / nuclear test / nuclear tests / nuclear test friday /

11441 cols/washington/000001.htm
Terms: message / pact / treaty / ban / point / prove / sample / sampling / test / atomic / nuclear / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban / comprehensive test ban / test ban treaty /

10802 cols/washington/000459.htm
Terms: treaty / ban / prohibition / demonstrate / test / nuclear / ban treaty / test ban treaty / test ban treaty / nuclear non-proliferation treaty / test ban

Figure 2. Website Term Browser interface

1. The system is available for testing at http://rayuela.lsi.uned.es/wtb

Approximate Ad-hoc Query Engine for Simulation Data[i]

Ghaleb Abdulla, Chuck Baldwin,
Terence Critchlow, Roy Kamimura,
Ida Lozares, Ron Musick, Nu Ai Tang
CASC, Lawrence Livermore National Laboratory
Livermore, CA 94551
{abdulla1, baldwin5, critchlow, kamimura1,
ilozares, tangn}@llnl.gov

Byung S. Lee, Robert Snapp
Department of Computer Science
University of Vermont
Burlington, VT 05405
(802)656-1919
{bslee,snapp}@cs.uvm.edu

ABSTRACT

In this paper, we describe AQSim, an ongoing effort to design and implement a system to manage terabytes of scientific simulation data. The goal of this project is to reduce data storage requirements and access times while permitting ad-hoc queries using statistical and mathematical models of the data. In order to facilitate data exchange between models based on different representations, we are evaluating using the ASCI common data model that is comprised of several layers of increasing semantic complexity. To support queries over the spatial-temporal mesh structured data we are in the process of defining and implementing a grammar for MeshSQL

KEYWORDS

Mesh Data, Scientific Data Management (SDM), Visualization, Data Integration, Query, Data Retrieval.

1. INTRODUCTION

Scientific data is commonly represented as a mesh. Mesh data is one of the most basic conceptual models for describing physical systems within computer models. A mesh breaks a surface or a volume down into an interconnected grid of 2D or 3D zones, each storing a set of computed variables. If the zones are small enough, the micro-scale properties and interactions can be modeled with sufficient accuracy to provide sufficient predictions of macro scale events. Storage and computation power requirements, however, increase with the number of zones. Current capabilities have simulations running for weeks, if not months, on massively parallel machines and produce meshes in the scale of a few billion zones; a more typical range is between tens of thousands, to tens of millions of zones. Saving these data sets for query processing is not an option because of storage limitations. For an elaborate description of the simulation mesh data please refer to [1].

Querying tera-scale data requires addressing several research challenges including the size of the data, multiple data formats, and supporting complex spatio-temporal queries. We are pursuing a multi-pronged approach to these issues. First, we create a hierarchical partitioning of the data, and model each partition, to create a multi resolution view. Currently we generate a statistical model and a wavelet model. Since obtaining a highly accurate response can require significant time, we provide the capability to trade accuracy for response time. Second, we use metadata associated with these models to facilitate processing the ad-hoc queries. This metadata helps match the user query to the appropriate model, allowing us to generate the most accurate answer within a user-specified error tolerance. We use the term "approximate" for the ad-hoc queries because of the described constraints. Third, we are evaluating a mathematical model that will take into account the relationship between physical systems and mathematics. It considers the relationship between common mathematical entities in simulation and discrete representations of them employed in computer algorithms. This model can be exploited for data management and, in particular, we will use it for query optimization purposes.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

Approximate query answering research is very active for OLAP data (see for example [2]). Chakrabarti et al [3] use wavelets for approximate query answering, however, our data is different in nature (scientific simulation data) and size. We create a multi-resolution index of the data before modeling it, and we use multiple models to support different kinds of queries.

The rest of the paper will describe the current system architecture and the research challenges we hope to address.

2. SYSTEM ARCHITECTURE

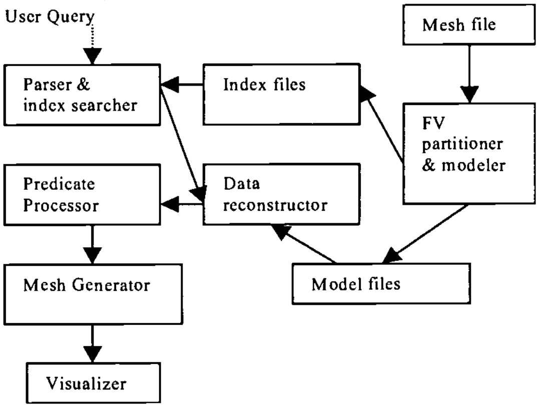


Figure 1 A simplified diagram of the current system architecture

Figure 1 shows a simplified diagram of the current architecture. On the right side of the diagram, we start with a mesh file, which is used to create a matrix of Feature Vectors (FVs), which contains all the spatio-temporal data. The matrix is then partitioned into smaller sets, generating an index tree of nodes.

Each node may have one or more data models describing the current partition. Each model contains the specific parameters required to regenerate the data and the accuracy of the regenerated data has. The results of this initial phase are the index file, describing the partitions, and the associated model data files. The generated files are smaller than the originals, however, they retain the information content of the original data at several resolutions. These files will be used for query processing, while the original data is moved to tertiary storage.

The left side of Figure 1 shows the query engine, from which queries are entered in SQL syntax. The query statement is passed to a parser, and the resulting predicate is used by the index searcher to locate a set of candidate partitions. The partitions are then passed to the Data Reconstructor (DR), which uses model information to reconstruct data points in the required partition. The predicate processor evaluates the user query against these points and creates mesh data that can be viewed by a visualization application. The user query can include functions that use several variables to generate an implicit relationship. The DR uses the information from the user query, metadata, and the error tolerance specified by the user to pick the suitable model for data reconstruction.

3. RESEARCH CHALLENGES

3.1 MODELING AND PARTITIONING ALGORITHMS

Partitioning has an obvious effect on model accuracy, but determining the best partitioning for a collection of models is a challenge. In our approach, a variety of models will be examined - some more for their data compression capability others for their ability to address a wide variety of queries. Optimal model performance may be impacted by the partitioning scheme selected and this interaction needs to be examined. The current partitioning uses an octree-like structure using the spatial and time coordinates as inputs (see Figure 2). We are using the current partitioning method as at test-bed to evaluate the initial set of models, which include wavelets and b-splines:

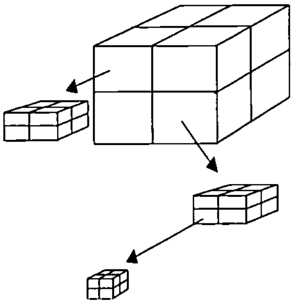


Figure 2 The partition algorithm.

3.2 DATA REDUCTION AND MODELING TECHNIQUES

Different data models are capable of answering different types of queries with different speeds. For example, users might be interested in following a certain region of the data over a certain period. Currently we are using statistical summary and wavelets to model the data. The statistical model allows us to quickly generate the data points, which can then be used to directly answer simple range queries. The wavelet model allows us to easily identify areas of high variability, at varying resolutions, which are often of interest to the scientists.

We will be adding other modeling techniques in the future as needed.

3.3 ERROR METRICS

Since not all queries need to be highly accurate, we will investigate the relationship between the expected error, query speed, and the use of different levels of multi-resolution models. We will perform a series of experiments to determine the difference between the time for theoretical cost models and the observed time. We will use the results to improve error metrics (partition and model errors) at the nodes. We will fine-tune the models and the retrieval algorithms based on the obtained results.

3.4 THE QUERY ENGINE

Scientists often want to perform complex analysis of their data, not just perform simple selections over it, so the ability to include user-defined functions as part of the predicate is required. This increases the complexity of the query engine and requires supporting queries that do not include explicit relationships between variables. The solution is to capture and characterize as much as we can about the models, and heuristically define mappings from queries to the models that can provide for the best answer.

4. CONCLUSIONS

We are currently developing the prototype as described in this paper. Initial results are promising and validate the concepts introduced here. Results imply that we can create a system that will support approximate ad-hoc queries over large data sets. Because of this work we will be able to save the time and space needed to ask complex queries over large simulation data sets.

5. REFERENCES

[1] R. Musick, T. Critchlow, "Practical Lessons in Supporting Large-Scale Computational Science," SIGMOD Record Vol 28 Num 4, Dec 1999.

[2] S. Acharya, P. B. Gibbsons, V. Poosala, and S. Ramaswamy. The aqua approximate query answering system. Proceedings of the ACM SIGMOD, pages 574-576, 1999.

[3] Kaushik Chakrabarti, Minos Garofalakis, Rajeev Rastogi, and Kyuseok Shim. "Approximate Query Processing Using Wavelets," Proceedings of VLDB'2000, Cairo, Egypt, September 2000, pp. 111-122.

[i] This work was performed under the auspices of the U. S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48. UCRL-JC-138042

Extracting Taxonomic Relationships from On-Line Definitional Sources using LEXING

Judith Klavans
Columbia University
540 W. 120th Street
NY, NY, 10027
212-854-7443
klavans@cs.columbia.edu

Brian Whitman
Columbia University
540 W. 120th Street
NY, NY, 10027
212-939-7108
bwhitman@minnowmatch.com

ABSTRACT

We present a system which extracts the genus word and phrase from free-form definition text, entitled LEXING, for Lexical Information from Glossaries. The extractions will be used to build automatically a lexical knowledge base from on-line domain specific glossary sources. We combine statistical and semantic processes to extract these terms, and demonstrate that this combination allows us to predict the genus even in difficult situations such as empty head definitions or verb definitions. We also discuss the use of 'linking prepositions' for use in skipping past empty head genus phrases. This system is part of a project to extract ontological information for energy glossary information.

Keywords

Ontologies, glossaries, definitions, lexical knowledge bases (LKB), information retrieval, natural language processing.

1. USING DEFINITIONS TO BUILD TAXONOMIES

The research we present is part of a larger project to improve access to a set of heterogeneous databases in the Energy Data Collection project [1] of the Digital Government Research Center (www.dgrc.org). As part of this project, a large domain specific ontology serves as the information broker to improve access across databases [2]. Our goal is to merge these domain specific glossary items into a larger ontology. We use glossary definitions since they are by their very nature pertinent only to the domain. For example, the following definition appears in our data set:

> **Motor Gasoline Blending Components:** Naphthas (e.g., straight-run gasoline, alkylate, reformate, benzene, toluene, xylene) used for blending or compounding into finished motor gasoline. These components include reformulated gasoline blendstock f or *{sic}* oxygenate blending (RBOB) but exclude oxygenates (alcohols, ethers), butane, and pentanes plus. *Note:* Oxygenates are reported as individual components and are included in the total for other hydrocarbons, hydrogens, and oxygenates.

Figure 1 - Sample Glossary Definition

Compare this relatively unstructured entry with a typical dictionary definition which has rich internal structure e.g. pronunciation, etymology, sense numbers, etc. In the dictionary analysis projects in [3], over fifty of such fields are identified. To achieve our goals for the EDC project, we must:

1. Identify the definitional material from a web page or other online source (*identification*)
2. Parse the definition for its most salient properties and features (*LKB generation, genus finding*)
3. Incorporate the structured information into a larger ontology, including linking and merging definitions from different agencies and sources.

2. GENUS TERM AND PHRASE FINDING

Our model of building a lexical knowledge base (LKB) from machine readable dictionaries (MRDs) is influenced by the work done in [3] and [4], in which the authors propose a hierarchical structure to represent the complex information found in MRDs. Richly embedded structures containing a head word with subordinate information including cross-references are derived from MRDs. The structured LKB can then be queried as a database and information which was previously inaccessible becomes available. The key feature of this research is that definitions (e.g. car: a vehicle with four wheels) consist of a genus term (vehicle), defined as the main noun or noun phrase which captures an is-a type relation, and differentia (e.g. with four wheels), which details how the defined term differs from related terms (e.g. motorcycle: a vehicle with two wheels).

The method LEXING uses to determine the genus uses knowledge gained from the part-of-speech tagging and noun-phrase (NP) chunking [5] components. We have developed a grammar of phrase identification from a manual study of various definitional sources, and have implemented a evaluation metric for comparing our system's results against a manually tagged set of 500 glossary definitions from 5 different sources.

Definition → (Head Term:) (Definition Text)
Definition Text → (Genus Phrase (GP)) (Remainder)
Remainder → Text
Genus Phrase → NP (of)? (Genus Phrase)
Genus Term → (last noun of first GP NP)

Figure 2 - Genus Phrase and Term Grammar

Since each domain could use domain specific semantic separators, we also introduce the notion of *automatically derived semantic*

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

attributes that are inferred simply from their frequency in the text. LEXING identifies separators such as *having a, used for,* or *containing a*, as "cue phrases" suitable for semantic chunking. As an example, the phrase *having a* was not in our original list of manually-derived separators, but after running a bigram analysis, we discovered its frequency and importance.

Below we show an abbreviated parse of our sample definition, showing the semantic separators *used-for* and *excludes* as well as the genus term (Napthas) and acronym.

```
(term: Motor Gasoline Blending Components
          (full-def: ...)
          (core-def: ...)
          (is-a: Napthas)
          (properties:
               (used-for: blending)
               (exlcludes: oxygenates))
          (acronym: RBOB)
)
```

Figure 3 – Sample Glossary Parse with LEXING

3. RESULTS OF LEXING EVALUATION

We have performed two evaluations on LEXING output:

1. Definition Content Analysis: we ran our system on various definitional sources to determine if our ideas of content were correct, and which fields are frequent.
2. LEXING Accuracy: we evaluated the genus term identification algorithm against a "gold standard".

For evaluation of our representation of definition content, the semantic separator components were tested using both unedited and edited definitions. Our main definition sets came from two gpvernment agencies: the Energy Information Administration (data on energy sources such as gasoline or coal), and the Environmental Protection Agency (glossaries on environmental concerns); we also tested over heart-disease related definitions from definition extraction work done in [6]. Inputs ranged from web pages to flat ASCII documents.

The results in Table 1 show that definitional content hinges largely on the genus. Semantic properties are found in well-edited glossaries (such as EIA edited). Note that the results above do not indicate LEXING's performance. Rather, this evaluation indicates the profile of source definitions in terms of complexity.

Table 1. Definitional Content Analysis

Table 2	Terms	Genus Phrases	Prop-erties	Quant -ifiers	Includes exclude
EIA Edited	19	18 **95%**	15 **79%**	7 **37%**	2 **11%**
EIA Web	127	121 **95%**	38 **30%**	50 **39%**	9 **7%**
EPA Web	1054	1029 **98%**	56 **5%**	24 **2%**	75 **7%**
Medical Auto	90	83 **92%**	0 **0%**	0 **0%**	0 **0%**

For the second evaluation to determine the accuracy of LEXING, we manually tagged 500 definitions, 100 each from 5 domains. (Civil Engineering, Computer Terms, Biomedical Information, General Medical Information, and Energy Information) in order to establish a measurement standard or "gold standard". To compute scoring for genus term identification, a match was defined as both the human tagger and the computer choosing the same genus term. We then computed the accuracy over each definitional set.

Table 2. Genus Term Finding Results

Domain:	Civil Eng	Comp.	Biomedical	Medical	Energy
Genus	93/99	81/101	93/102	100/103	85/102
Term	**94%**	**80%**	**92%**	**97%**	**83%**

Through the first experiment outlined in the above section, we continually isolated and fixed errors related to the semantic attributes and separators. Although some of our definitional sets did not contain as much extractable content as we would have liked, the results on the edited sets were more fruitful. Through this feedback mechanism (run with known set, check results, and fix) we also improved our genus term extraction to achieve the strong results shown in the second experiment.

4. CONCLUSION

We show that a combination of semantic knowledge (the semantic separators) and statistical methods (the automatically-derived semantic attributes) can together provide a methodology for deriving lexical knowledge bases and specifically genus terms, from free-form glossary sources. What is novel is our processing of heterogenous glossary input from different sources to be merged into a large on-line ontology.

5. REFERENCES

[1] Hovy, Eduard, et. al. Simplifying Data Access: The Energy Data Collection (EDC) Project. IEEE Computer 34 (2), Special Issue on Digital Government, February 2001.

[2] Gey et. al. Advanced Search Technologies for Unfamilar Metadata. Proc of 3rd IEEE Metadata Conf. 1999, Bethesda MD.

[3] Byrd, R.J., B.K. Boguraev, J.L. Klavans and M.S. Neff. From Structural Analysis of Lexical Resources to Semantics in a Lexical Knowledge Base. U. Zernik (eds.) Proc of the First International Workshop on Lexical Acquisition, Detroit, Michigian, 1989.

[4] Neff, Mary and Bran Boguraev. Dictionaries, dictionary grammars and dictionary entry parsing. Proc 27th Meeting of the ACL, Vancouver, Canada, 1989.

[5] Evans, D., Klavans, J. and Wacholder, N. (2000) *Document Processing with LinkIT.* RIAO Paris, France, 1336-1345.

[6] Klavans, J.L., Muresan S. DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-Line Text. Proc. AMIA 2000. 1096.

Hierarchical Indexing and Document Matching in BoW

Maayan Geffet and Dror G. Feitelson
School of Computer Science and Engineering
The Hebrew University, 91904 Jerusalem, Israel
{mary,feit}@cs.huji.ac.il

ABSTRACT

BoW is an on-line bibliographical repository based on a hierarchical concept index to which entries are linked. Searching in the repository should therefore return matching topics from the hierarchy, rather than just a list of entries. Likewise, when new entries are inserted, a search for relevant topics to which they should be linked is required. We develop a vector-based algorithm that creates keyword vectors for the set of competing topics at each node in the hierarchy, and show how its performance improves when domain-specific features are added (such as special handling of topic titles and author names). The results of a 7-fold cross validation on a corpus of some 3,500 entries with a 5-level index are hit ratios in the range of 89-95%, and most of the misclassifications are indeed ambiguous to begin with.

1. INTRODUCTION

An obvious and natural approach to organize a large corpus of data is a hierarchical index — akin to a book's table of contents. The type of corpus we deal with is a bibliographical repository, with entries from a limited domain (our prototype is on "parallel systems"). Given such an index, it is desirable that search results point to relevant locations in the hierarchy, rather than just providing a flat list of entries. This is useful not only to support user searching, but also as an aid suggesting possible places to link new entries that are inserted into the repository.

1.1 BoW – Bibliography on the Web

The goal of the BoW project [9] is to create a convenient environment for using and maintaining an on-line bibliographic repository. The key idea is that this be a communal effort shared by all the users. Thus every user can benefit from the input and experience of other users, and can also make contributions. In fact, the system tabulates user activity, so merely searching through the repository and exporting selected items already contributes to the ranking of items in terms of user interest. A prototype implementation is available at http://www.bow.cs.huji.ac.il.

The heart of the BoW repository is a deep (multi-level) hierarchical index spanning the whole domain. The nodes in the hierarchy are called *concept pages*. Pages near the top of the hierarchy represent broad concepts, while those near the bottom represent more narrow concepts. The depth of the hierarchy should be sufficient so that the bottommost pages only contain a handful of tightly related entries (as opposed to Web search engines and scientific literature databases like CORA [5] which contain a relatively shallow directory). A subtrees containing all the concept pages reachable from a certain (high level) concept page is referred to as a *topic*. Entries can be linked to multiple concept pages, if they pertain to multiple concepts. Likewise, they can be linked at different levels of the hierarchy, depending on their breadth and generality.

The index is navigated using a conventional browser. Normally three frames are available (Fig. 1). The first shows the hierarchical index, and the currently selected concept page. The second lists entries linked to this concept page, and allows for the selection of a specific entry. the third displays the surrogate of the chosen entry, including all the bibliographical data (authors, title, where and when published), user annotations, and additional links (e.g. to where the full text is available). Available operations on the current entry include marking it for export, adding an annotation, and adding links. This includes links from additional concept pages to the entry, links between this entry and related entries (e.g. from a preliminary version of a paper to the final version), and links to external resources such as the full text.

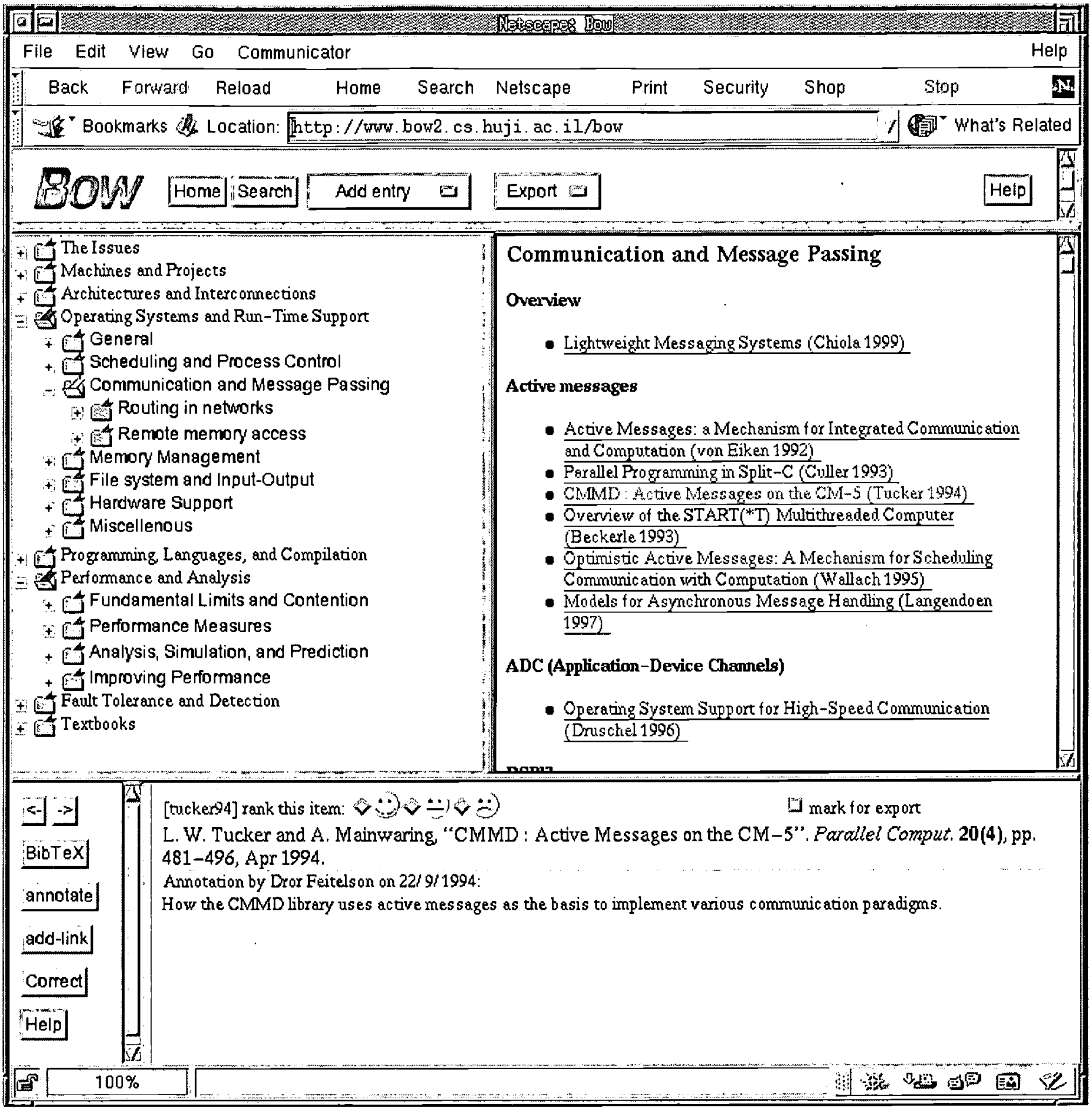
The index structure is created by the site editor. The vocabulary used in the index and annotations is uncontrolled by the system, and users also query the system using natural language [2]. Indexing is simplified by the fact that we use concise surrogates, rather than full text documents [13]. We make up for the reduction in data by enlisting users to verify indexing suggestions. Thus, when a user introduces a new entry, the system uses the text of the entry as a query, and finds concept pages that contain similar entries. But the actual decision to link the new entry to these concept pages is left to the discretion of the user.

The indexing described in this paper is based on lexical analysis of concept pages and entries linked to them. For each topic, we create a list of keywords that differentiate it from other topics that have the same parent. The indexing then proceeds from the root, choosing the most suitable sub-topic(s) at each point. As only contending topics are considered, the complexity of the search is reduced [14, 20].

1.2 Related Work

There are three basic approaches for textual documents processing [15]: lexical, syntactic, and semantic analysis. A number of systems using syntactic and semantic analysis have been developed and are being used for research, such as DR-LINK [18], CLARIT [8] and TREC [7, 31]. However, they are typically not significantly

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.


Netscape: Bow
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.bow2.cs.huji.ac.il/bow
What's Related
BoW
Home
Search
Add entry
Export
Help
<-
->
BibTeX
annotate
add-link
Correct
Help
100%

- The Issues
- Machines and Projects
- Architectures and Interconnections
- Operating Systems and Run-Time Support
  - General
  - Scheduling and Process Control
  - Communication and Message Passing
    - Routing in networks
    - Remote memory access
  - Memory Management
  - File system and Input-Output
  - Hardware Support
  - Miscellenous
- Programming, Languages, and Compilation
- Performance and Analysis
  - Fundamental Limits and Contention
  - Performance Measures
  - Analysis, Simulation, and Prediction
  - Improving Performance
- Fault Tolerance and Detection
- Textbooks

Communication and Message Passing

Overview

- Lightweight Messaging Systems (Chiola 1999)

Active messages

- Active Messages: a Mechanism for Integrated Communication and Computation (von Eiken 1992)
- Parallel Programming in Split-C (Culler 1993)
- CMMD : Active Messages on the CM-5 (Tucker 1994)
- Overview of the START(*T) Multithreaded Computer (Beckerle 1993)
- Optimistic Active Messages: A Mechanism for Scheduling Communication with Computation (Wallach 1995)
- Models for Asynchronous Message Handling (Langendoen 1997)

ADC (Application-Device Channels)

- Operating System Support for High-Speed Communication (Druschel 1996)

[tucker94] rank this item: ◇ ☺ ◇ 😐 ◇ ☹ □ mark for export

L. W. Tucker and A. Mainwaring, "CMMD : Active Messages on the CM-5". *Parallel Comput.* **20(4)**, pp. 481-496, Apr 1994.

Annotation by Dror Feitelson on 22/9/1994:

How the CMMD library uses active messages as the basis to implement various communication paradigms.

Figure 1: Screen dump of BoW showing partially opened hierarchical index.

better than the best lexical analyzers. We will discuss various lexical analyzers throughout the paper, in relation to our work.

Very little has been done so far on hierarchical indexing. In general, it has been shown that hierarchical indexing methods outperform traditional flat algorithms [20, 14]. However, these studies were based on a very wide domain and a relatively shallow hierarchy (e.g. two levels). our work, in contrast, requires a very fine classification, as the bottom levels of the hierarchy only contain a small number of entries each.

Search and browsing based on a hierarchy was suggested in [24]. However, in this case the hierarchy is very strict and depends on nested key phrases (e.g. "forest fires" is under "forest"), which allows it to be automated. We take the opposite approach: the hierarchy is created by humans so as to capture pertinent concepts, and the automation comes in trying to find what characterizes this structure.

2. OFF-LINE PREPARATION OF KEYWORD VECTORS

The hierarchical indexing mechanism consists of two parts. The first is an off-line traversal of the whole repository, repeated at reg-

Topic	Number of clusters	Hit ratio	
		5-grams	Whole words
Cooking recipes	10	87%	53%
Linux	16	85%	47%

Table 1: Clustering hits ratio for two given documents collections using 5-grams versus whole words.

ular intervals (e.g. once a day) in order to compute keyword vectors for all the topics. The second is a matching scheme that compares new entries or queries with these pre-computed keyword vectors.

The off-line part is executed recursively for every level of the index, top-down. The main idea is that each topic encompasses all the concept pages in a sub-tree of the index, therefore all of them should be taken into account while constructing its keywords vector. The group of sibling topics, located at the same level and having the same parent in the index are called a *competitive topics set*, since they compete for keywords with each other. The algorithm generates keywords vectors in five steps: *parse* all the pages in the topic's sub-tree, *merge* them into one vector, *unify* the resulting vectors to include the same words, *normalize* the weights of the words in all the vectors, and *choose* the most relatively frequent ones to represent the corresponding topics.

2.1 Parsing

The first stage is parsing the text of concept pages, with the goal of creating a vector of all the words in the given concept page [30, 33], denoted by Voc_{page}. This of course requires us to define "word".

The natural definition is a completely separated meaningful string. This has the well-known disadvantages of treating related words as being different, and the well-known solutions such as stemming (e.g. [23, 19]). An alternative is to use n-grams (substrings of length n of words: for example, "algorithm" will be turned into "algor", "lgori", "gorit", "orith", and "rithm") [1]. We prefer the latter, and specifically use 5-grams, based on a separate study[1] in which documents were clustered automatically based on similarity and this was compared with manual clustering (Table 1). But in order to avoid 5-grams that are largely based on common suffixes and therefore meaningless, we also use stemming first.

Note that longer words are represented by more 5-grams in the vocabulary vector than shorter ones, which gives them more weight in the comparisons. Thus it would be interesting to check if similar results would be obtained by using whole words, and weighting them according to length.

In any case, from now on the word "word" will mean a 5-gram.

2.2 Merging

After parsing all the concept pages in a topic's sub-tree, the resulting vocabulary vectors are merged. The resulting vector includes the complete vocabulary of the topic:

$$Voc_{topic} = Voc_{page1} \cup Voc_{page2} \cup \ldots \cup Voc_{pagen}.$$

The counters indicating how many times each word appears are summed as described below.

2.3 Unification

In order to compare a query with a set of competitive topics, the vocabulary vectors of these topics must span the same space. We therefore create a unified vocabulary that includes all the words that appear in any of the competitive topics:

$$Voc_{compet-set} = Voc_{topic1} \cup Voc_{topic2} \cup \ldots \cup Voc_{topick}.$$

We then normalize the vocabulary vectors of the individual topics to include all these words, by adding the missing ones with a count of zero. The resulting normalized vectors will be denoted by $NormVoc_{topic}$.

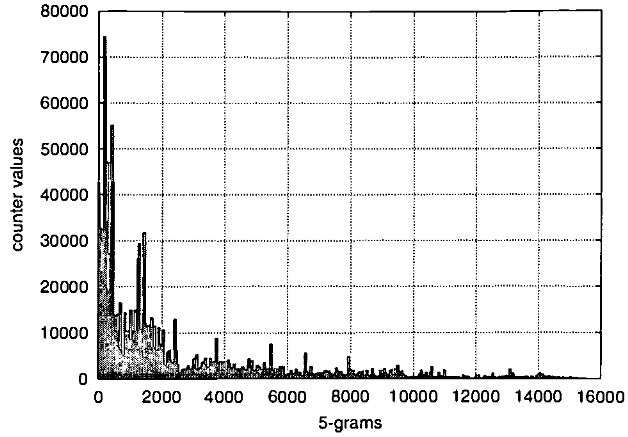


Figure 2: Example of counter values for 5-grams in the vocabulary of a top-level topic.

2.4 Counters Normalization

In order to select meaningful keywords, we need to consider the number of times each word appears in each topic. As shown in Fig. 2, these values vary considerably. But they also suffer from the scaling effect problem [15]: the counter values in "small" topics are generally lower than in "big" topics, leading to an assignment of all the keywords to the bigger topics. To compensate for this, we need to normalize the counters based on the size of each topic.

The simplest approach is to divide the counter values by the total number of the words in the topic. However, according to Zipf's formula [34], $rank \times count \approx constant$ (where the words in the text are ranked in order of decreasing count), so the number of distinct words in the text grows much slower than their counts. Practically, about 50% of the regular text content consists of the same 250 words [15]. Therefore this method does not lead to good normalization (Fig. 3).

The most popular algorithm is TFIDF (Term Frequency Inverse Document Frequency) [27, 26, 6]. However, this technique does not take into account the frequency of term occurrences in other documents in the collection, based on the assumption that there are very many documents. In our case, we are trying to distinguish between a small set of topics, so an adjustment is needed. When applied directly, TFIDF did not produce good results (Fig. 3).

Our chosen approach is to normalize the counters on-the-fly during the previous three steps. Since we are interested in defining a topic's vocabulary, words which occur frequently in one particular entry within it should not have a higher weight. Thus, we count each word only once for every entry containing it in the concept page. For example, given a topic with 5 entries, the maximal weight of a word is 5 if it appears in all the entries, but if it appears twice in one entry and three times in another, its weight will only be 2. This normalization is implemented as part of the parsing algorithm. To

[1] In cooperation with E. Boncheck.

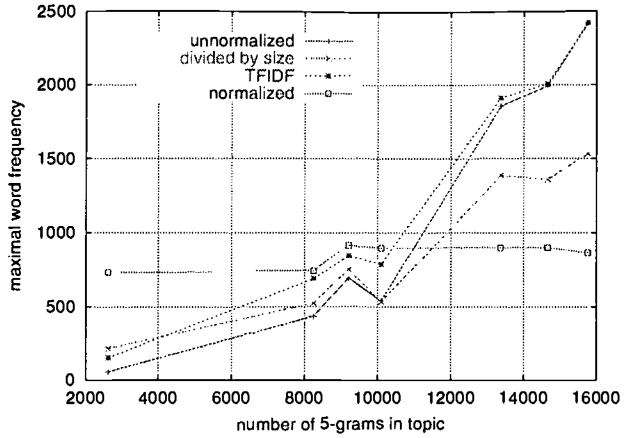


Figure 3: The influence of various normalization strategies on the 5-grams frequencies in the top level competitive set of 7 topics.

deal with the fact that concept pages have different sizes, the counters are further normalized by dividing by the number of entries in a page or topic. This is done as part of the merging and unification.

The comparative results of this method are illustrated in Fig. 3. As shown in the graph the maximal weights have reached the uniform distribution irrespective of the topic size.

2.5 Keywords Selection Heuristic

A keyword is a word that characterizes a concept and differentiates one topic from others [15]. Thus, in order to decide whether a word is a keyword of some topic, one should consider its frequency (weight) in this topic, and also compare with its weights in all the competitive topics. The basic idea is that if a word is extremely frequent in one particular topic and relatively rare in others, then we may use it as a keyword for this topic. If a word has similar weight in all the topics, then it does not represent any of them, even if its weight is high [29].

One way to assess the discriminatory power of a word is based on the difference between its maximal and minimal counter values in different topics in the competitive set. More formally, the algorithm is as follows (where $NormVoc_t(w)$ denotes the counter value for word w in the normalized vector of topic t):

1. For each topic in the competitive set, find those words that achieve their maximal counter value in this topic:

   $$Max_t = \{w | \forall i,\ i \neq t :\ NormVoc_t(w) > NormVoc_i(w)\}.$$

2. For these words, find the range of counter values:

   $$\forall w, w \in Max_t,$$
   $$Dif(w) = max_{i \neq t}\{(NormVoc_t(w) - NormVoc_i(w))\}.$$

3. sort the words in Max_t according to $Dif(w)$ in a descending order.

4. Choose the top 10% of the words (those with the biggest difference values) and place them in the keywords vector $Tkeys_t$.

A possible problem with this definition is that the difference can be large because the minimal value is very small. An alternative is therefore to use the difference between the two top counter values in step 2. The definition then becomes

$$Dif(w) = NormVoc_t(w) - max_{i \neq t}\{NormVoc_i(w)\}.$$

Heuristic used	Hit ratio
extreme values differences	48%
two highest values differences	62%
if (max_weight > avg+std_dev)	87%

Table 2: Correct classification rate when using alternative heuristics for keyword selection.

This version selects the words with significantly greater weight in one particular topic than in all the others, but may miss cases in which a word has a high count in 2 or 3 topics (which may happen as shown in Fig. 4). Specifically, in the BoW corpus the gap between the two highest values is the largest in 65-79% of the cases, but the gap between the 2nd and 3rd is the largest in another 15-22%.

Another disadvantage of this heuristic is the percentage of words to be chosen as the most significant: we decided to choose an empirically-determined 10% threshold, but maybe for other repositories it will be reasonable to use another threshold. An alternative is to choose the most significant words according to their statistics. Specifically, we propose to select those words whose counter value is larger than the average plus one standard deviation:

1. For each word calculate the average counter value:

   $$average(w) = \frac{1}{n} \sum_{1 \leq i \leq n} NormVoc_i(w)$$

   (where n is the number of topics in the competitive set).

2. Calculate the standard deviation:

   $$std\_dev(w) = \frac{1}{n}\sqrt{\sum_{1 \leq i \leq n} (NormVoc_i(w) - average(w))^2}.$$

3. If $max\_weight(w) > average(w) + std\_dev(w)$ then the word w is a keyword of the maximal weight topic, otherwise it does not represent any topic since it is almost equally frequent in all of them.

To check if the word should be a keyword for other topics as well, the highest value is removed and the procedure repeated for the remaining topics.

To compare the above heuristics we used them to classify 200 entries from the BoW prototype repository. The results are shown Table 2, and indicate that the last heuristic (using the average and standard deviation) is the best.

2.6 Optimizations

2.6.1 Stop-lists

A well-known optimization in classifications based on lexical analysis is the definition of a stop-list — a list of common words that should be ignored. In order to generate the list automatically, a threshold distinguishing the most common words should be found. Numerous studies of documents show that 30% of general English text encompassing millions of words is made up of only 18 distinct words [15]. Usually, stop-lists contain about 250-300 terms [32, 25, 10]. However, our repository is limited to a focused scientific domain, so its language is rather limited, and may vary among topics. Thus, the stop-list should contain only those words which are

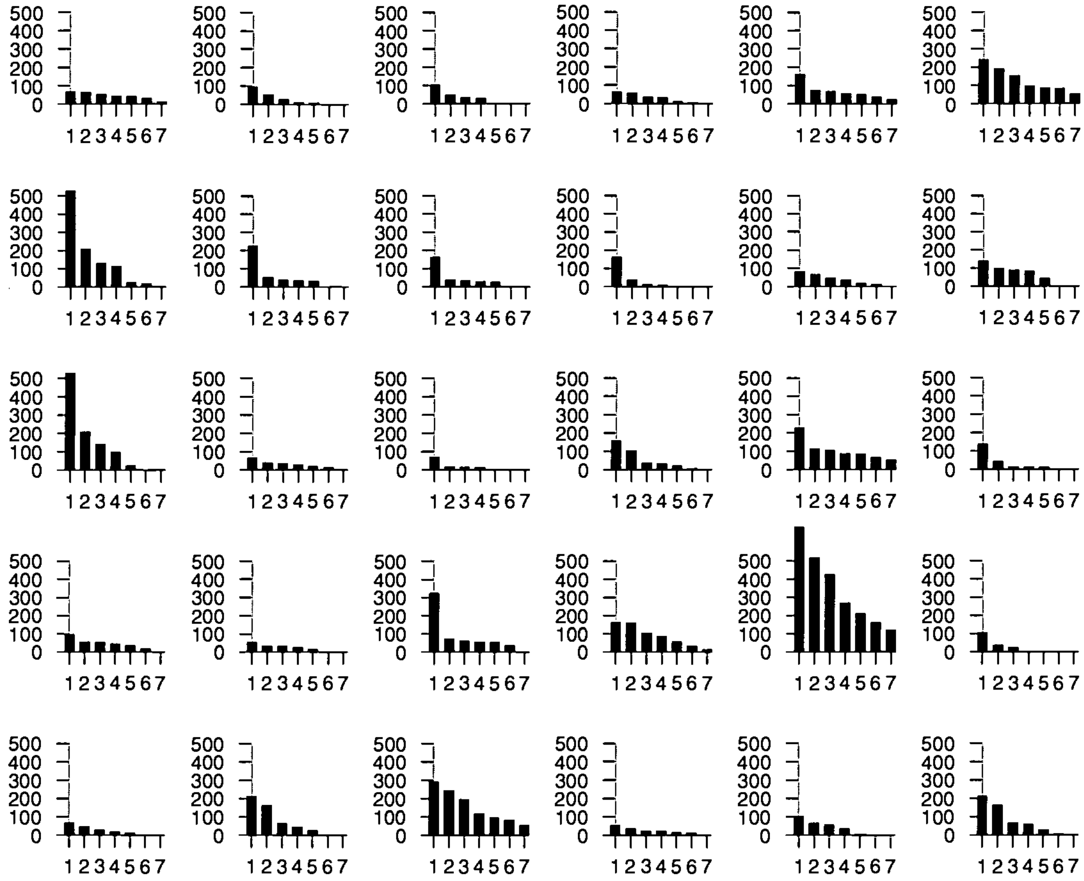


Figure 4: Distribution of word counts in the top level 7 topics for 30 selected 5-grams. The counters are sorted in a descending order.

common in all the topics in the competitive set. This leads us to the following method for stop words identification:

$$stop\_list = \{w | \forall i, NormVoc_i(w) > \tau\},$$

where τ is an empirically selected threshold on the counter values.

According to our observation of the common words values distribution, the upper value at the top level is greater than at lower levels. The best τ value is 80 for the highest level of the hierarchy, 60 for the second one, and 50 for the rest, where the average maximal counters are 320, 240 and 200, respectively. Thus the empirically obtained rule is that a stop-words lower bound threshold is a quarter of the average maximal frequencies for the given competitive set of topics.

Another interesting question is whether the stop-lists at various locations in the hierarchy will differ. It is reasonable to expect that words like "network", "software", and "language" will be important at the highest level of the hierarchy, since each of them leads to an appropriate broad topic, such as "architecture and interconnections", "operating systems and run-time support", and "programming, languages, and compilation" (see Fig. 1). Obviously, inside the topic "programming, languages, and compilation" the words "programming" and "languages" should be the first ones to go to the stop-list. However, our observation of the parallel systems repository has shown that most of the stop-words at all the levels were the same, while for every lower level several additional common stop-words were added. The total number of stop-words is around 200 with slight differences for various competitive sets.

2.6.2 Special Treatment for Selected Fields

Another means for optimization is using domain-specific knowledge. In our case the domain is a bibliographical repository, which is classified into topics. Thus special fields like authors names and topic titles may carry special significance.

For example, the topics and sub-topics title fields may be expected to reflect the contents of the topic, and this is based on a semantic understanding by a human editor. It is therefore desirable to use these words as keywords, even if the counter-based algorithm described above does not recognize them as such.

The special treatment of author names is founded on the assumption that usually scientists tend to concentrate their work in a rather narrow area of research. Therefore if several of the given author's publications appear in one specific topic of the competitive set, but not in the others, then it is sensible to suggest that the new article will also belong to this topic. As most of the author names appear too rarely and thus do not survive the keyword filtering process, special treatment is required. Just as in the case of topic titles, we simply treat author names explicitly as keywords. For this purpose, the first and last names are concatenated and treated as a single term.

2.6.3 Thesauri

The final major problem to be considered here is the use of similar or related terms (synonyms). Thus the use of thesauri in order to recognize variants or to control the vocabulary has been suggested [3]. A specific feature of our index is that it contains a lot of names of projects, systems, and tools, which are often referred to by acronyms. Text observations show that typically such terms occur in one of the following formats at least once [16]:

1. The full term words with capital letters and then the acronym consisting of the same first capital letters in parenthesis.
2. The acronym is followed by the parenthesized full term words interpretation.

Based on this we developed a thesaurus-builder which is responsible for lexical text analysis and extracting the full expressions and their acronyms, and used it to construct a dictionary of acronyms. This was used during parsing to check if the acronym or its interpretation occur in any particular concept vocabulary, and if so it was explicitly entered into the keywords vector. User queries are also checked against the thesaurus, and expanded in a similar manner.

3. ON-LINE SEARCHING

Given the keyword vectors for all the repository's topics, those matching queries can be found. This is done in two cases: when a user issues a search by specifying authors and/or keywords, and when a user inserts a new entry into the repository. In this latter case, the goal is to recommend topics to which the new entry may be linked.

An important goal is that a retrieved set will be of "reasonable" size — large enough to give the user a choice but not too large. BoW therefore doesn't retrieve a set of individual documents in response to a query. Instead, it returns whole concept pages. Moreover, if many of these concept pages belong to the same higher-level topic, that topic is returned rather than listing the lower level ones.

3.1 Matching and Ranking

Matching and ranking go together — we want to find the topics that match the query to the highest degree. Several methods for such ranking exist [21]. The most popular are based on the TFIDF algorithm described in section 2.4 [26, 30, 28, 27] and will be rejected here for the same reasons. An alternative approach which is usually used in clustering (e.g. in Isodata Clustering) is to compute the distances between the keyword vectors. This can be applied in our case, by comparing the distances between the query vector and the competitive set vectors. However, the query is typically so short that it is not reasonable to weight its terms [11] so the terms relative frequencies distance between the query and the index vectors is not useful in our model. Thus we have to use a boolean ranking method [17], rather than a vector space algorithm.

Our matching process works as follows:

1. Check the query data against the acronyms thesaurus, and insert both acronyms and their full interpretation into the initially empty query vocabulary vector $QVoc$.
2. Parse the query (or new entry) and insert the resulting 5-grams into the vocabulary vector $QVoc$ (with no terms weights considerations).
3. Starting from the highest level topics, measure the similarity of the query to all topics in the competitive set by counting the number of common words in the vectors:

$$score_{topic} = | QVoc \cap TKeys_{topic} | .$$

4. Select the topics with the highest score, and continue recursively to lower levels. The selection criterion is that the score be higher than the average plus a standard deviation, as was done in section 2.5. This gives good results because in 84%-91% of the queries the biggest gap is between the highest and the next topic, or between the second highest and the third one (Fig. 5).

Note that we don't examine all the tree branches, but only those which survive the filtering criteria, thus reducing the computational cost. This technique, called tree pruning, was also employed by others [14, 20], except that they choose only the single most suitable sub-topic at each level. The main disadvantage of such aggressive "single-path" pruning is that a failure at one of the higher levels will cause all the classification process to fail, whereas pruning that keeps two or three branches for further examination attains almost the same accuracy as full tree evaluation. Therefore, our ranking scheme does not suffer from the irrecoverable errors occurrence problem. Choosing more than one also meets our expectation that an article may refer to several categories in the bibliography.

3.2 Output Representation

Observe that the total number of selected topics may grow exponentially while descending the tree, if most subtopics are selected at each stage. To avoid showing the user such a long list of hits, we replace them all by their shared father. As the result, the more general (higher level) topic will be returned to the user. The condition for such output compression is that at least 50% of the particular topic's children and more than two of them are in the resulting list. The compressing routine is performed recursively from the bottom to the root of the index. The results of output compression are demonstrated in Fig. 6. The output was compressed for about 25% of the queries, where the majority of the compressed output sets were those including 14 links and more, only 10% of them remained untouched. On the other hand, only 10% of smaller sets (up to 13 links) were compressed. The compression ratio is quite big, and the size of compressed output sets was decreased by half in average.

Given the topics selected by the ranking process, and remaining after output compression, the question is how to display them on the screen. The dilemma is how to reconcile two contradicting considerations: keep both the concept pages' topological locations in the hierarchy (as in the Berkeley Cha-Cha Search Engine [4]), and their respective ranking with regard to this query (as is typically done in search engines, e.g. Northernlight [22]). Our solution is to display the original index tree, with the selected links opened and marked with different colors and font sizes according to their relevance to the query.

4. EVALUATION

In order to check the final algorithm performance we have conducted a sequence of 5 experiments employing 7-fold cross validation over a corpus of about 3,500 bibliographic entries. The corpus is focused on the domain of parallel systems, with an index that has an average depth of 5 and an average branching factor of 6. Every experiment was based on about 500 randomly chosen entries, which were extracted from the repository. The automatic off-line

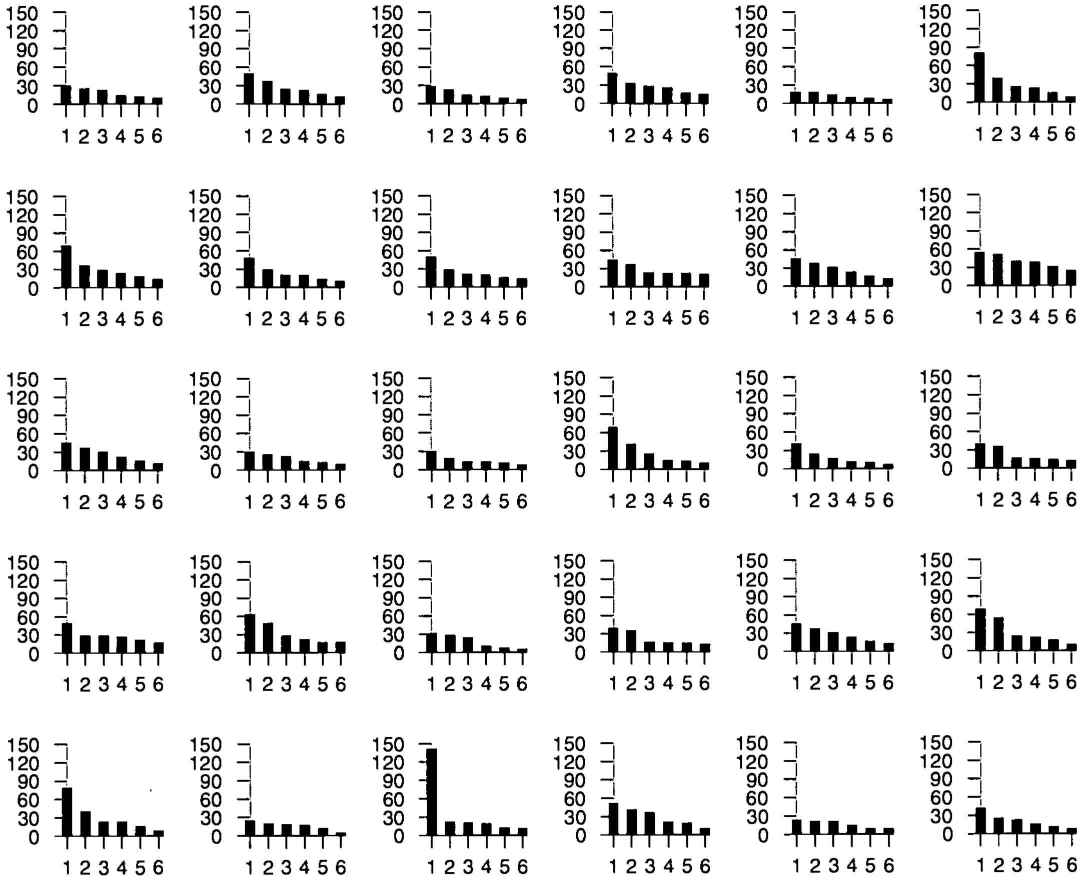


Figure 5: Distribution of topics scores (in a competitive set of 6 topics) for 30 sample queries. The topics are sorted in a descending order for each one.

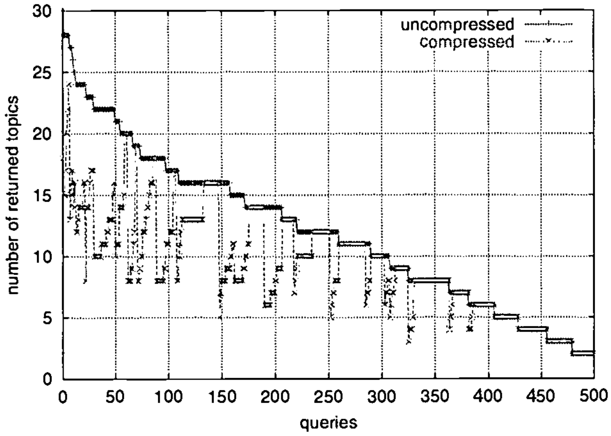


Figure 6: Results of the compression procedure for 500 queries from one of the 7-fold cross validation experiments (described below), sorted in descending order by the uncompressed output sets sizes.

indexing was performed on the remaining 3,000 entries, and the resulting keyword vectors used to re-insert the 500 entries that were extracted. The hit ratio for each case was computed by comparing the algorithm's classification of these entries with their original manual classifications (Fig. 7). Manually checking those that were misclassified revealed that in many cases they were indeed ambiguous, and had very short annotations that only included very general terms.

Our experimental results have corroborated those of McCallum et al. that larger vocabulary sizes generally perform better. For larger branches of the index our algorithm selects more keywords, and the classification reached its highest accuracy (near 100%). For example, the "Operating Systems and Run-Time Support" topic, which is one of the biggest topics in the repository with over 7,700 distinct five-grams vocabulary, got 100% hit ratio, whereas "Algorithms and Applications" which is a smaller topic, containing about 3,700 keywords, attained only 92% hit ratio. Another evidence is the decrease in hits percentage for lower levels, due to the smaller number of entries and therefore the smaller number of keywords, as shown in Fig. 8.

Generally, the results indicate that the more information is available about each concept and each query, the better the matching

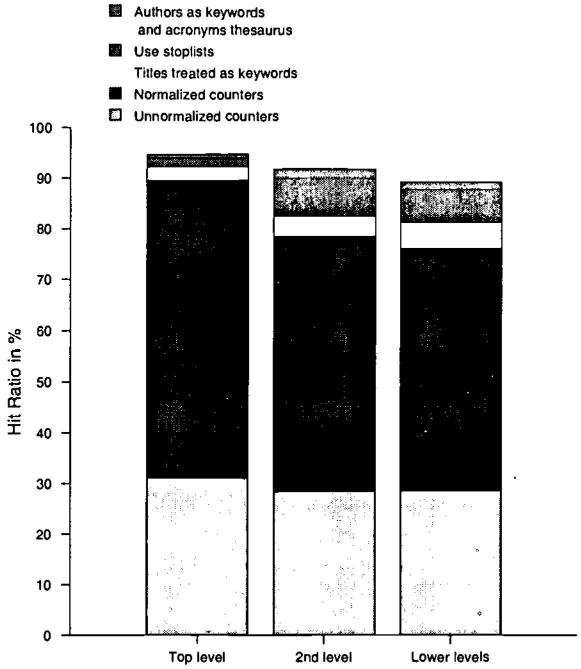


Figure 7: The hit ratios achieved at different levels of the index hierarchy, and how they depend on different parts of the classification algorithm.

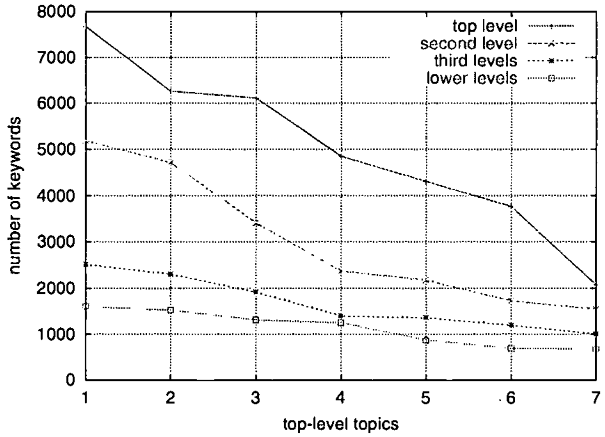


Figure 8: Distribution of keywords under the 7 top-level topics (which are sorted by size).

that is achieved. However, we find that even a relatively short annotation of 2-3 lines is enough for a reasonably good classification.

5. CONCLUSIONS

We have developed and presented the details of a data classification algorithm for effective concept-based storage and retrieval of scientific papers in multi-level hierarchical repositories. The three main features of the algorithm are its homogeneity, scale independence, and self-updateability. The algorithm is homogeneous in that it produces good results at all levels of the hierarchical index, and does not depend on the index depth. It is scale independent due to the normalization of the keyword vectors, resulting in fair judgments for various-sized concept pages. It updates the keyword vectors regularly, thus keeping them current and adjusting to changes in the repository contents. This is done at selected intervals, rather than on-line for each new entry, because every local change in an individual concept page causes changes in the entire topic's vocabulary, and so in the selection of keywords across the entire competitive set; moreover, this effect can propagate up the hierarchy.

Results of experimentation with the BoW prototype repository on parallel systems are very promising. At the top level, nearly 95% of the entries were classified correctly, and this dropped to just under 90% for the lowest levels. Remarkably, this was achieved with only the entry details (mainly title and authors), and very short annotations typically between one and three sentences long. There was no access to or use of full text. The entries that were misclassified were found to be ambiguous and had short or missing annotations.

In the future we hope to test our algorithm on additional repositories. Possible extensions include automatic construction of a full thesaurus for all the words and phrases in the given corpus. A bigger challenge is automatic index creation from scratch. Our suggestion is to use one of the hierarchical clustering methods [12] combined with the described automatic indexing algorithm.

Acknowledgements

This work was supported by the Ministry of Science, Culture & Sport under the program to develop scientific and technological infrastructure.

6. REFERENCES

[1] Adamson, G. and J. Boreham. 1974. "The use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles," *Information Storage and Retrieval*, 10, 253-260.

[2] Blair, David. C. 1990. *Language and Representation in information retrieval.* N.Y.: Elsevier.

[3] Chen, H., Martinaz, J., Kirchhoff, A., Ng, T. D., and Schatz, B. R. "Alleviating search uncertainty through concept association: automatic indexing, co-occurrence analysis, and parallel computing." *J. Am. Soc. Inf. Sys.* 49(3), 216-206.

[4] Chen, M., Hearst, M., Hong, J., and Lin, J., 1999. "Cha-Cha: A System for Organizing Intranet Search Results". In *2nd USENIX Symp. Internet Technologies & Systems.*

[5] CORA - Computer Science Research Paper Search Engine, http://cora.whizbang.com

[6] Doszkocs, T.E. 1982. "From Research to Application: The CITE Natural Language Information Retrieval System," in *Research and Development in Information Retrieval*, eds. G. Salton, and H. J. Schneider, pp. 251-62. Berlin: Springer-Verlag.

[7] Dumais, Susan T. 1995. "Latent semantic indexing (LSI): TREC-3 report." In *Overview of 3rd Text Retrieval Conference (TREC-3).* Donna K. Harman, ed. 1995. Washington, D. C.: Nist Special Publication.

[8] Evans, David A., Robert G. Lefferts, Gregory Gregenstette, S. Henderson, William Hersh, and A. Archbold. 1993. "CLARIT TREC design, experiments, and results." In *1st Text Retrieval Conference (TREC-1).* ed. Donna K. Harman, pp. 251-286. 1993. Washington, D. C.: Nist Special Publication, 500-207.

[9] Feitelson, D. G., 2000. "Cooperative Indexing, Classification, and Evaluation in BoW". In *7th IFCIS Intl. Conf. Cooperative Information Syst.*, O. Etzion and P. Scheuermann (eds.), Springer-Verlag LNCS Vol. 1901, pp. 66-77.

[10] Fox, C. 1990. "A Stop List for General Text". *SIGIR Forum.*

[11] Frakes, W. B., Baeza-Yates, R. eds. 1992. *Information Retrieval - Data Structures and Algorithms*, Englewood Cliffs, N. J.: Prentice Hall.

[12] Jardine, N. and C. J. vanRijsberngen. 1971. "The Use of Hierarchic Clustering in Information Retrieval." *Information Storage and Retrieval*, 7(5), 217-40.

[13] Kerner, C. J., and T. F. Lindsley. 1969. "The value of abstracts in normal text searching." In *The information bazaar: Proc. 6th Ann. Nat'l Colloq. Information Retrieval*, Philadelphia, pp. 437-440.

[14] Koller, D., and Sahami, M. 1997. "Hierarchically classifying documents using very few words". In *Proc. 14th Int'l Conf Machine Learning (ML-97)*, pp. 170-178, Nashville, Tennessee.

[15] Korfhage, R. R., 1997. *Information Storage and Retrieval*, N.Y.: John Wiley and Sons.

[16] Larkey, Leah S., Ogilvie, Paul, Price, A. Andrew, and Tamilio, Brenden, 2000. "Acrophile: an automated acronym extractor and server". In 5th *ACM Conf. Digital Libraries*, pp. 205–214.

[17] Lee, Joon Ho, Myoung Ho Kim, and Yoon Hoon Lee. 1993. "Ranking documents in thesaurus-based Boolean retrieval systems." *Information Processing & Management* 30(1), 79-91.

[18] Liddy, Elizabeth D., and Sung H. Myaeng. 1993. "DR-LINK's linguistic-conceptual approach to document detection." In *1st Text Retrieval Conference (TREC-1)*. Donna K. Harman, ed. Washington, D. C.: Nist Special Publication, 500-207, pp. 113-130.

[19] Lovins, J. B. 1968. "Development of the Stemming Algorithm." *Mechanical Translation and Computation Linguistics*, 11(1-2), 22-23.

[20] McCallum, A., Rosenfeld, R., Mitchell, T., and Ng, A. Y. 1998. "Improving Text Classification by Shrinkage in a Hierarchy of Classes". In *Proc. 15th Int'l Conf Machine Learning (ML-98)*, Madison, Wisconsin.

[21] McGill, M. et al. 1979. *An Evaluation of Factors Affecting Document Ranking by information retrieval systems*. Project report. Syracuse, New York: Stracuse University School of Information Studies.

[22] The Northernlight Search Engine, http://www.northernlight.com.

[23] Paice, Chris. 1990. "Another stemmer." *SIGIR Forum* 24(1), 53-61.

[24] Paynter, G. W., I. H. Witten, S. J. Cunningham, and G. Buchanan, "Scalable browsing for large collections: a case study". In 5th *ACM Conf. Digital Libraries*, pp. 215–223, Jun 1999.

[25] Salton, G., and M. McGill. 1983. *Modern Information Retrieval*. New-York: McGraw-Hill.

[26] Salton, G. and C. S. Yang. 1973. "On the Specification of Term Values in Automatic Indexing." *J. Documentation* 29(4), 351-72.

[27] Salton, G. 1971. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Englewood Cliffs, N. J.: Prentice Hall.

[28] Salton, G. and C. Buckley. 1988. "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management* 24(5), 513-23.

[29] Salton, G., H. Wu, and C. T. Yu. 1981. "The Measurement of Term Importance in Automatic indexing." *J. Am. Soc. Inf. Sys.* 32(3), 175-86.

[30] Salton, G. and J. Allan. 1994. "Text retrieval using the vector processing model." *Proc. 3rd Ann. Symp. Document analysis and information retrieval*, Las Vegas, Nevada, pp. 9-22.

[31] Smeaton, Alan F., R. O'Donnel, and F. Kelledy. 1995. "Indexing structures derived from syntax in TREC-3: system description". In *Overview of 3rd Text Retrieval Conference (TREC-3)*. Donna K. Harman, ed. 1995. Washington, D. C.: Nist Special Publication.

[32] vanRijsberngen, C. J. 1975. *Information Retrieval*. London: Butterworths.

[33] Wong, S.K.M., and W. Ziarko. 1985. "On generalized vector space model in information retrieval." *Annals of the Society of Mathematics of Poland, Series 4: Fundamentals of information* 8(2), 253-267.

[34] Zipf, George Kinglsey. 1949. *Human behavior and the principle of least effort*. Cambridge, Massachusetts: Addison-Wesley.

Scalable Integrated Region-based Image Retrieval using IRM and Statistical Clustering*

James Z. Wang[†]
School of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16801
jwang@ist.psu.edu

Yanping Du[‡]
School of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16801
ydu@cse.psu.edu

ABSTRACT

Statistical clustering is critical in designing scalable image retrieval systems. In this paper, we present a scalable algorithm for indexing and retrieving images based on region segmentation. The method uses statistical clustering on region features and IRM (Integrated Region Matching), a measure developed to evaluate overall similarity between images that incorporates properties of all the regions in the images by a region-matching scheme. Compared with retrieval based on individual regions, our overall similarity approach (a) reduces the influence of inaccurate segmentation, (b) helps to clarify the semantics of a particular region, and (c) enables a simple querying interface for region-based image retrieval systems. The algorithm has been implemented as a part of our experimental SIMPLIcity image retrieval system and tested on large-scale image databases of both general-purpose images and pathology slides. Experiments have demonstrated that this technique maintains the accuracy and robustness of the original system while reducing the matching time significantly.

Keywords

Content-based image retrieval, wavelets, clustering, segmentation, integrated region matching.

1. INTRODUCTION

As multimedia information bases, such as the Web, become larger and larger in size, scalability of information retrieval system has become increasingly important. According to a report published by Inktomi Corporation and NEC Research in January 2000 [13], there are about 5 million unique Web sites (± 3%) on the Internet. Over one billion web pages (± 35%) can be down-loaded from these Web sites. Approximately one billion images can be found on-line. Searching for information on the Web is a serious problem [16, 17]. Moreover, the current growth rate of the Web is exponential, at an amazing 50% annual rate.

1.1 Image retrieval

Content-based image retrieval is the retrieval of relevant images from an image database based on automatically derived features. The need for efficient content-based image retrieval has increased tremendously in many application areas such as biomedicine, crime prevention, the military, commerce, culture, education, entertainment, and Web image classification and searching.

Content-based image retrieval has been widely studied. Space limitations do not allow us to present a broad survey. Instead we try to emphasize some of the work that is most related to what we propose. The references below are to be taken as examples of related work, not as the complete list of work in the cited area.

In the commercial domain, IBM QBIC [8, 25] is one of the earliest developed systems. Recently, additional systems have been developed at IBM T.J. Watson [34], VIRAGE [10], NEC C&C Research Labs [23], Bell Laboratory [24], Interpix (Yahoo), Excalibur, and Scour.net.

In academia, MIT Photobook [26, 27] is one of the earliest. Berkeley Blobworld [5], Columbia VisualSEEK and WebSEEK [33], CMU Informedia [35], University of Illinois MARS [22], University of California at Santa Barbara NeTra [21], the system developed by University of California at San Diego [14], Stanford WBIIS [36], and Stanford SIMPLIcity [38, 40]) are some of the recent systems.

Many indexing and retrieval methods have been used in these image retrieval systems. Some systems use keywords and full-text descriptions to index images. Others used features such as color histogram, color layout, local texture, wavelet coefficients, and shape to index images. In this paper, we focus on region-based retrieval of images.

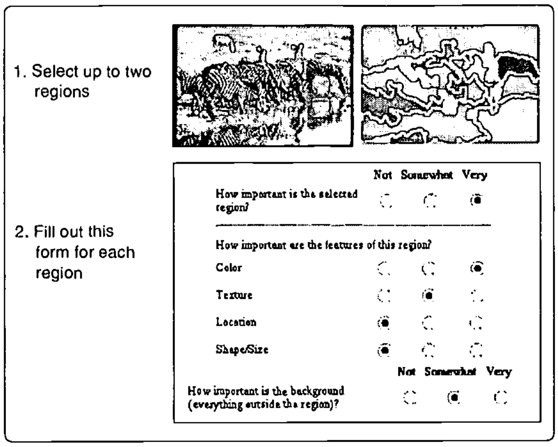
*An on-line demonstration is provided at URL: `http://wang.ist.psu.edu`

[†]Also of Department of Computer Science and Engineering and e-Business Research Center. Research started when the author was with the Departments of Biomedical Informatics and Computer Science at Stanford University.

[‡]Also of Department of Electrical Engineering, . Now with Cisco Systems, Inc.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

1.2 Region-based retrieval


1. Select up to two regions

2. Fill out this form for each region

	Not	Somewhat	Very
How important is the selected region?	○	○	●
How important are the features of this region?			
Color	○	○	●
Texture	○	●	○
Location	●	○	○
Shape/Size	●	○	○

	Not	Somewhat	Very
How important is the background (everything outside the region)?	○	●	○

Figure 1: Query procedure of the Blobworld system developed at the University of California, Berkeley.

Before the introduction of region-based systems, content-based image retrieval systems used color histogram and color layout to index the content of images. Region-based approach has recently become a popular research trend. Region-based retrieval systems attempt to overcome the deficiencies of color histogram and color layout search by representing images at the object-level. A region-based retrieval system applies image segmentation to decompose an image into regions, which correspond to objects if the decomposition is ideal. The object-level representation is intended to be close to the perception of the human visual system (HVS).

Many earlier region-based retrieval systems match images based on individual regions. Such systems include the Netra system [21] and the Blobworld system [5]. Figures 1 and 2 show the querying interfaces of Blobworld and Netra. Querying based on a limited number of regions is allowed. The query is performed by merging single-region query results. The motivation is to shift part of the comparison task to the users. To query an image, a user is provided with the segmented regions of the image, and is required to select the regions to be matched and also attributes, e.g., color and texture, of the regions to be used for evaluating similarity. Such querying systems provide more control for the users. However, the query formulation process can be very time consuming.

1.3 Integrated region-based retrieval

Researchers are developing similarity measures that combine information from all of the regions. One effort in this direction is the querying system developed by Smith and Li [34]. Their system decomposes an image into regions with characterizations pre-defined in a finite pattern library. With every pattern labeled by a symbol, images are then represented by region strings. Region strings are converted to composite region template (CRT) descriptor matrices that provide the relative ordering of symbols. Similarity between images is measured by the closeness between the CRT descriptor matrices. This measure is sensitive to object shifting since a CRT matrix is determined solely by the ordering of symbols. Robustness to scaling and rotation is also not considered by the measure. Because the definition of the CRT descriptor matrix relies on the pattern library, the system performance depends critically on the library. The performance degrades if region types in an image are not represented by patterns in the library. The system in [34] uses a CRT library with patterns described only by color. In particular, the patterns are obtained by quantizing color space. If texture and shape features are used to distinguish patterns, the number of patterns in the library will increase dramatically, roughly exponentially in the number of features if patterns are obtained by uniformly quantizing features.

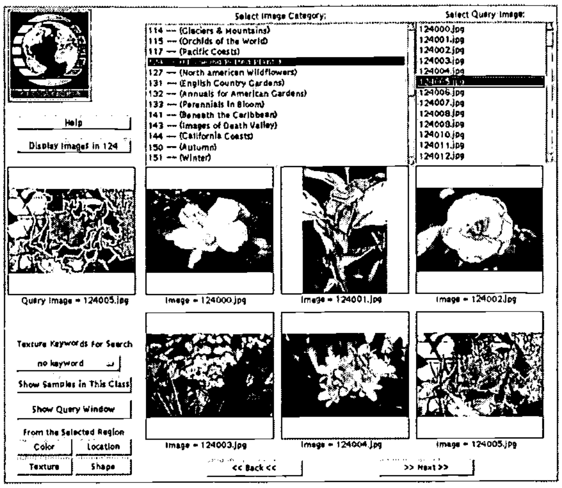


Figure 2: Query interface of the NeTra system developed at the University of California, Santa Barbara.

Li et al. of Stanford University recently developed SIMPLIcity (Semantics-sensitive Integrated Matching for Picture LIbraries) [37]. SIMPLIcity uses semantics type classification and an integrated region matching (IRM) scheme to provide efficient and robust region-based image matching [18]. The IRM measure is a similarity measure of images based on region representations. It incorporates the properties of all the segmented regions so that information about an image can be fully used. With IRM, region-based image-to-image matching can be performed. The overall similarity approach reduces the adverse effect of inaccurate segmentation, helps to clarify the semantics of a particular region, and enables a *simple* querying interface for region-based image retrieval systems. Experiments have shown that IRM is comparatively more effective and more robust than many existing retrieval methods. Like other region-based systems, the SIMPLIcity system is a linear matching system. To perform a query, the system compares the query image with all images in the same semantic class.

1.4 Statistical clustering

There are many efforts made to statistically cluster the high dimensional feature space before the actual searching using various tree structures such as K-D-B-tree [28], quadtree [9] R-tree [11], R^+-tree [31], R^*-tree [1], X-tree [3], SR-tree [15], M-tree [6], TV-tree [19], and hB-tree [20]. As mentioned in [4, 2, 1, 15, 41], the speed and accuracy of these algorithms degrade in very high dimensional spaces.

This is referred to as the *curse of dimensionality.* Besides, many of the clustering and indexing algorithms are designed for general purpose feature spaces such as Euclidean space. We developed our own algorithm for clustering and indexing image databases because we wanted the system to be suitable to our IRM region matching scheme.

1.5 Overview

In this paper, we present an enhancement to the SIMPLIcity system for handling image libraries with million of images. The targeted applications include Web image retrieval and biomedical image retrieval. Region features of images in the same semantic class are clustered automatically using a statistical clustering method. Features in the same cluster are stored in the same file for efficient access during the matching process. IRM (Integrated Region Matching) is used in the query matching process. Tested on large-scale image databases, the system has demonstrated high accuracy, robustness, and scalability.

The remainder of the paper is organized as follows. In Section 2, the similarity matching process based on segmented regions is defined. In Section 3, we describe the experiments we performed and provide results. We discuss limitations of the system in Section 4. We conclude in Section 5.

2. THE SIMILARITY MEASURE

In this section, we describe the similarity matching process we developed. We briefly describe the segmentation process and related notations in Section 2.1. The feature space analysis process is described in Section 2.2. In Section 2.3, we give details of the matching scheme.

2.1 Region segmentation

Semantically-precise image segmentation is extremely difficult and is still an open problem in computer vision [32, 39]. We attempt to develop a robust matching metric that can reduce the adverse effect of inaccurate segmentation. The segmentation process in our system is very efficient because it is essentially a wavelet-based fast statistical clustering process on blocks of pixels.

To segment an image, we partitions the image into blocks with $t \times t$ pixels and extracts a feature vector for each block. The k-means algorithm is used to cluster the feature vectors into several classes with every class corresponding to one region in the segmented image. We dynamically determine k by starting with $k = 2$ and refine if necessary to $k = 4$, etc. k is dynamically determined based on the complexity of the image. We do not require the clusters to be locationally contiguous because we rely on a robust matching process. The details of the segmentation process is described in [18].

Six features are used for segmentation. Three of them are the average color components in a $t \times t$ block. The other three represent energy in high frequency bands of wavelet transforms [7], that is, the square root of the second order moment of wavelet coefficients in high frequency bands. We use the well-known LUV color space, where L encodes luminance, and U and V encode color information (chrominance). The LUV color space has good perception correlation properties. We chose the block size t to be 4 to compromise between the texture detail and the computation time.

Let N denote the total number of images in the image database. For the i-th image, denoted as R_i, in the database, we obtain a set of n_i feature vectors after the region segmentation process. Each of these n_i d-dimensional feature vectors represents the dominant visual features (including color and texture) of a region, the shape of that region, the rough location in the image, and some statistics of the features obtained in that region.

2.2 Feature space analysis

The new integrated region matching scheme depends on the entire picture library. We must first process and analyze the characteristics of the d-dimensional feature space.

Suppose feature vectors in the d-dimensional feature space are $\{x_i : i = 1, ..., L\}$, where L is the total number of regions in the picture library. Then $L = \sum_{i=1}^{N} n_i$.

The goal of the feature clustering algorithm is to partition the features into k groups with centroids $\hat{x}_1, \hat{x}_2, ..., \hat{x}_k$ such that

$$D(k) = \sum_{i=1}^{L} \min_{1 \leq j \leq k} (x_i - \hat{x}_j)^2 \qquad (1)$$

is minimized. That is, the average distance between a feature vector and the group with the nearest centroid to it is minimized. Two necessary conditions for the k groups are:

1. Each feature vector is partitioned into the cluster with the nearest centroid to it.
2. The centroid of a cluster is the vector minimizing the average distance from it to any feature vector in the cluster. In the special case of the Euclidean distance, the centroid should be the mean vector of all the feature vectors in the cluster.

These requirements of our feature grouping process are the same requirements as those of the Lloyd algorithm [12] to find k cluster means with the following steps:

1. Initialization: choose the initial k cluster centroids.
2. Loop until the stopping criterion is met:
   (a) For each feature vector in the data set, assign it to a class such that the distance from this feature to the centroid of that cluster is minimized.
   (b) For each cluster, recalculate its centroid as the mean of all the feature vectors partitioned to it.

If the Euclidean distance is used, the k-means algorithm results in hyper-planes as cluster boundaries (Figure 3. That is, for the feature space $\mathbb{R}^d$, the cluster boundaries are hyperplanes in the $d-1$ dimensional space $\mathbb{R}^{d-1}$.

Both the initialization process and the stopping criterion are critical in the process. We initialize the algorithm adaptively by choosing the number of clusters k by gradually increasing k and stop when a criterion is met. We start with $k = 2$. The k-means algorithm terminates when no more feature vectors are changing classes. It can be proved that the k-means algorithm is guaranteed to terminate, based on the fact that both steps of k-means (i.e., assigning vectors to nearest centroids and computing cluster centroids) reduce the average class variance. In practice, running to completion may require a large number of iterations. The cost for each iteration is $O(kn)$, for the data size n. Our stopping criterion is to stop after the average class variance is smaller than a threshold or after the reduction of the class variance is smaller than a threshold.

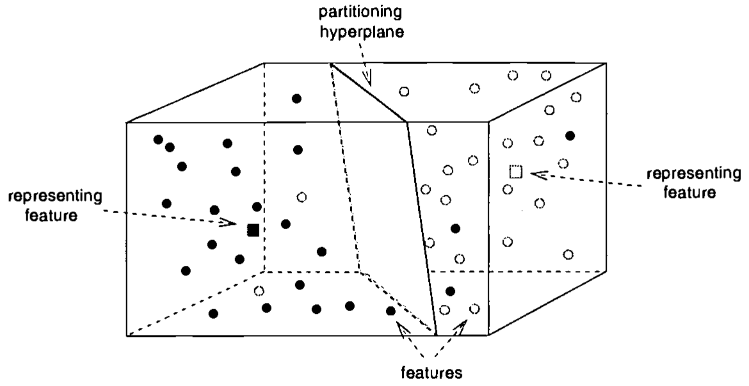


Figure 3: The k-means algorithm partitions the feature space using hyper-planes.

2.3 Image matching

To retrieve similar images for a query image, we first locate the clusters of the feature space to which the regions of the query image belong. Let's assume that the centroids of the set of k clusters are $\{c_1, c_2, ..., c_k\}$. We assume the query image is represented by region sets $R_1 = \{r_1, r_2, ..., r_m\}$, where r_i is the descriptor of region i. For each region feature r_i, we find j such that

$$d(r_i, c_j) = \min_{1 \leq l \leq k} d(r_i, c_l)$$

where $d(r_1, r_2)$ is the region-to-region distance defined for the system. This distance can be a non-Euclidean distance. We create a list of clusters, denoted as $\{c_{r_1}, c_{r_2}, ..., c_{r_k}\}$. The matching algorithm will further investigate only these 'suspect' clusters to answer the query.

With the list of 'suspect' clusters, we create a list of 'suspect' images. An image in the database is a 'suspect' image to the query if the image contains at least one region feature in these 'suspect' clusters. This step can be accomplished by merging the cluster image IDs non-repeatedly.

To define the similarity measure between two sets of regions, we assume that the image R_1 and image R_2 are represented by region sets $R_1 = \{r_1, r_2, ..., r_m\}$ and $R_2 = \{r'_1, r'_2, ..., r'_n\}$, where r_i or r'_i is the descriptor of region i. Denote the distance between region r_i and r'_j as $d(r_i, r'_j)$, which is written as $d_{i,j}$ in short. To compute the similarity measure between region sets R_1 and R_2, $d(R_1, R_2)$, we first compute all pair-wise region-to-region distances in the two images. Our matching scheme aims at building correspondence between regions that is consistent with our perception. To increase robustness against segmentation errors, we allow a region to be matched to several regions in another image. A matching between r_i and r'_j is assigned with a significance credit $s_{i,j}$, $s_{i,j} \geq 0$. The significance credit indicates the importance of the matching for determining similarity between images. The matrix $S = \{s_{i,j}\}$, $1 \leq i \leq n$, $1 \leq j \leq m$, is referred to as the significance matrix.

The distance between the two region sets is the summation of all the weighted matching strength, i.e.,

$$d_{IRM}(R_1, R_2) = \sum_{i,j} s_{i,j} d_{i,j} \ .$$

This distance is the integrated region matching (IRM) distance defined by Li et al. in [18].

To choose the significance matrix S, a natural issue to raise is what constraints should be put on $s_{i,j}$ so that the admissible matching yields good similarity measure. In other words, what properties do we expect an admissible matching to possess? The first property we want to enforce is the fulfillment of significance. Assume that the significance of r_i in Image 1 is p_i, and r'_j in Image 2 is p'_j, we require that

$$\begin{aligned} \sum_{j=1}^{n} s_{i,j} &= p_i, \ i = 1, ..., m \\ \sum_{i=1}^{m} s_{i,j} &= p'_j, \ j = 1, ..., n \ . \end{aligned}$$

A greedy scheme is developed to speed up the determination of the matrix $S = \{s_{i,j}\}$. Details of the algorithm can be found in [18].

2.4 The RF*IPF weighting

For applications such as biomedical image retrieval, local feature is critically important in distinguishing the semantics between two images. In this section, we present the Region Frequency and Inverse Picture Frequency (RF*IPF) weighting, a relatively simple weighting measure developed to further enhance the discriminating efficiency of IRM based on the characteristics of the entire picture library. This weighting can be used to emphasize uncommon features.

The definition of RF*IPF is in some way close to the definition of the Term Frequency and Inverse Document Frequency (TF*IDF) weighting [30], a highly effective techniques in document retrieval. The combination of RF*IPF and IRM is more effective than the IRM itself in a variety of image retrieval applications. Additionally, this weighting measure provides a better unification of content-based image retrieval and text-based image retrieval.

The RF*IPF weighting consists of two parameters: the Region Frequency (RF) and the Inverse Picture Frequency (IPF).

For each region feature vector x_i of the image R_j, we find the closest group centroid from the list of k centroids computed in the feature analysis step. That is, we find c_0 such that

$$\| x_i - \hat{x}_{c_0} \| = \min_{1 \leq c \leq k} \| x_i - \hat{x}_c \| \ . \qquad (2)$$

Let's denote N_{c_0} as the number of pictures in the database with at least one region feature closest to the centroid $\hat{x}_{c_0}$ of the image group c_0. Then we define

$$IPF_i = log\left(\frac{N}{N_{c_0}}\right) + 1 \qquad (3)$$

where IPF_i is the Inverse Picture Frequency of the feature x_i.

Now let's denote M_j as the total number of pixels in the image R_j. For images in a size-normalized picture library, M_j are constants for all j. Denote $P_{i,j}$ as the area percentage of the region i in the image R_j. Then, we define

$$RF_{i,j} = log(P_{i,j}M_j) + 1 \qquad (4)$$

as the Region Frequency of the i-th region in picture j. Then RF measures how frequently a region feature occurs in a picture.

We can now assign a weight for each region feature in each picture. The RF*IPF weight for the i-th region in the j-th image R_j is defined as

$$W_{i,j} = RF_{i,j} * IPF_i \ . \qquad (5)$$

Clearly, the definition is close to that of the TF*IDF (Term Frequency times Inverse Document Frequency) weighting in text retrieval.

After computing the RF*IPF weights for all the L regions in all the N images in the image database, we store these weights for the image matching process.

We now combine the IRM distance with the RF*IPF weighting in the process of choosing the significance matrix S. A natural issue to raise is what constraints should be put on $s_{i,j}$ so that the admissible matching yields good similarity measure. In other words, what properties do we expect an admissible matching to possess? The first property we want to enforce is the fulfillment of significance. We computed the significance W_{i,R_1} of r_i in image R_1 and r'_j in image R_2 is W_{j,R_2}, we require that

$$\begin{aligned}\sum_{j=1}^{n} s_{i,j} &= p_i = \frac{W_{i,R_1}}{\sum_{l=1}^{m} W_{l,R_1}}, \ i = 1, ..., m \\ \sum_{i=1}^{m} s_{i,j} &= q_j = \frac{W_{j,R_2}}{\sum_{l=1}^{n} W_{l,R_2}}, \ j = 1, ..., n \ .\end{aligned}$$

3. EXPERIMENTS

This algorithm has been implemented and compared with the first version of our experimental SIMPLIcity image retrieval system. We tested the system on a general-purpose image database (from COREL) including about 200,000 pictures, which are stored in JPEG format with size 384 × 256 or 256 × 384. To conduct a fair comparison, we use only picture features in the retrieval process.

3.1 Speed

On a Pentium III 800MHz PC using the Linux operating system, it requires approximately 60 hours to compute the feature vectors for the 200,000 color images of size 384 × 256 in our general-purpose image database. On average, one second is needed to segment an image and to compute the features of all regions. Fast indexing has provided us with the capability of handling outside queries and sketch queries in real-time.

The feature clustering process is performed only once for each database. The Lloyd algorithm takes about 30 minutes CPU time and results in clusters with an average of 1100 images. Our image segmentation process generates an average of 4.6 regions per image. That is, on average a 'suspect' list for a query image contains at most 1100 × 4.6 = 5060 images.

Category	IRM	fast IRM	EMD2	EMD 1
1. Africa	0.475	0.472	0.288	0.132
2. Beach	0.325	0.323	0.286	0.134
3. Buildings	0.330	0.307	0.233	0.160
4. Buses	0.363	0.389	0.267	0.108
5. Dinosaurs	0.981	0.635	0.914	0.143
6. Elephants	0.400	0.390	0.384	0.169
7. Flowers	0.402	0.447	0.416	0.113
8. Horses	0.719	0.669	0.386	0.096
9. Mountains	0.342	0.335	0.218	0.198
10. Food	0.340	0.340	0.207	0.114

Table 1: The average performance for each image category evaluated by average precision (p).

The matching speed is fast. When the query image is in the database, it takes about 0.15 seconds of CPU time on average to sort all the images in the 200,000-image database using our similarity measure. This is a significant speed-up over the original system which runs at 1.5 second per query. If the query is not in the database, one extra second of CPU time is spent to process the query.

Figures 4 and 5 show the results of sample queries. Due to the limitation of space, we show only two rows of images with the top 11 matches to each query. In the next section, we provide numerical evaluation results by systematically comparing several systems.

Because of the fast indexing and retrieval speed, we allow the user to submit any images on the Internet as a query image to the system by entering the URL of an image (Figure 6). Our system is capable of handling any image format from anywhere on the Internet and reachable by our server via the HTTP protocol. The image is downloaded and processed by our system on-the-fly. The high efficiency of our image segmentation and matching algorithms made this feature possible[1]. To our knowledge, this feature of our system is unique in the sense that no other commercial or academic systems allow such queries.

3.2 Accuracy on image categorization

We conducted extensive evaluation of the system. One experiment was based on a subset of the COREL database, formed by 10 image categories, each containing 100 pictures. Within this database, it is known whether any two images are of the same category. In particular, a retrieved image is considered a match if and only if it is in the same category as the query. This assumption is reasonable since the 10 categories were chosen so that each depicts a distinct semantic topic. Every image in the sub-database was tested as a query, and the retrieval ranks of all the rest images were recorded.

For each query, we computed the precision within the first 100 retrieved images. The recall within the first 100 retrieved images was not computed because it is proportional to the precision in this special case. The total number of se-

[1] It takes some other region-based CBIR system [5] approximately 8 minutes CPU time to segment an image.

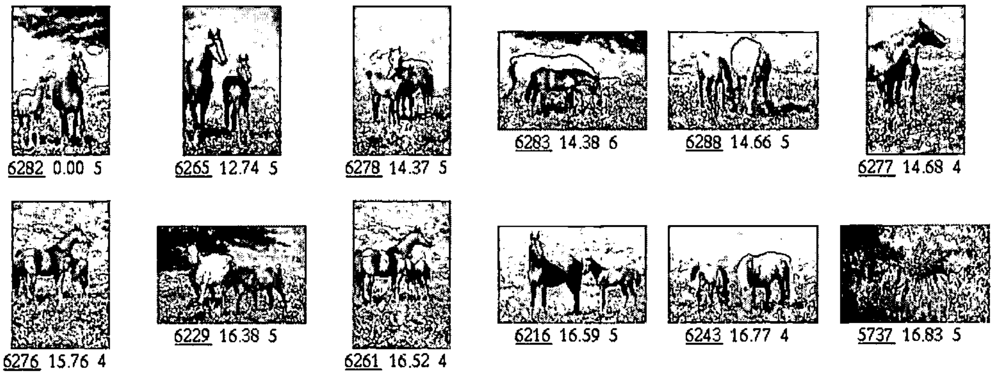
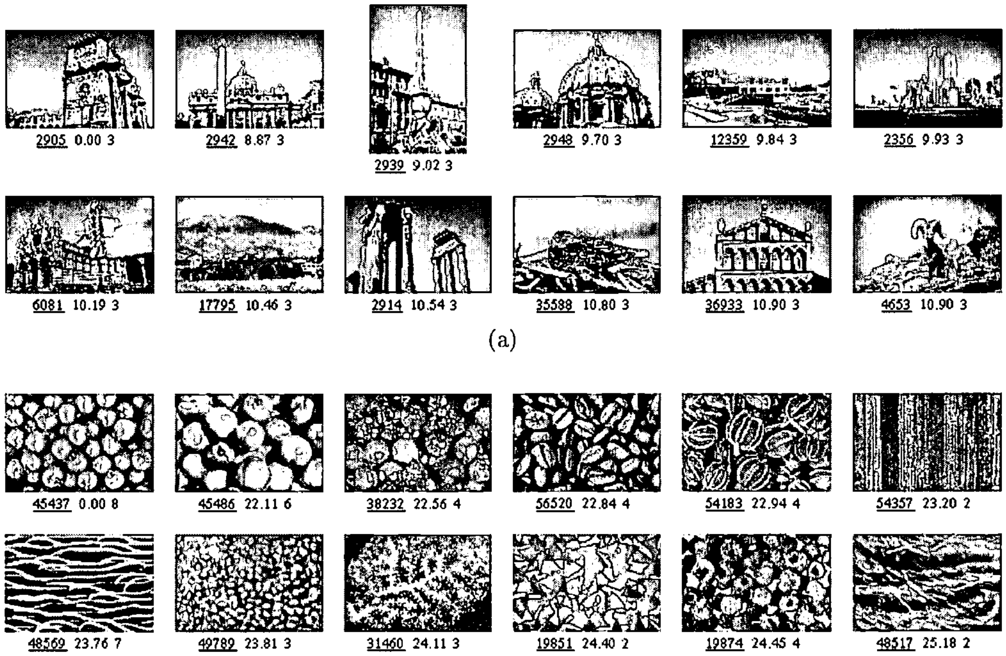


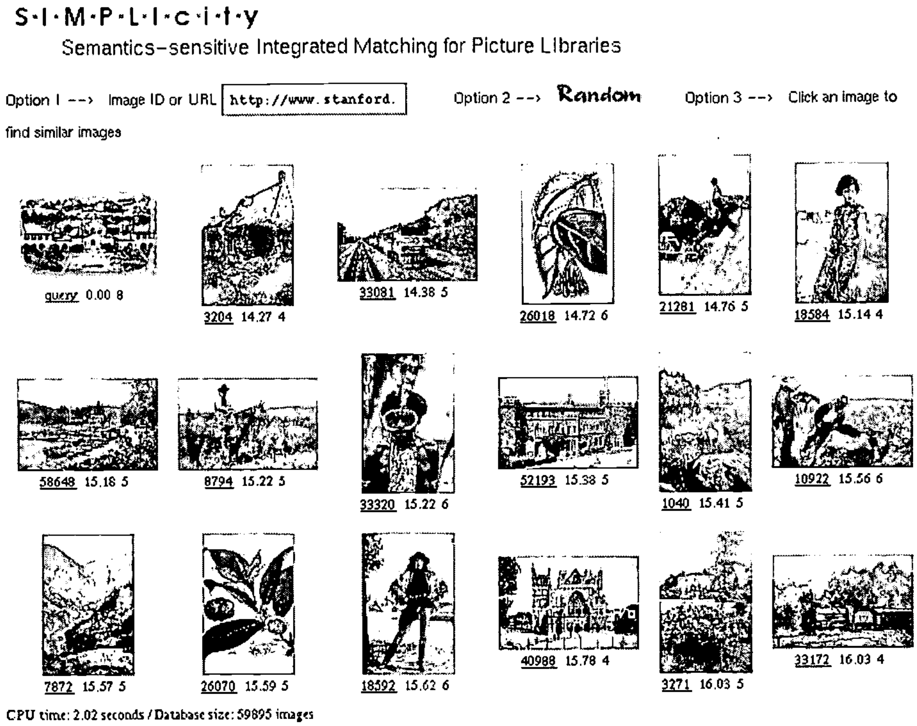
Figure 4: Best 11 matches of a sample query. The database contains 200,000 images from the COREL image library. The upper left corner is the query image. The second image in the first row is the best match.



(a)

(b)

Figure 5: Two other query examples.



S·I·M·P·L·I·c·i·t·y

Semantics-sensitive Integrated Matching for Picture Libraries

Option 1 --> Image ID or URL http://www.stanford. Option 2 --> Random Option 3 --> Click an image to find similar images

query 0.00 8

3204 14.27 4

33081 14.38 5

26018 14.72 6

21281 14.76 5

18584 15.14 4

58648 15.18 5

8794 15.22 5

33320 15.22 6

52193 15.38 5

1040 15.41 5

10922 15.56 6

7872 15.57 5

26070 15.59 5

18592 15.62 6

40988 15.78 4

3271 16.03 5

33172 16.03 4

CPU time: 2.02 seconds / Database size: 59895 images

Figure 6: The external query interface. The best 17 matches are presented for a query image selected by the user from the Stanford top-level Web page. The user enters the URL of the query image (shown in the upper-left corner, http://www.stanford.edu/home/pics/h-quad.jpg) to form a query.

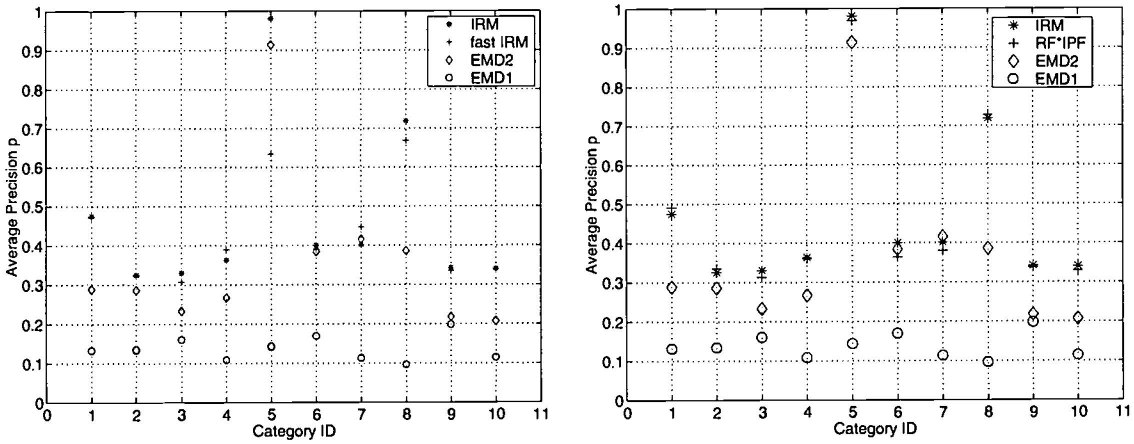


Figure 7: Comparing with color histogram methods on average precision p. Color Histogram 1 gives an average of 13.1 filled color bins per image, while Color Histogram 2 gives an average of 42.6 filled color bins per image. SIMPLIcity partitions an image into an average of only 4.3 regions.

mantically related images for each query is fixed to be 100. The average performance for each image category in terms of the average precision is listed in Table 1, where p denotes precision. For a system that ranks images randomly, the average p is about 0.1.

We carried out similar evaluation tests for color histogram match. We used LUV color space and a matching metric similar to the EMD described in [29] to extract color histogram features and match in the categorized image database. Two different color bin sizes, with an average of 13.1 and 42.6 filled color bins per image, were evaluated. We call the one with less filled color bins the Color Histogram 1 system and the other the Color Histogram 2 system. Figure 7 shows the performance as compared with the Lloyd-based SIMPLIcity system. Clearly, both of the two color histogram-based matching systems perform much worse than the Lloyd-based system in almost all image categories. The performance of the Color Histogram 2 system is better than that of the Color Histogram 1 system due to more detailed color separation obtained with more filled bins. However, the Color Histogram 2 system is so slow that it is impossible to obtain matches on larger databases. The original SIMPLIcity runs at about twice the speed of the faster Color Histogram 1 system and gives much better searching accuracy than the slower Color Histogram 2 system.

The overall performance of the Lloyd-based system is close to that of the original system which uses IRM and area percentages of the segmented regions as significant constraints. Both the regular IRM and the fast IRM algorithms are much more accurate than the EMD-based color histogram. Experiments on a database of 70,000 pathology slides demonstrated similar comparison results.

3.3 Robustness

Similar to the original SIMPLIcity system [38], the current system is exceptionally robust to image alterations such as intensity variation, sharpness variation, intentional color distortions, intentional shape distortions, cropping, shifting, and rotation.

The system is fairly robust to image alterations such as intensity variation, sharpness variation, intentional color distortions, other intentional distortions, cropping, shifting, and rotation. On average, the system is robust to approximately 10% brightening, 8% darkening, blurring with a 15 × 15 Gaussian filter, 70% sharpening, 20% more saturation, 10% less saturation, random spread by 30 pixels, and pixelization by 25 pixels. These features are important to biomedical image databases because usually visual features of the query image are not identical to the visual features of those semantically-relevant images in the database because of problems such as occlusion, difference in intensity, and difference in focus.

4. DISCUSSIONS

The system has several limitations. (1) Like other CBIR systems, SIMPLIcity assumes that images with similar semantics share some similar features. This assumption may not always hold. (2) The shape matching process is not ideal. When an object is segmented into many regions, the IRM distance should be computed after merging the matched regions. (3) The querying interfaces are not powerful enough to allow users to formulate their queries freely. For different user domains (e.g., biomedicine, Web image retrieval), the query interfaces should ideally provide different sets of functions.

In our current system, the set of features for a particular image category is determined empirically based on the perception of the developers. For example, shape-related features are not used for textured images. Automatic derivation of optimal features is a challenging and important issue in its own right. A major difficulty in feature selection is the lack of information about whether any two images in the database match with each other. The only reliable way to obtain this information is through manual assessment, which is formidable for a database of even moderate size. Furthermore, human evaluation is hard to be kept consistent from person to person. To explore feature selection, primitive studies can be carried with relatively small databases. A database can be formed from several distinctive groups of images, among which only images from the same group are considered matched. A search algorithm can be developed to select a subset of candidate features that provides optimal retrieval according to an objective performance measure. Although such studies are likely to be seriously biased, insights regarding which features are most useful for a certain image category may be obtained.

The main limitation of our current evaluation results is that they are based mainly on precision or variations of precision. In practice, a system with a high overall precision may have a low overall recall. Precision and recall often trade off against each other. It is extremely time-consuming to manually create detailed descriptions for all the images in our database in order to obtain numerical comparisons on recall. The COREL database provides us rough semantic labels on the images. Typically, an image is associated with one keyword about the main subject of the image. For example, a group of images may be labeled as "flower" and another group of images may be labeled as "Kyoto, Japan". If we use the descriptions such as "flower" and "Kyoto, Japan" as definitions of relevance to evaluate CBIR systems, it is unlikely that we can obtained a consistent performance evaluation. A system may perform very well on one query (such as the flower query), but very poorly on another (such as the Kyoto query). Until this limitation is thoroughly investigated, the evaluation results reported in the comparisons should be interpreted cautiously.

5. CONCLUSIONS AND FUTURE WORK

We have developed a scalable integrated region-based image retrieval system. The system uses the IRM measure and the Lloyd algorithm. The algorithm has been implemented as part of the the IRM metric in our experimental SIMPLIcity image retrieval system. Tested on a database of about 200,000 general-purpose images, the technique has demonstrated high efficiency and robustness. The main difference between this system and the previous SIMPLIcity system is the statistical clustering process which significantly reduces the computational complexity of the IRM measure.

The clustering efficiency can be improved by using a better statistical clustering algorithm. Better statistical modeling and matching scheme is likely to improve the matching accuracy of the system. We are also planning to apply the methods to special image databases (e.g., biomedical), and very large multimedia document databases (e.g., WWW, video).

6. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation's Digital Library II initiative under Grant No. IIS-9817511 and an endowment from the PNC Foundation. The authors wish to thank the help of Jia Li, Xiaoming Huo, Gio Wiederhold, Oscar Firschein, and anonymous reviewers.

7. REFERENCES

[1] N. Beckmann, H.-P. Kriegel, R. Schneider, B. Seeger, "The R*-tree: An efficient and robust access method for points and rectangles," *Proc. ACM SIGMOD*, pp. 322-331, Atlantic City, NJ, 23-25 May 1990.

[2] J. Bentley, J. Friedman, "Data structures for range searching," *ACM Computing Surveys*, vol. 11, no. 4, pp. 397-409, December 1979.

[3] S. Berchtold, D. Keim, H.-P. Kriegel, "The X-tree: An index structure for high-dimensional data," *Proc. Int. Conf. on Very Large Databases*, pp. 28-39, 1996.

[4] S. Berchtold, C. Bohm, B. Braunmuller, D. Keim, H.- P. Kriegel, "Fast parallel similarity search in multimedia databases," *Proc. ACM SIGMOD*, pp. 1-12, Tucson, AZ, 1997.

[5] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, J. Malik, "Blobworld: a system for region-based image indexing and retrieval," *Proc. Int. Conf. on Visual Information Systems*, D. P. Huijsmans, A. W.M. Smeulders (eds.), Springer, Amsterdam, The Netherlands, June 2-4, 1999.

[6] P. Ciaccia, M. Patella, P. Zezula, "M-tree: An efficient access method for similarity search in metric spaces," *Proc. Int. Conf. on Very Large Databases*, Athens, Greece, 1997.

[7] I. Daubechies, *Ten Lectures on Wavelets*, Capital City Press, 1992.

[8] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, W. Equitz, "Efficient and effective querying by image content," *Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies*, vol. 3, no. 3-4, pp. 231-62, July 1994.

[9] R. Finkel, J. Bentley, "Quad-trees: A data structure retrieval on composite keys," *ACTA Informatica*, vol. 4, no. 1, pp. 1-9, 1974.

[10] A. Gupta, R. Jain, "Visual information retrieval," *Communications of the ACM*, vol. 40, no. 5, pp. 70-79, May 1997.

[11] A. Guttman, "R-trees: A dynamic index structure for spatial searching," *Proc. ACM SIGMOD*, pp. 47-57, Boston, MA, June 1984.

[12] J. A. Hartigan, M. A. Wong, "Algorithm AS136: a k-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100-108, 1979.

[13] "Web surpasses one billion documents," *Inktomi Corporation Press Release*, January 18, 2000.

[14] R. Jain, S. N. J. Murthy, P. L.-J. Chen, S. Chatterjee "Similarity measures for image databases", *Proc. SPIE*, vol. 2420, pp. 58-65, San Jose, CA, Feb. 9-10, 1995.

[15] N. Katayama, S. Satoh, "The SR-tree: An index structure for high-dimensional nearest neighbor queries," *Proc. ACM SIGMOD* pp. 369-380, Tucson, AZ, 1997.

[16] S. Lawrence, C.L. Giles, "Searching the World Wide Web," *Science*, vol. 280, pp. 98, 1998.

[17] S. Lawrence, C.L. Giles, "Accessibility of information on the Web," *Nature*, vol. 400, pp. 107-109, 1999.

[18] J. Li, J. Z. Wang, G. Wiederhold, "IRM: Integrated region matching for image retrieval," *Proc. ACM Multimedia Conference*, pp. 147-156, Los Angeles, ACM, October, 2000.

[19] K.-I. Lin, H. Jagadish, C. Faloutsos, "The TV-tree: An index structure for high-dimensional data," *The VLDB Journal*, vol. 3, no. 4, pp. 517-549, October 1994.

[20] D. Lomet, "The hB-tree: A multiattribute indexing method with good guaranteed performance," *ACM Transactions on Database Systems*, vol. 15, no. 4, pp. 625-658, December 1990.

[21] W. Y. Ma, B. Manjunath, "NaTra: A toolbox for navigating large image databases," *Proc. IEEE Int. Conf. Image Processing*, pp. 568-71, 1997.

[22] S. Mehrotra, Y. Rui, M. Ortega-Binderberger, T.S. Huang, "Supporting content-based queries over images in MARS," *Proc. IEEE International Conference on Multimedia Computing and Systems*, pp. 632-3, Ottawa, Ont., Canada 3-6 June 1997.

[23] S. Mukherjea, K. Hirata, Y. Hara, "AMORE: a World Wide Web image retrieval engine," *World Wide Web*, vol. 2, no. 3, pp. 115-32, Baltzer, 1999.

[24] A. Natsev, R. Rastogi, K. Shim, "WALRUS: A similarity retrieval algorithm for image databases," *Proc. ACM SIGMOD*, Philadelphia, PA, 1999.

[25] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, G. Taubin, "The QBIC project: querying images by content using color, texture, and shape," *Proc. SPIE*, vol. 1908, pp. 173-87, San Jose, February, 1993.

[26] A. Pentland, R. W. Picard, S. Sclaroff, "Photobook: tools for content-based manipulation of image databases," *Proc. SPIE*, vol. 2185, pp. 34-47, San Jose, February 7-8, 1994.

[27] R. W. Picard, T. Kabir, "Finding similar patterns in large image databases," *Proc. IEEE ICASSP*, Minneapolis, vol. V, pp. 161-64, 1993.

[28] J. Robinson, "The k-d-b-tree: A search structure for large multidimensional dynamic indexes," *Proc. ACM SIGMOD* pp. 10-18, 1981.

[29] Y. Rubner, L. J. Guibas, C. Tomasi, "The earth mover's distance, Shimulti-dimensional scaling, and color-based image retrieval," *Proc. ARPA Image Understanding Workshop*, pp. 661-668, New Orleans, LA, May 1997.

[30] G. Salton, M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, NY, 1983.

[31] T. Sellis, N. Roussopoulos, C. Faloustos. "The R+-tree: A dynamic index for multi-dimensional objects," *Proc. Int. Conf. on Very Large Databases*, pp. 507-518, Brighton, England, 1987.

[32] J. Shi, J. Malik, "Normalized cuts and image segmentation," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 731-7, San Juan, Puerto Rico, June, 1997.

[33] J. R. Smith, S.-F. Chang, "An image and video search engine for the World-Wide Web," *Proc. SPIE*, vol. 3022, pp. 84-95, 1997.

[34] J. R. Smith, C. S. Li, "Image classification and querying using composite region templates," *Journal of Computer Vision and Image Understanding*, 2000.

[35] S. Stevens, M. Christel, H. Wactlar, "Informedia: improving access to digital video," *Interactions*, vol. 1, no. 4, pp. 67-71, 1994.

[36] J. Z. Wang, G. Wiederhold, O. Firschein, X. W. Sha, "Content-based image indexing and searching using Daubechies' wavelets," *International Journal of Digital Libraries*, vol. 1, no. 4, pp. 311-328, 1998.

[37] J. Z. Wang, J. Li, D. , G. Wiederhold, "Semantics-sensitive retrieval for digital picture libraries," *D-LIB Magazine*, vol. 5, no. 11, DOI:10.10 45/november99-wang, November, 1999. http://www.dlib.org

[38] J. Z. Wang, J. Li, G. Wiederhold, "SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, 2001. to appear.

[39] J. Z. Wang, J. Li, R. M. Gray, G. Wiederhold, "Unsupervised multiresolution segmentation for images with low depth of field," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, pp. 85-90, 2001.

[40] J. Z. Wang, *Integrated Region-Based Image Retrieval*, Kluwer Academic Publishers, 190 pp., 2001.

[41] R. Weber, Hans-J. Schek, Stephen Blott, "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces," *Proc. Int. Conf. on Very Large Databases*, New York, 1998.

Cumulating and Sharing End Users Knowledge to Improve Video Indexing in a Video Digital Library

Marc Nanard
LIRMM
161 rue Ada, 34392 Montpellier France
Phone: (33) 467 41 85 17, Fax: (33) 467 41 85 00
mnanard@lirmm.fr

Jocelyne Nanard
LIRMM
161 rue Ada, 34392 Montpellier France
Phone: (33) 467 41 85 17, Fax: (33) 467 41 85 00
jnanard@lirmm.fr

ABSTRACT

In this paper, we focus on a user driven approach to improve video indexing. It consists in cumulating the large amount of small, individual efforts done by the users who access information, and to provide a community management mechanism to let users share the elicited knowledge. This technique is currently being developed in the "OPALES" environment and tuned up at the "Institut National de l'Audiovisuel" (INA), a National Video Library in Paris, to increase the value of its patrimonial video archive collections. It relies on a portal providing private workspaces to end users, so that a large part of their work can be shared between them. The effort for interpreting documents is directly done by the expert users who work for their own job on the archives. OPALES provides an original notion of "point of view" to enable the elicitation and the sharing of knowledge between communities of users, without leading to messy structures. The overall result consists in linking exportable private metadata to archive documents and managing the sharing of the elicited knowledge between users communities.

Categories and Subject Descriptors

H.3.5[INFORMATION STORAGE AND RETRIEVAL]: Online Information Services - *Data bank sharing*

General Terms

Design

Keywords

Video annotation. Video indexing. Private workspaces. Users communities. Knowledge sharing.

1. INTRODUCTION

It is now well admitted that retrieval of relevant images or video segments among large collections requires taking advantage of semantically rich metadata associated to small information chunks. A lot of efficient techniques for automatically elaborating metadata from text documents are now well mastered. References on that topic can be found, for instance, in conferences on information retrieval [19].

At the opposite, automatically elaborating semantically relevant metadata from images and, moreover, from video is a far harder task [1] which currently is a challenge for further development of the information technologies and multimedia digital libraries. The cause is obvious: contrary to texts which, as a natural language representation, have the power and all of the features of a formal knowledge representation scheme, images only rely on an iconic representation scheme [18]. They rely only on suggestive, emotional communication modes. They do not usually embed any syntactic or semantic structures likely to be elicited by a machine for elaborating semantically rich metadata. As a consequence, and unfortunately, human interpretation of video still is the only one technique which enables precise semantic indexing at scene level.

Automatic image indexing techniques have huge difficulties in accessing the semantics of an image. The simplest image indexing techniques do not care at all for image semantics. They are based on signal processing. They focus only on physical and graphical properties of the image [3] such as the color histogram, the textures, image similitude, and so on, without any interpretation. A more elaborated approach takes advantage of image recognition. Such techniques currently remain limited to simple cases such as very typical faces recognition [5], [13], situation recognition (sitting/standing), familiar object recognition (cars, planes, tables). Nevertheless, very little semantics can be elicited from image analysis. A far more efficient approach consists in taking advantage of multimodality between image and sound tracks in movies or in TV news broadcast for cross fertilizing the document analysis. In the Informedia project [12], [17], the recognition of a subset of relevant words such as politicians or country names in the sound track of news may let attach, for instance, to a landscape image a metadata telling that the image concerns "Afghanistan", since this word has been recognized in the voice commentary. This technique also helps contextually solving ambiguities in image recognition. For instance, let us suppose the system recognizes the presence of a face but cannot identify it further. Famous names recognition in an associated commentary on the sound track may help the system improve recognition, solve ambiguities and let it suppose it is, for instance, Marilyn Monroe's face. This quite efficient technique for automatically indexing news is already available on the market place. Nevertheless, none of these automatic techniques can fully succeed in automatically indexing a large variety of archive documents. Either there is no or too few associated multimodal data, or the commentary is only very loosely related to the image, like this is unfortunately the case in many news report. Therefore only the mixing of several approaches can lead to a better indexing of images and of videos. In most cases, correct indexing of images and of video requires human interpretation of the situations.

Conference JCDL '00, May, 2000, Virginia.

In this paper, we focus on a typically different approach to improve video indexing. The approach does not intend at all to be a substitute to other. Rather it is a complementary strategy for drastically improving the overall efficiency of the end user's work in the trend of social navigation [7], [8]. It consists in cumulating the large amount of small, individual efforts done by the users who access information, and in providing a community management mechanism to let users share the elicited knowledge. This technique is currently being developed in the OPALES environment and tuned up at the Institut National de l'Audiovisuel in Paris (INA) to increase the value of its video archive collections. It relies on a portal providing private workspaces to end users, so that a large part of their work can be shared between them. The effort for interpreting documents is directly done by the expert users who work for their own job on the archives. OPALES provides with an original notion of "point of view" to enable the elicitation and the sharing of knowledge between communities of users, without leading to messy structures. The overall result consists in linking exportable private metadata to archive documents and managing the sharing of the elicited knowledge between user communities.

The paper first describes the context of the study and its design rationale. Then it focuses on a specific point of the project: the management of user elicited knowledge. The notion of "point of view" enables to reduce the problem complexity. It helps manage smaller knowledge clusters specific to user communities.

2. CONTEXT OF THE WORK

In any domain of industry, companies usually keep track of their own production, most often for technical or commercial reasons, but sometimes also as archives considered as a memory of patrimony. We name these kinds of archives "patrimonial archives". For instance, car producers build large museums to exhibit tracks of their creative activity. In any cases, these archives represent a very small part of their production. Contrary to goods manufacturers, information producers deal with such a huge amount of data that keeping all of it for a long time is a hard and costly choice. Whereas policies for archiving printed documents for the long term are now ruled at national level in many countries, video production is not yet concerned with such rules. Storage is often handled directly by producers, and thus storage strategies may be subject to opportunistic variations. As a consequence, a large part of TV production is discarded once it has been broadcast. In many cases, just the best part, or the reusable part is preserved. Even, in TV or radio companies where systematic archiving is often the rule, heavy storage cost, lack of room for storage, inconsistencies in the storage strategy or changes in the management sometimes lead to later discard archives which had been preserved for years. For instance, such a situation had already occurred, leading a few years ago a famous broadcasting company to discard a large part of its records of daily news of the fifties.

2.1 INA, multimedia archive provider

The Institut National de l'Audiovisuel (INA), created in Paris in the early seventies, is in charge of keeping records of national French TV broadcasts. A law voted in June 1992 defines the "*dépot legal* (official and mandatory storage)" which requires copies of any national radio or TV production to be deposited at INA as patrimonial archives. Storage does not concern simply the items themselves (e.g.: TV series as such) but also the context in which they have been broadcast. This enables rich sociologic studies, for instance studies of correlation between the focus of advertisements and the contents of the film they break. Similarly, the context associated to the audio and video contents provides historians with a far more precise record of our way of life than separate items would do. Furthermore, INA has inherited from the archives of the previous national broadcasting company "ORTF". Currently, INA deals with more than one and a half billion of hours of TV and radio and more than one billion of still pictures stored on more than fifteen miles of shelves. INA already has started to convert its data to digital format. 200 000 hours of TV and 300 000 hours of radio are now available, thus making it the repository of one of the largest collection of audio-video archives, like those of BBC and RAI.

INA's main function is to be an information provider for TV producers, and for any other professionals. INA is famous in France for its authentic and watermarked archive sources included in TV news. It also serves as a patrimonial archive library for researchers such as historians, sociologists, economists, politicians, and so on, who study historical facts. Since INA is just the archivist but is not the copyright owner of all of deposited archives, it often operates just as a partner between buyers and information owners.

Efficiently accessing such a huge amount of archives is an increasingly important challenge for INA. Like in any library, the video archives have been indexed once for all when they were stored. This initial indexing is obviously sufficient for most of professional use: everyday TV producers access the INA video-library to search and buy archive sequences. Of course, it is not possible, nor suitable to make changes in this primary indexing scheme to improve it.

One way to offer better service to users is to build a new separate indexing, based on more efficient and more precise techniques such as NCG [4], enabling video indexing at different levels of granularity and stratification of indexing [20]. Unfortunately, the cost for re-indexing the entire set of archive documents is far beyond the possibilities of the organization. So, the planned solution is to let it be done by the users themselves and to incite them to cumulate their individual efforts to improve the overall service.

2.2 The OPALES project

2.2.1 Overview

OPALES is an ongoing R&D project, initiated by the French ministry of Economy in 2000, scheduled to be operational in the fall 2001. It aims at developing a new service empowered by digital video and hypermedia technology, and intended to incrementally increase the value of the multimedia archives accessed through it. It consists of a distributed environment able to support the activity of virtual communities of experts working on the INA patrimonial video archives. OPALES is a private portal. It enables its users to directly work on archive documents in private workspaces, to share elicited knowledge about studied documents, and to collaborate anonymously as well as within explicit groups. The basic assumption is that the results of the work of expert groups can be made available to others, thus boosting their own work. The return of business generated by knowledge exchange between experts is also business for the archive provider itself.

2.2.2 Target users

Access to the OPALES portal is currently restricted to a group of researchers who participate to the R&D project. Beside INA, several institutions participate to its elaboration and evaluation: the "Cité des Sciences et de l'Industrie" in Paris, the MSH "Maison des Sciences de l'Homme", the CNDP "National Center for Distance Learning", and the BPS "Program and Service Bank" of the 5th TV Channel. They provide expert users as well as video data. The targeted users are typically knowledge workers. For the first steps of the project, researchers in human sciences and teachers have been chosen as representatives of future users of the system. They access documents and study them with the purpose of elaborating new knowledge, either for their own usage or for transmitting it to others.

2.2.3 Corpus

In order to make experiments easier and cheaper, the corpus currently used to bootstrap the project only contains copyright free documents. Handling copyright issues is of course one of the usual INA business. But this point is beyond the scope of the first stage of the project.

2.2.4 Task

The task supported by OPALES is called "active reading". Researchers usually practice active reading in libraries. They act as readers and writers at the same time. They annotate, extract, search, etc. Such a task consists of alternated reading and writing steps deeply intermingled, thus producing a gloss bound to the document. Although the term "active reading" had been coined for working on printed documents, this task also concerns video documents. Actively reading a video is fundamentally different from simply "looking at" it. It supposes the will to understand the document in its depth, to connect facts with others, compare sequences, and so on. To do so, the reader needs to create private notes, to link them directly onto segments of the read video, exactly like a researcher annotates a private copy of a paper. Active readers also frequently wish to know what other readers think about the studied documents. Of course, the reader is usually an author who writes her own documents, inserts archive items into them, and annotates them in the same manner. For instance, a history teacher at a university enjoys preparing her own video from highly relevant archive segments selected to illustrate her discourse.

All of these considerations make the INA portal quite different in its purpose from portals of most of Internet access providers.

3. DESIGN RATIONALE OF OPALES

The OPALES project relies on the following assumptions:

- Sharing one's knowledge with other people improves one's work efficiency [22].
- One uses a tool only when the return is greater than the effort to use the tool.
- To be efficient on a machine, a user needs interacting seamlessly with the objects (s)he studies as well as those (s)he produces.

To do so, OPALES provides each of its registered users with a private workspace. The purpose of the workspace is threefold:

- Enable the user to work on archive documents and on other documents as freely as if they were private copies, and to use them as raw material for their own use.
- Keep track not simply of the "production", but also of the work, e.g. the interpretation of facts observed on the videos. We call it "elaborated knowledge".
- Manage the sharing of elaborated knowledge with other users. This last point implies the use of efficient but flexible open collaboration techniques in order to facilitate structure emergence from the end users efforts [7].

The overall result is also threefold:

- The user produces for her own use new documents and new knowledge from the archives. This is supposed to be the basic reason why (s)he works on the system. No one sustains a long effort when there is no personal return.
- The effort done by a user at work is capitalized by sharing it with others. This results in a direct return from the OPALES system which incrementally improves the available knowledge about documents.
- Knowledge sharing between users can be done either for free or be accounted, in this case generating knowledge business. Some expert group may import knowledge about the archive documents from other expert groups to improve their own understanding of documents and provide other experts with this improved knowledge. Dealing with knowledge business is out of the current scope of OPALES whereas knowledge sharing accounting is already handled in the system.

These considerations match the initial goals:

- First, capitalizing and sharing user knowledge in the system boosts everyone's efficiency. This idea was strongly promoted by Douglas Engelbart. One may consider OPALES as an implementation of a NIC (Network Improved Collectivities) [9].
- Second, the result of users work directly benefits to the owner of the portal: the elaborated and capitalized knowledge constitutes an added value to the documents, which makes them more attractive and more valuable for access by new users through the portal.
- Third, users access documents on the OPALES portal for working and preparing their own documents. The workspace offers seamless interaction with any kind of document: from archives documents to users' own documents and even to documents built as shared knowledge.

4. THE POINT OF VIEW NOTION

4.1 Design rationale of the point of view notion

4.1.1 A shared ontology

Sharing knowledge implies that the users agree on the meaning of some vocabulary. This is done by representing knowledge in the system according to a shared ontology [10], [11]. This ontology is used internally in OPALES for indexing documents and computing on indexing.

Nevertheless, two major problems must be solved for cumulating user efforts:

- Providing users with an extensible representation mechanism for freely representing their own knowledge.
- Inducing a strong structure of the resulting knowledge in a non intrusive way.

4.1.2 Extensible Ontology

The first problem implies that the ontology cannot be static. Although OPALES is a restricted access system open to people who share the same need to understand and interpret archive contents, there is no restriction on the topics on which experts focus. Moreover, the diversity of expertise domains is precisely the interest of the system, because no library could afford such a large panel of experts to index the documents.

When annotating video sequences, experts in a given domain need to be allowed to handle concepts specific to their domain, which are mostly specialization of existing ones. As a consequence, they must be allowed to enhance the shared ontology accordingly, under some control.

4.1.3 Non intrusive interaction scheme

The second problem implies finding a good balance between constraints and freedom. This is one of the originalities of OPALES. If the structure is too strongly constrained by the system, in an intrusive manner, the user in hampered. Her activity reduces and the overall efficiency collapses. Conversely, if the structure is too weak, the knowledge elaborated by some users may become soon incompatible with the knowledge elicited by others, leading to messy and unusable results. As a consequence, regulation mechanisms based on community management are needed to avoid an anarchic evolution of the ontology. This mechanism is provided in OPALES, owing to the choice of an internal knowledge representation scheme directly computable. It enables the system to control for example the evolution of the ontology and to make users who edit the ontology aware of the existence of concepts similar to those they want to add.

4.1.4 Points of view as knowledge clusters

To deal with these problems, OPALES introduces the original notion of "point of view" which enables to virtually organize the users work into dynamically adaptable virtual communities in order to manage clusters of locally consistent knowledge. Dealing with inconsistency is a complex and delicate problem, even for humans. It becomes harder and harder as and when the scope of the knowledge widens and the amount of metadata increases, which is the case in OPALES. In order to keep the inconsistency in reasonable and manageable limits, we have made the choice to break it down, by dynamically identifying smaller scopes of knowledge in which sets of users can locally manage by themselves the consistency of their sub-domain. The result is that knowledge is self-organizing in locally consistent small clusters which directly reflect the structure of user expert groups. For instance, if some users have expertise in "fashion and dressing in the sixties" and need to introduce new concepts in the ontology, it is easier to them to locally manage the suitable extension. Thus, evolution of the ontology remains local and does not conflict with extensions needed by other experts, for instance those of "horses races". In order to insulate the clusters and organize their overall structure, a technique similar to XML namespace is used: we call it a "point of view". The extensions of the ontology and of the elicited knowledge are explicitly attached to the domain for which they have been added: they belong to a "point of view".

OPALES provides means to create at will clusters called "authoring points of view" and to elicit knowledge into them. It symmetrically provides means to take advantage of knowledge elicited according to different points of view, so that a reader may mix the knowledge elaborated by several communities.

4.2 Virtual communities

Most of OPALES users are experts, for instance in history, sociology and so on. Their expertise makes them, implicitly or explicitly, belong to "virtual communities". A community is said virtual when its members do not need to know each other. A virtual community exists as soon as some people have identified and named their concern, thus making explicit to others some knowledge, some interest, some hobby, and wish to share it, anonymously or not, with others [15], [6]. Virtual communities emerge on the web everyday. We call such communities virtual to stress the fact that belonging to a community does not require to be introduced, to pay for it, nor to adhere to some predefined ideas. A virtual community exists when a topic is made explicit by naming it and precisely identifying it, and when some people feel concerned by it. In OPALES, a virtual community is implicitly created when an author defines a new point of view and makes it public. At that moment, other users can feel concerned with writings related to this point of view as readers or as authors.

4.3 The notion of "point of view" in OPALES

4.3.1 Definition

The term "point of view" seems quite familiar but is used in OPALES with a very precise and restrictive meaning. We define it as a statement of the author about her authoring activity which sets the document in the concerns of a virtual community. Contrary to some familiar meaning, the "authoring point of view" of a document is not the semantics of the document itself. For instance, two experts may annotate a video on "Cashmere War" with completely contradictory interpretations, whereas they share a same vocabulary to express it, and have the same concern. In OPALES, their annotations belong to the same point of view: "India and Pakistan matters experts" regardless to the actual content of the annotation. Conversely, the same video may be annotated with the point of view of a "video reporter school teacher" who would comment the narrative structure, the framing of shots, the choice of images and so on. "India and Pakistan matters" and "video reporter school teacher" are quite distinct points of view. They can be used to annotate the same document. A "Economical international relationships expert" would annotate the same document in a quite distinct manner.

The notion of point of view in OPALES enables writers to explicitly tell to which virtual community their writings are dedicated. It induces clustering of knowledge and enables to use the specific community vocabulary which is appended to the shared ontology as depending on the point of view. It implicitly defines in this way local namespaces which drastically reduce ambiguities.

4.3.2 Using and managing points of view

The kernel of OPALES internal architecture handles private and public documents, points of view, annotations and indexing in a unified, reflexive, and consistent manner. Consequently, we use the term "piece of information" rather than the term "document" which could be understood with some restrictive meaning. To any piece of information is attached a resource descriptor which includes an "authoring point of view" stamp, an owner stamp, a type, and a status tag, and so on. For portability reasons, resources are externally described as RDF descriptors [21], [14]. A "workspaces database" keeps track of all the resources and of their interdependencies. Points of view are implemented like stamps attached to any piece of information. They characterize in which context information makes sense. For reflexivity reason, points of views are also considered as "pieces of information": a unique document of type "point of view" (which is primitive in the system) is associated to each point of view, as its informal description. This document is indexed by a precise indexing pattern, which enables the system to retrieve points of views. Thereby, there is strictly no difference between indexing points of view and other documents. The same mechanism applies for retrieving them.

The role of this mandatory indexing pattern associated to each point of view is to formally characterize it with respect to the shared part of the ontology from which the point of view is visible. It enables any author both to retrieve existing points of view defined by other authors and to declare new ones so that other authors can be aware of their existence. For many reasons, which are out of the scope of this paper, the OPALES internal knowledge representation formalism is NCG, the "nested conceptual graphs model" [4]. NCG enables a more precise indexing than keywords. For instance, NCG makes it very simple to distinguish between "transportation of sailing boats", "transportation by sailing boat", and "transportation of sails of boats". Another important result about NCG is a fuzzy matching algorithm [16] used for comparing NCG representations; it takes advantage of specialization, generalization and composition relationships in the ontology. It enables to compute distances between NCGs and thus to determine which are the closest points of view to a given one. For instance, an expert analyzing a movie of the 2nd World War can annotate it from a "medical expert" point of view or from one of its specialization as "nutrition expert" or as "psychiatry expert". As a consequence, the search engine would retrieve psychiatry annotations as specialization of medical expert annotations. Points of view and vicinity of points of view are the base for retrieving annotated documents and annotations, which are meaningful for a virtual community. This is the internal basement for the points of view and virtual community management in OPALES.

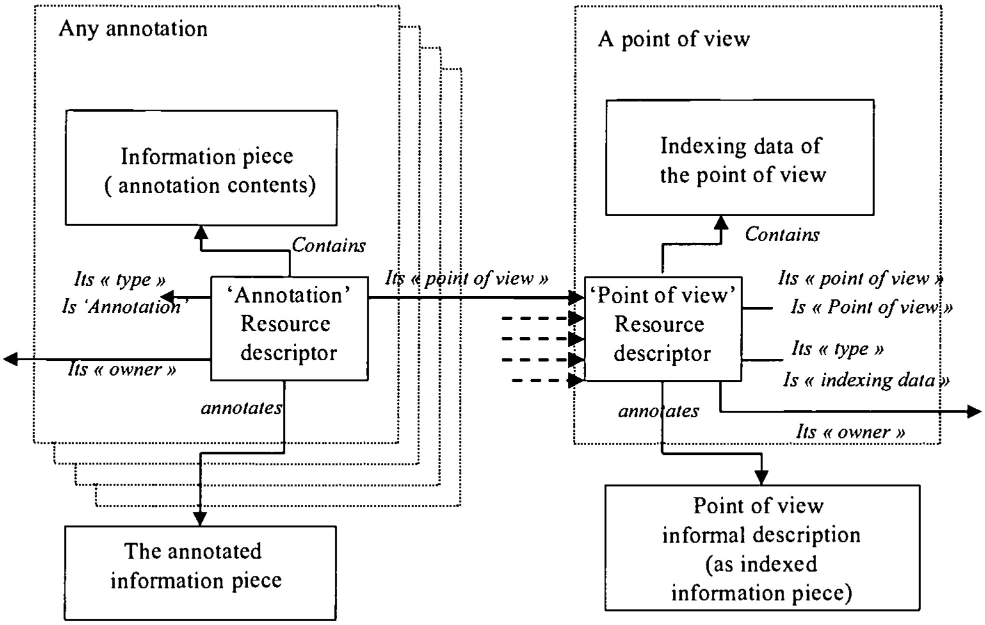


Figure 1:

Reflexivity in OPALES internal structure: annotations, indexing, points of view... are handled in a unified manner.

4.4 How authors interact with points of view

4.4.1 Selecting or defining a point of view

One of the requirements of OPALES design is a very low overhead for users. The point of view management sub-system is designed so that it provides users with more return than it requires efforts to put it in action. Any created piece of information (annotation, document, indexing) automatically becomes a resource stamped with the point of view associated to the window in which it was edited, and typed by the editor's type.

When a user logs in OPALES, her private workspace displays the last state in which the user logged out. Thereby, the list of her favorite authoring points of view, as created in previous sessions, is already available. A "current" point of view is kept marked in the list. It is assigned to any new window for stamping any editing actions taking place in it. A pop up menu enables to easily change the "current" point of view of a window whenever needed.

As for any other document, retrieval of a point of view not in the favorite list is achieved by means of a query. OPALES interface helps elaborating the query according to the ontology, by contextually selecting the vocabulary. Points of view close to the favorite ones can also be directly accessed in a browser interface. If the user considers that no existing point of view matches her current authoring situation, she creates a new one, most often by specialization of an existing one. Let us remark that, if no relevant point of view can be found, the query itself is very close to the formal indexing of the new point of view, thus making the burden to create new points of view quite limited.

All this just requires the author is conscious of the context in which she works. This assumption is fully compatible with OPALES users groups.

In most of cases, annotating existing documents or creating new ones does not require the author explicitly deals with points of view, since the current point of view is automatically assigned by default when an information chunk is created.

4.4.2 Exporting points of view

Any information piece (or document) in OPALES has a status tag which indicates whether the chunk is public or private. A private document can be accessed only by its author, whilst a public document can be read by anyone but edited only by its author. For consistency internal reasons and use of reflexivity in the implementation architecture, points of view are handled as documents. For sure they are so, because they have a content (their informal description), they are indexed exactly like any other document, they have an author who created the point of view, and a point of view ("point of view creator" which is primitive in the system). As a consequence, like any document, a point of view can be either private or public. Making a document or a point of view public is called "exporting" it. This makes it potentially visible to other users. This enables users to privately handle their annotations in their private workspace and later export them as well as the associated points of view.

4.4.3 Owners of documents

Any piece of information resource in OPALES has an owner and a point of view. No one except its owner may edit a piece of information. For consistency reasons, this applies to archive documents as well as to annotations and private documents. The term owner must be understood not as the copyright ownership but as the person or the institution who is responsible of the storage of the information in the system. An archive (video, image, sound record, text...) is under the responsibility of an institution (INA, MSH,...) who added it to the portal ; the institution is its OPALES "Owner". The point of view of an archive document is "archive" which is primitive in the system. This is quite consistent with the notion of point of view: for instance, an indexing with the "archive" point of view precisely is the genuine "INA" indexing associated to the document. Like any other document an archive can be public or private. In this last case, it is not visible for the end users, but may be handled by its owner. This feature is useful for instance during the first indexing stages of documents done before exporting them.

4.5 Annotating videos with OPALES

4.5.1 Stratified annotations

OPALES allows stratified [20] indexing and annotation of video. Freely stratified annotations are independent annotations whose anchoring in a document may overlap at will. Although automatic scene recognition tools easily provide a primary segmentation of video, it is now well known that this kind of segmentation is insufficient for precise indexing. For instance, in news, topics are announced and start with the speaker face on the screen. Automatic scene separation suggests starting a new segment when the image changes from the speaker to another image, whereas such an event may occur in the middle of a sentence. Breaking it or shortening it may deeply alter its semantics. This kind of segmentation is visual but, not at all, semantic, like those which are the concerns of OPALES. Because users index and annotate documents themselves, they are allowed to freely define segments and annotate them. For instance a specialist of body language may study hand motion of politicians during speeches. The segments she needs in order to put her expertise in action are quite different from those needed by a specialist of rhetoric. Stratified indexing is suitable so that annotations can freely overlap.

4.5.2 Annotation versus indexing

An annotation is an informal metadata, i.e. any information piece linked to a document. In OPALES there is no constraints on its content. An annotation can be simply the name of a person on an image of a group of guys and a link with a geometrical anchor to locate the person on the image. It may also be a long and argued discussion about some events of the currently selected segment. It can be a typed link towards another document.

At the other extreme, indexing is a formal data anchored into a document, and internally represented as a NCG. Formally indexing a document consists in providing typed annotations (type is "indexing", which is primitive) containing computable metadata which enables the internal search engine to retrieve it. Since indexing is just a specialization of annotations, as many private indexing, with specific points of view can complement the archive indexing of a document and thus describe richer semantics on specific segments as well as on the whole document.

Indexing a video segment or any part of a document is achieved by making a selection in the information piece and opening an annotation window of type "indexing". A specific NCG based indexing tool opens in the annotation windows. Indexing patterns can be defined by communities of users and attached to points of view in order to help indexing and ensure consistency of indexing rules within a point of view. Regulation mechanisms are provided by the user community management sub-system. Some virtual groups may become explicit, work closer together and elect moderators. This is a problem of user management, which is out of the scope of the paper.

5. EXPLOITING THE NOTION OF POINT OF VIEW

5.1 Reading versus authoring points of view

The notion of point of view would have no interest if it were not the key feature for readers working on documents. It is used to improve the information retrieval mechanism and provide finer access to the annotation base. We distinguish the notions of "authoring point of view" and of "reading point of view".

On the one hand, an authoring point of view characterizes the virtual community dedicated by an author to an annotation when he creates it. An annotation or an indexing is characterized by only one authoring point of view. On the other hand, a reading point of view characterizes which sources of annotations a reader wants to see linked as complements to a displayed document, and which complementary indexing information the OPALES search engine will use to retrieve more relevant documents. A reader can use different reading points of view to observe annotations and indexing of video segments.

Therefore, authoring points of view and reading points of view are distinct notions handled separately by the system. Let us suppose a reader wishes to integrate sociologic and economic sources as complementary information in her studies in order to get a deeper understanding of the studied videos. For retrieving more relevant videos, she also mixes in the queries concepts defined in extension on the ontology part associated to these points of view. The union of "economy" and "sociology" corresponds to her "reading point of view". Her authoring point of view simply is "childhood expert" which is her specialty. She considers her neither as a sociology expert nor as an economy expert and would not write annotations or indexing as such. She imports these points of view in her workspace just to constitute a "reading point of view". She may export her annotations written with the "childhood expert" point of view, inducing in this way a kind of knowledge commerce between users.

5.2 Defining a reading point of view

In a user's workspace, any editor or browser window has an associated "reading point of view" which acts as a filter to enhance its contents. The favorite reading points of view of a user are kept in a list in order to enable her to quickly set the point of view associated to her windows. Defining a new reading point of view is usually achieved by specifying an ordered set of authoring points of view. The reader just drags and drops some authoring points of view to define this new reading point of view. She can explicitly name it for further reuse. She can also explicitly define it in the same manner as a new authoring point of view, for instance by taking advantage of generalization mechanisms.

A list of annotations selected according to the reading point of view associated to a window is dynamically associated to the currently displayed document. The listed annotations are those which have been authored in one of the points of view referenced in the reading point of view, and which were linked as annotations anchored to the current selection in the displayed document. For instance, let us suppose the reader has selected some segment of an archive video as an answer to a search query, and looks at it. Since she observes it from a given reading point of view, all the available annotations for this point of view which are linked to any segment of this video that includes the current time code are listed. When seeking the video, the annotation list is dynamically updated according to the current position. Moving the cursor over the list displays a short preview of the selected annotation, thus avoiding unnecessarily link firing. When an annotation is geometrically anchored into the video, moving the mouse over its reference in the annotation list shows its anchorage directly on the video, under the condition the video is in the paused mode. This feature is extremely pleasant, for instance for scanning names of participants on a picture of a group.

5.3 OPALES system architecture

OPALES system architecture, as shown on figure 2, relies on the cooperation of three servers. The main server delivers archive video data and icons of selected shots. The workspace server stores all private and shared information pieces which are not archives, and uses a database for managing descriptors. It delivers enhanced information according to the selected reading point of view. Most of interactions are locally handled by a plug-in on the client browser. The knowledge server is based on a NCG engine developed at LIRMM [16]. It stores the ontology and all the indexing data.

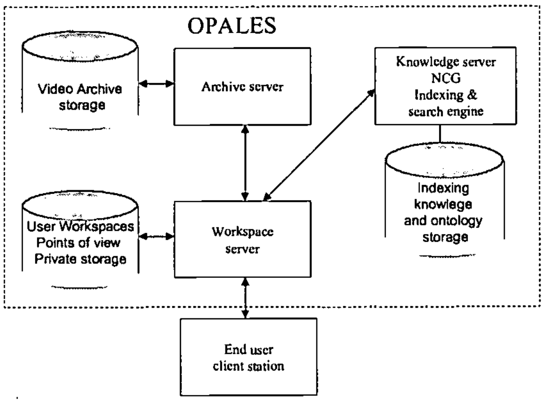

OPALES
Video Archive storage
Archive server
Knowledge server NCG Indexing & search engine
Indexing knowlege and ontology storage
User Workspaces Points of view Private storage
Workspace server
End user client station

Figure 2: OPALES system architecture.

6. DISCUSSION

The structure of users' work with OPALES emerges as the consequence of using a very simple set of rules associated with the private workspaces:

- Each user feels like working privately on her own copies of documents.
- If a reader selects the "archive" point of view, she only sees genuine information.
- If a reader imports some points of view, the displayed documents are enhanced with annotations accordingly.
- Searching for points of view is done in the same manner as searching for documents.
- Only the owner of an information may alter it. Imported information is inalterable.
- All information pieces created by a user keep track of the point of view in which they were created.
- A user may export and import points of views.

As a consequence,

- Any information made public is always, *de facto*, organized into a structure based on the point of view description in the ontology. When it is exported, it is cumulated in the system in an organized and non intrusive manner for the users, which induces very little overhead.
- The cumulated effort is made available to the collectivity of users in such a way that a user may focus only on her sub-domains. The reading point of view acts as a dynamically adjustable filter, which spares the burden to express complex queries. Furthermore, the point of view notion is far richer to express semantics than keywords are, since it precisely expresses the author's intention, whether or not relevant keywords are present in the annotation.

7. CONCLUSION

Patrimonial video archives contain considerable amounts of highly valuable information about our society. Contrary to books, which can be automatically analyzed once digitized for enhancing their indexing, digital video still requires human expertise to be relevantly indexed. The OPALES project offers a solution to enhancing the elicited knowledge about a part of the INA archive library.

Relying on users' work is a challenge. The web has assessed the outstanding power of users collaborating together. The Semantic Web Project [2] trusts this assumption as well. ÓPALES design aims at providing users with both simple and efficient mechanisms to share their knowledge. Ease of use seems to us a strict prerequisite to bootstrap knowledge sharing between users, and to cumulate it in the library. The concept of "point of view" and its implementation in OPALES are a key for reducing the complexity of huge amounts of knowledge independently elicited by groups of users. Although OPALES has been designed for enhancing video archives, the described techniques are directly transposable to other types of digital libraries.

Experiments for observing users' behavior and adjusting mechanisms are on the way.

REFERENCES

[1] Aigrain, P., Petkovic, D., & Zhang, H.J. Content-based representation and retrieval of visual data: a state of the art review, Multimedia Tools and Application, Special issue on representation and retrieval of visual media, 1996.

[2] Berners-Lee, T. Semantic Web Road Map, http://www.w3.org/DesignIssues/Semantic.html, 1998.

[3] Chang, S.F. *et al.*, VideoQ: An automated content-based video search system using visual cues. In Proc. ACM Multimedia'97 (1997), pp. 313-324.

[4] Chein, M., Mugnier, M.L., & Simonet., G. Nested Graphs: A Graph-based Knowledge Representation Model with FOL Semantics, in Proc. 6th International Conference on Principles of Knowledge Representation and Reasoning (KR'98), (1998), pp. 524-534, Morgan Kaufmann Publishers.

[5] Crowley, J.L. & Berard, F. Multi-Modal Tracking of Faces for Video Communications, IEEE Conference on Computer Vision and Pattern Recognition, CVPR '97, Puerto Rico, (1997).

[6] Davenport, G. & Pan, P. I-Views: a Community-oriented System for Sharing Streaming Video on the Internet, in Proc. WWW9 Conference (1999).

[7] Dieberger, A. Supporting Social Navigation on the World-Wide Web. International Journal of Human Computer Studies: Special Issue on Novel Applications of the WWW (in press).

[8] Dieberger, A., Dourish, P., Höök, K., & Wexelblat, A. Social Navigation: Techniques for Building More Usable Systems, Interactions, Vol. VII.6, 2000.

[9] Engelbart, D., Networked Improved Communities, Keynote at ACM Conf. Hypertext'98, (1998). ACM SIGWEB video archive. See also http://www.bootstrap.org

[10] Garino, N. Formal ontology, conceptual analysis and knowledge representation. Int. Journal of Human-Computer Studies, 43 (5/6), pp. 625-640, 1995.

[11] Gruber, T.R. Toward principles for the design of ontologies used for knowledge sharing. In *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Nicola Guarino and Roberto Poli, editors, Kluwer Academic, in preparation. Original paper presented at the *International Workshop on Formal Ontology*, March 1993. Available as Stanford Knowledge Systems Laboratory Report KSL-93-04. On line: http://ksl-web.stanford.edu/knowledge-sharing/papers/onto-design.rtf

[12] Hauptmann, A. and Smith, M. Text, Speech, and Vision for Video Segmentation: The Informedia Project, AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision, 1995. See also: http://www.informedia.cs.cmu.edu/

[13] Houghton, R. Named Faces: Putting Names to Faces. IEEE Intelligent Systems Magazine, Vol 14, No. 5, pp. 45-50, 1999.

[14] Kahan, J., Koivunen, M.R., Prud'Hommeaux, E., & Swick R.R. Annotea: An Open RDF Infrastructure for Shared Web Annotations, in *Proc. of the WWW10 Int. Conference*, Hong Kong, (2001).

[15] Martin Röscheisen, M. & Winograd, T. Beyond browsing: shared comments, soaps, trails, and on-line ccommunities, in Proc. WWW3 Int. Conference (1995).

[16] Mugnier, M.L. Knowledge Representation and Reasonings Based on Graph Homomorphism, in Proc. 9th International Conference on Conceptual Structures (ICCS), (2000).

[17] Olligschlaeger, A., Hauptmann, A. Multimodal Information Systems and GIS: The Informedia Digital Video Library, ESRI User Conference (1999).

[18] Peirce, C.S. Ecrits sur le signe, Editions du Seuil, Paris, 1978.

[19] Staab, S., Erdmann, M., Maedche, A., & Decker, S. An Extensible Approach for Modeling Ontologies in RDF(S), Workshop on Semantic Web associated to ECDL'2000.

[20] SIGIR conferences, ACM Press.

[21] Smith, A., & Davenport, G. The Stratification System: A Design Environment for Random Access Video. In ACM Workshop on Networking and Operating System Support for Digital Audio and Video, San Diego, California (1992).

[22] Stone, V.E. Social Interaction and Social Development in Virtual Environments. Presence 2, 2 (Spring 1993), pp. 153-161.

XSLT for Tailored Access to a Digital Video Library

Michael G. Christel
CS Dept. and HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213
1-412-268-7799
christel@cs.cmu.edu

Bryan Maher
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA 15213
1-412-268-8970
bsm@cs.cmu.edu

Andrew Begun
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA 15213
1-412-268-2029
apb@andrew.cmu.edu

ABSTRACT

Surrogates, summaries, and visualizations have been developed and evaluated for accessing a digital video library containing thousands of documents and terabytes of data. These interfaces, formerly implemented within a monolithic stand-alone application, are being migrated to XML and XSLT for delivery through web browsers. The merits of these interfaces are presented, along with a discussion of the benefits in using W3C recommendations such as XML and XSLT for delivering tailored access to video over the web.

Categories and Subject Descriptors

H.5.1 **[Information Interfaces and Presentation]**: Multimedia Information Systems – *video.* H.3.7 **[Information Storage and Retrieval]**: Digital Libraries – *standards, dissemination, user issues.*

General Terms

Design, Human Factors, Standardization.

Keywords

Digital video library, XML, XSLT, surrogate.

1. INFORMEDIA INTERFACES

The Informedia Project at Carnegie Mellon University has created a multi-terabyte digital video library consisting of thousands of hours of video, segmented into over 50,000 stories, or documents. Since Informedia's inception in 1994, numerous interfaces have been developed and tested for accessing this library, including work on multimedia abstractions or *surrogates* which represent a video document in an abbreviated manner [4, 5]. The utility and efficiency of these surrogates have been reported in detail elsewhere [1, 2, 3, 14], validated through a number of usability methods, including transaction log analysis, formal empirical studies, contextual inquiry, heuristic evaluation, and cognitive walkthroughs. This paper begins with an introduction to a few of these interfaces and their implementation history. The promise of web technologies is then discussed, particularly the recommendations of the World Wide Web Consortium (W3C), leading to a presentation of the Informedia digital video library delivered through a web browser via XML and XSLT. Emphasis is placed on the tailored accessibility offered by this information architecture, with specific examples given as evidence. The paper concludes with a discussion of next steps planned for the Informedia library work.

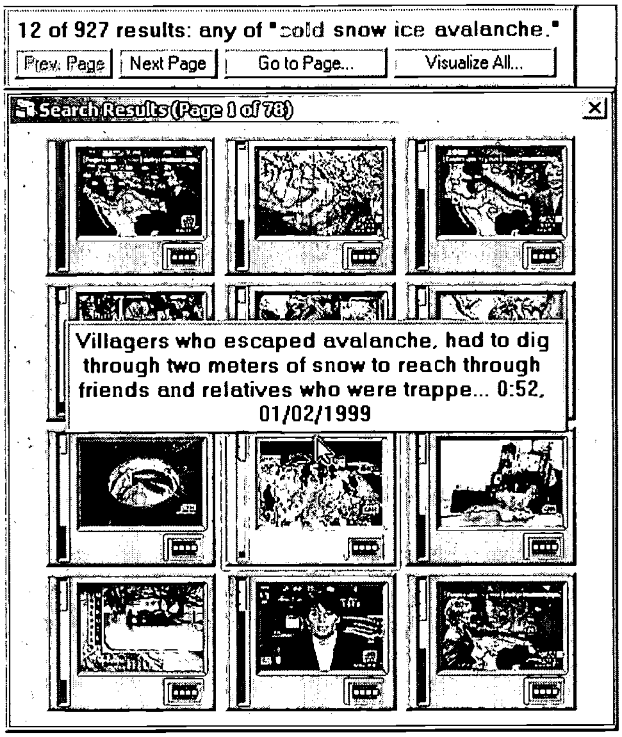
JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.

1.1 Informedia Surrogates

Video is an expensive medium to transfer and view. MPEG-1 video, the compressed video format used in the Informedia library, consumes 1.2 Megabits per second, and looking through a ten minute video for a section of interest could take a viewer ten minutes of time. Surrogates can help users focus on precisely which video documents are worth further investigation, reducing viewing and video data transfer time. Example Informedia surrogates include brief titles and single thumbnail image overviews, as shown in Figure 1 for 12 documents.

The Figure 1 interface shows query-based thumbnail images: the image is selected from the neighborhood of the document where the highest match scores occurred. In this example, the first few documents show weather maps, indicating that most of the matching to the query "cold snow ice avalanche" occurred in portions of the documents where weather maps were shown. By contrast, the ninth document shows a snowplow, indicating footage of snow and a plow where the query terms are discussed most frequently in the story. Past work showed the utility of choosing thumbnails based on context rather than simply choosing the first visual for a document, and for packing the result set with thumbnails rather than solely listing text titles, document durations and broadcast dates [1].

The vertical bar to the left of each thumbnail indicates relevance to the query, with color-coding used to distinguish contributions of each of the query terms. The document surrogate under the mouse cursor, the eighth result, has its title text displayed in a pop-up window, and the query word display is also adjusted to reflect this particular document. The document is part of the results set primarily because it mentions "avalanche" frequently with some mention of "snow." In Figure 1, "cold" and "ice" are grayed out to show they don't apply to the currently focused document, and the vertical relevance bar for the document shows only two colors: a small patch for "snow" and a large extent for "avalanche." Hence, the display of Figure 1 makes use of relevance bars, query word color-coding, context-specific thumbnail selection, and additional pop-up text information to present a page of documents to the user.

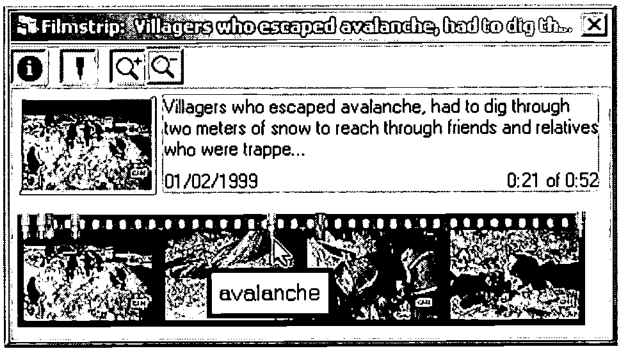


12 of 927 results: any of "cold snow ice avalanche."

Villagers who escaped avalanche, had to dig through two meters of snow to reach through friends and relatives who were trappe... 0:52, 01/02/1999

Figure 1. Thumbnail results page for 12 documents, with one pop-up title shown.

From Figure 1's interface, clicking on the filmstrip icon ▭ for a document displays a storyboard surrogate with the visual flow of that document, along with locations of matches to a query, as shown in Figure 2.

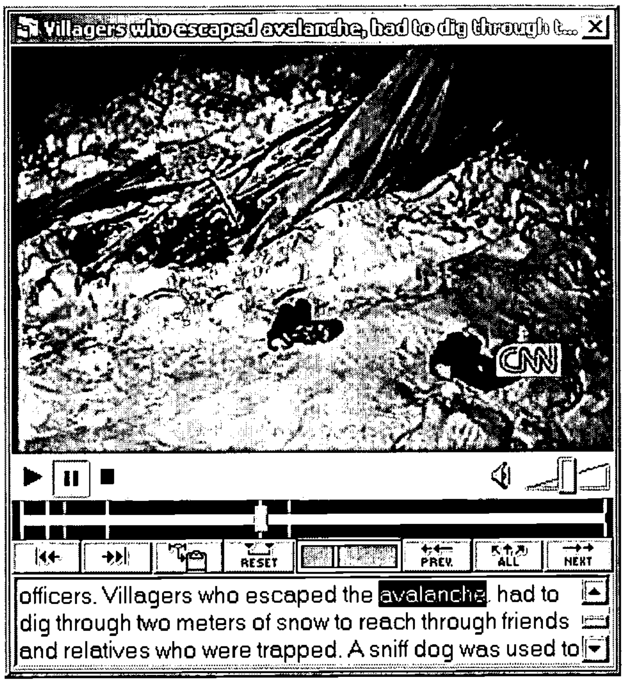

Filmstrip: Villagers who escaped avalanche, had to dig th...
avalanche

Villagers who escaped avalanche, had to dig through two meters of snow to reach through friends and relatives who were trappe...

01/02/1999 0:21 of 0:52

Figure 2. Storyboard, showing that "avalanche" is discussed 21 seconds into the 52-second video document.

Such an interface is equivalent to drilling into a document to expose more of its details before deciding whether it should be viewed. Storyboards are also navigation aids, allowing the user to click on an image to seek to and play the video document from that point forward. For example, Figure 3 shows the video playback window for this document, complete with synchronized transcript, started at this point by clicking on the Figure 2 storyboard's second image. These surrogates are built from metadata automatically extracted by Informedia speech, image, and language processing modules, including transcript text, shot boundaries, key frames for shots, and synchronization information associating the data to points within the video [14].



officers. Villagers who escaped the avalanche, had to dig through two meters of snow to reach through friends and relatives who were trapped. A sniff dog was used to

Figure 3. Video playback window, complete with match lines and scrolling transcript.

Figures 1, 2, and 3 show the typical interaction progression of users during the first years of the library. A text search was entered, results were returned as in Figure 1, titles and thumbnails were browsed, with optionally more detailed surrogates as that of Figure 2 examined, leading to some videos being played with the interface of Figure 3. Many fewer videos were actually played compared to the total number returned by text searches.

While the surrogates were put to use, they were not sufficient to deal with the richness of a growing library. As the Informedia collection grew from tens to thousands of hours, the results set from queries grew from tens to hundreds or thousands of documents. Whereas a query on "cold snow ice avalanche" might have produced 30 results that could all be shown on a single screen, later queries against years of CNN news produced too many documents to afford a direct examination of each thumbnail. Figure 1 shows the results of a query against 1998 and 1999 news, producing 927 results. Visualization techniques were added to provide overviews of the full result set and to enable user-directed inquiries into spaces of interest within this result set.

1.2 Informedia Visualization Techniques

The three main visualization techniques employed in the Informedia library interface are:

- Visualization by Example (VIBE), developed to emphasize relationships of result documents to query words [12].
- Timelines, emphasizing document attributes to broadcast date [4].
- Maps, emphasizing geographic distribution of the events covered in video documents [5].

Each technique is supplemented with dynamic query sliders, allowing ranges to be selected for attributes such as document size, date, query relevance, and geographic reference count. The visualizations shown here convey semantics through positioning, but could be enriched to overlay other information dimensions through size, shape, and color, as detailed elsewhere [4, 5].

By combining multiple techniques, users can refine large document sets into smaller ones and better understand the result space. For example, the 927 documents of the query in Figure 1 produce the VIBE plot shown in Figure 4. By dragging a rectangle bounding only the points between words, and excluding the points at just a single query word, the user can reduce the result set to just those documents matching two or more of the terms "cold snow ice avalanche." This operation is shown in Figure 4, reducing the focused result set from 927 documents to 281.

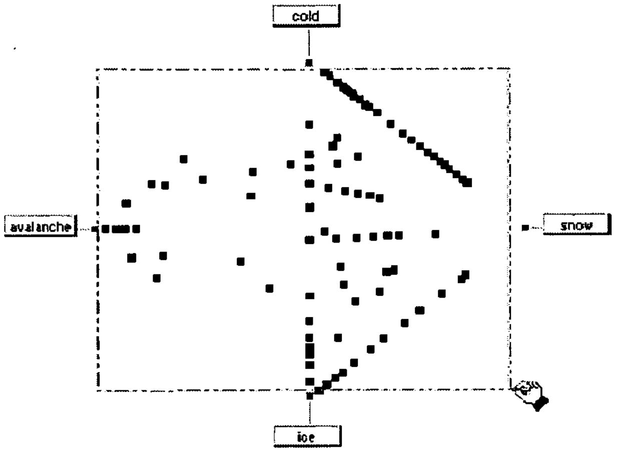


Figure 4. Selecting area of VIBE plot mapping to "two or more of the terms 'cold snow ice avalanche'".

VIBE allows users unfamiliar or uncomfortable with Boolean logic to be able to manipulate results based on their query word associations. For video documents such as a news corpus, there are other attributes of interest besides keywords, such as time and geography. Figure 5 shows a timeline that portrays the obvious (considering that the news corpus originates in the Northern Hemisphere): results from the "cold snow ice avalanche" query cluster in the winter months of November to March.

Figure 6 shows a snapshot of a sequence of interactions that trim down the 281 documents from Figure 4's interaction to a very manageable set of 11. A map view of the results shows a number of highlighted countries, some mentioned only once peripherally in news stories discussing two or more of "cold snow ice avalanche." By highlighting only countries mentioned 4 or more times, tangential references are given less consideration. The user can drag a time window, through the date slider shown below, to set a time period for which to plot results. The user can also manipulate the map, zooming into Europe as a region of focus. In this manner, the user discovers that when looking at February 1999 the documents are concentrated in Austria and Switzerland.

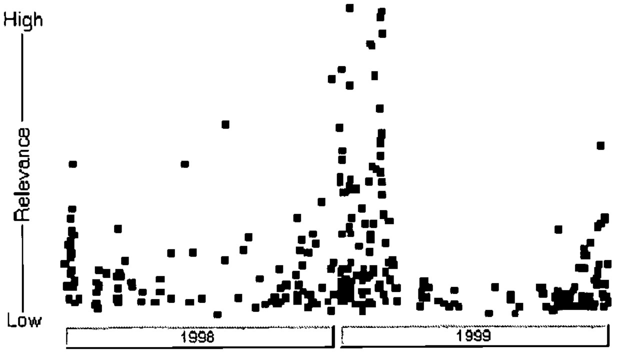
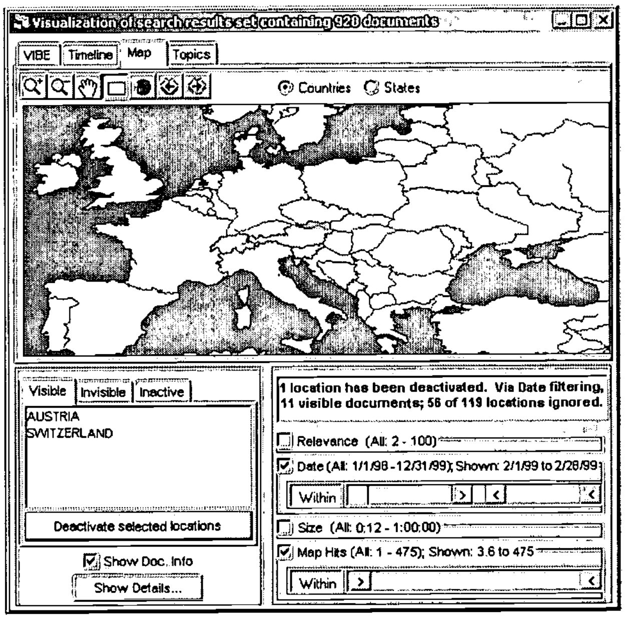


Figure 5. Timeline plot for Figure 4 subset, showing density of results in winter months.


Visualization of search results set containing 920 documents
VIBE
Timeline
Map
Topics
Visible
Invisible
Inactive
Deactivate selected locations
Show Details...
Within
Within

○ Countries ○ States

AUSTRIA
SWITZERLAND

☑ Show Doc. Info

1 location has been deactivated. Via Date filtering, 11 visible documents; 56 of 119 locations ignored.

☐ Relevance (All: 2 - 100)

☑ Date (All: 1/1/98 - 12/31/99); Shown: 2/1/99 to 2/28/99

☐ Size (All: 0:12 - 1:00:00)

☑ Map Hits (All: 1 - 475); Shown: 3.6 to 475

Figure 6. Map plot and dynamic query sliders, showing two European countries within February 1999 time focus.

This section has outlined through example the evolution of Informedia digital video library interface work. This work began with surrogates to enable the exploration of a single video document without the need to download and play the video data itself, and migrated to visualization techniques to allow the interactive exploration of sets of documents. A monolithic, Visual Basic Windows application provides these interfaces, allowing users to query or browse through text, image, and map searches, refine the result space with visualization techniques, and browse through surrogates such as titles, thumbnails, and storyboards.

The developments of the past year, particularly new W3C Recommendations and their implementation in major Web browsers, provided the opportunity to migrate this video library work to the Web. The remainder of this paper discusses this migration, emphasizing the benefits offered and the flexible library interface front-end provided to the user.

2. XML AND XSLT

"The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential as a forum for information, commerce, communication, and collective understanding" (verbatim from www.w3c.org). A number of key W3C Recommendations were published in 1999, enabling the separation of authoring from presentation in a standardized manner. In the case of the Informedia library, these recommendations allow the separation of video metadata from the library interface. Last year saw gradual implementation and support for these recommendations, documented at the W3C web site. The Informedia work described below makes use of the Microsoft XML Parser 3.0, an Internet Explorer add-on released by Microsoft in November 2000. The W3C Recommendations used in migrating Informedia interfaces to a Web browser include the following:

- XML (Extensible Markup Language): the universal format for structured documents and data on the Web, W3C Recommendation February 1998 [16].
- XML Schema: express shared vocabularies for defining the semantics of XML documents, not yet a full W3C Recommendation as of January 2001 [18].
- XSLT (XSL Transformations): a language for transforming XML documents, W3C Recommendation Nov. 1999 [19].
- XPath (XML Path Language): a language for addressing parts of an XML document, used by XSLT, W3C Recommendation November 1999 [17].

Other emerging standards for synchronized media metadata, such as MPEG-7 [9] and SMIL [15], will be tracked and incorporated as they become adopted by video streaming services and web browsers.

"Metadata" describes an information resource; it is "data about other data" [8]. A metadata record consists of a set of attributes necessary to describe the resource in question. For the Informedia video library, some attributes such as the producer, copyright holder, and broadcast date are given. A number of other attributes, such as start and end times, shot sequences, thumbnails, and transcripts, are automatically derived as input video is processed, segmented into documents, and catalogued.

The Informedia metadata is stored in a relational database and accessed through the application overviewed in Section 1. Such a closed system makes interoperability with other digital libraries difficult. A separate video collection might be described with a different set of metadata, or have that metadata stored in a different fashion.

An idealistic vision is to have a standard video metadata scheme, so that all video collections could be described to the same level of detail, accessed in the same manner, and have identical surrogates and interfaces built from the common scheme. However, video genres like news, sports, situation comedies, travel, lectures, and conference presentations have such diverse features that deriving a detailed, general video library metadata scheme will be a difficult if not impossible task. More likely, a common metadata framework will evolve, probably with input from professional societies in related disciplines like the Association of Moving Image Archivists. Using this common metadata framework as a foundation, more specific metadata could be added to more accurately describe resources in particular video collections.

The Dublin Core Metadata Initiative provides a fifteen-element set for describing a wide range of resources. While the Dublin Core "favors document-like objects (because traditional text resources are fairly well understood)" [8], it has been tested against moving-image resources and found to be generally adequate [7]. The Dublin Core is also extensible, and has been used as the basis for other metadata frameworks, such as an ongoing effort to develop interoperable metadata for learning, education and training, which could then describe the resources available in libraries like the Digital Library for Earth System Education (DLESE) [6]. Hence, Dublin Core is an ideal candidate for a high-level metadata scheme for the Informedia video library. An outside library service, with likely support for Dublin Core, would be able to make use of information drawn from the Informedia video library expressed in the Dublin Core element set.

The Dublin Core metadata for Informedia documents can be expressed as XML and validated through the use of a data type definition, or XML schema. More detailed metadata is necessary to produce the interfaces shown in Figures 1 through 6, but this metadata too can be expressed as XML and validated through a more comprehensive XML schema. In fact, a richly detailed XML document can be transformed into a minimal Dublin Core view, or transformed into views like those shown in Figures 1 through 6, with transformations performed via XSLT. Multiple XSLT transformations, e.g., one for low bandwidth users, another for high bandwidth users, optional additional ones for specific languages, age groups, etc., allow the video data to be widely disseminated in different forms based on W3C standards.

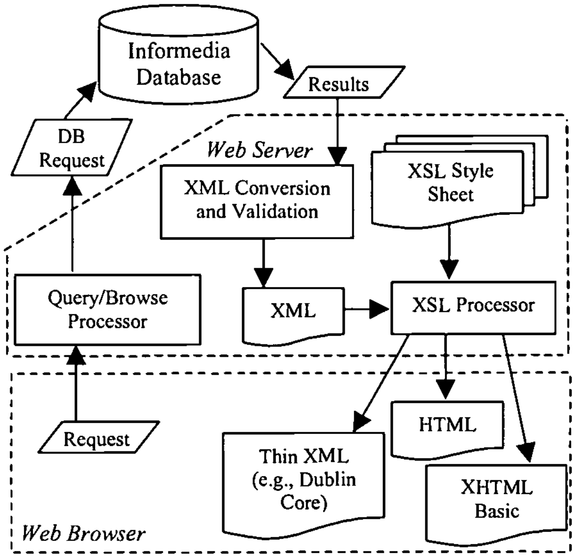


Figure 7. Architecture showing multiple outputs from XSL processing.

Figure 7 shows the process of a query or browse request against the Informedia database, producing XML results that are validated and data-typed via an XML schema. These XML results can be processed with different XSL style sheets to produce different library interfaces, such as an XML view consisting of Dublin Core elements, an HTML view that may look like Figure 1, or an XHTML Basic view suitable for display everywhere, including tiny PDAs. The next section gives specific examples, and discusses how tailored library access can be enhanced with XSL processing done in the client web browser rather than the web server.

3. TAILORED ACCESS TO DIGITAL VIDEO LIBRARY MATERIALS

In a recent editorial on "informationitis", Ramesh Jain notes that today's Web users and digital library patrons are overwhelmed by too much information. The traditional means for retrieving information has been keyword indexing and search, but abstracting the search level to keywords removes a great deal of relevant context for multimedia documents. In addition, presenting a list of documents returned from a keyword query involves perhaps a painstaking linear traversal of the list to find a document, with no gestalt view of the query space nor the results, i.e., no understanding of the relationship between result documents [11]. The editorial reinforces the Informedia interface conclusions drawn in the opening section: as the library contents increase in quantity, information visualization approaches need to be employed to facilitate understanding and navigation through larger document sets.

Speech recognition, image processing, and natural language processing allow automatic derivation of metadata to use as building blocks for subsequent generation of interfaces such as those shown in Section 1 [14]. The same metadata can be stored as XML and converted into numerous views through XSLT, where the views are tailored to a user's needs and bandwidth requirements. This section presents examples of XML and XSLT that implement such views, and discusses an architecture fostering quick presentation of multiple views into the digital video library, based on user selection. Users drive the library exploration and navigation, highlighting different aspects of document context to address their information needs and overcome "informationitis."

3.1 Informedia Access through XML and XSLT

Consider Figure 1 once again, showing a thumbnail view for a set of documents retrieved through an Informedia search service. These documents could be described in XML, as follows (listing shows only first and eighth result for Figure 1, to save space):

```
<IDVSet xmlns:im="x-
    schema:idvSchema.xml">
  <im:doc>
    <im:id>160814</im:id>
    <im:pos>1</im:pos>
    <im:shot>1961294</im:shot>
    <im:d_yr>1999</im:d_yr>
    <im:d_mo>1</im:d_mo>
    <im:d_day>14</im:d_day>
    <im:score>100</im:score>
    <im:dur>151250</im:dur>
    <im:mmss>2:31</im:mmss>
    <im:title>On Monday that cold air in place over
      upper midwest and great lakes with
      showers over midwest and snow in great
      lakes ...</im:title>
  </im:doc>
  <im:doc>
    <im:id>157053</im:id>
    <im:pos>8</im:pos>
    <im:shot>1931480</im:shot>
    <im:d_yr>1999</im:d_yr>
    <im:d_mo>1</im:d_mo>
    <im:d_day>2</im:d_day>
    <im:score>80</im:score>
    <im:dur>52120</im:dur>
    <im:mmss>0:52</im:mmss>
    <im:title>Villagers who escaped avalanche,
      had to dig through two meters of snow to
      reach through friends and relatives who
      were trappe...</im:title>
  </im:doc>
</IDVSet>
```

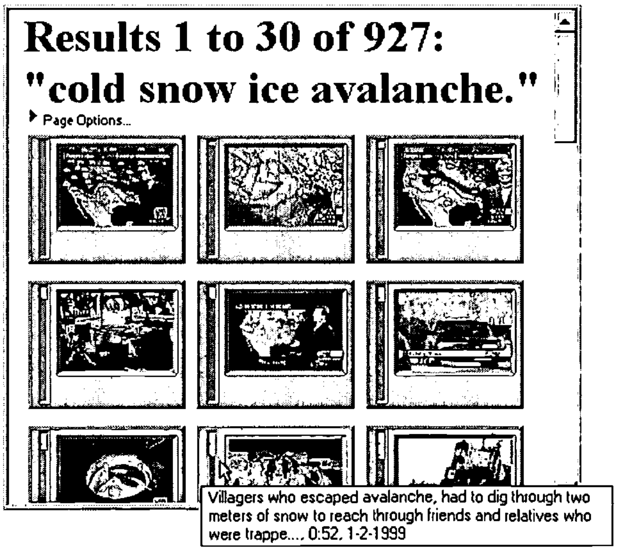
The referenced schema "idvSchema.xml" is used to validate and provide data type semantics for this XML text. Consider this subset of contents from idvSchema.xml:

```
<?xml version="1.0" ?>
<Schema name="IDVResultsSchema"
    xmlns="urn:schemas-microsoft-com:xml-data"
    xmlns:dt="urn:schemas-microsoft-
    com:datatypes">
  <ElementType name="score" content="textOnly"
      dt:type="ui1" />
  <ElementType name="doc" content="mixed">
      <element type="score" maxOccurs="1" />
  </ElementType>
</Schema>
```

These schema definitions limit "score" to appearing at most once for each document "doc", with "score" being an unsigned one-byte integer. The schema defines other requirements and types for "IDVSet." The validated XML can be transformed into the view shown in Figure 8 through the following XSL style sheet, which loops through each im:doc document metadata and converts it into appropriate HTML:

```
<xsl:stylesheet xmlns:xsl=
  "http://www.w3.org/1999/XSL/Transform"
  version="1.0" xmlns:im=
  "x-schema:idvSchema.xml">
  <xsl:output method="xml" indent="yes"
     omit-xml-declaration="yes" />
  <xsl:template match="/">
      <xsl:apply-templates />
  </xsl:template>
  <xsl:template match="IDVSet">
   <xsl:for-each select="im:doc">
    <xsl:sort select="im:score" order="descending"
       data-type="number" />
    <span class="resultStamp" id="R{im:pos}"
       rdb_id="{im:id}"
       onclick="stampClick(this);"
       onmouseover="stampChangeOver(this);"
       onmouseout="stampChangeOut(this);">
    <img id="Stamp_{im:pos}"
       src="graphics/Gstamp.gif" alt=""
       orgsrc="graphics/Gstamp.gif"
```

```
        oversrc="graphics/Gltstamp.gif"
        width="112" height="91" />
    <xsl:variable name="ScoreHt"
        select="round(im:score * 0.8)" />
    <!-- map 100 score to 80 px (im:score .le.100) -->
    <img id="Th_{im:pos}" src="graphics/red.gif"
        alt="">
      <xsl:attribute name="style">
       position:absolute; left:9; width:4; top:
        <xsl:value-of select="85-$ScoreHt" />
       ; height:
        <xsl:value-of select="$ScoreHt" />
       ;
      </xsl:attribute>
    </img>
    <img id="I_{im:pos}"
        style="position:absolute; left:23; top:9"
        width="80" height="55">
      <xsl:attribute name="src">
        <xsl:choose><xsl:when test="im:shot[. !=
                0]">GetShot.asp?<xsl:value-of
                select="im:shot" />
        </xsl:when>
        <xsl:otherwise>Graphics/viddeflt.gif
        </xsl:otherwise></xsl:choose>
      </xsl:attribute>
    </img>
    <img id="tip" src="Graphics/1p-trans.gif"
        style="position:absolute; left:0; top:0"
        width="112" height="91">
      <xsl:attribute name="alt">
        <xsl:value-of select="im:title" />,
        <xsl:value-of select="im:mmss" />,
        <xsl:value-of select="im:d_mo" />-
        <xsl:value-of select="im:d_day" />-
        <xsl:value-of select="im:d_yr" />
      </xsl:attribute>
    </img></span></xsl:for-each>
  </xsl:template>
</xsl:stylesheet>
```



Results 1 to 30 of 927: "cold snow ice avalanche."

Villagers who escaped avalanche, had to dig through two meters of snow to reach through friends and relatives who were trappe..., 0:52, 1-2-1999

Figure 8. Browser display of XSL-transformed XML into HTML (a view similar to Figure 1).

XSLT is itself an XML document, and so the style sheet above reads as a jumble of starting and ending XML tags that essentially do the following: for each Informedia document, create a green stamp area (Gstamp.gif) with the relevance score in red on a vertical bar, a thumbnail image if given a valid nonzero identifier, and pop-up title text, duration, and broadcast date information. The produced html from this XSLT for document 8 is as follows:

```
<span class="resultStamp" id="R8" rdb_id="157053"
    onclick="stampClick(this);"
    onmouseover="stampChangeOver(this);"
    onmouseout="stampChangeOut(this);"
    xmlns:im="x-schema:idvSchema.xml">
  <img id="Stamp_8" src="graphics/Gstamp.gif"
      alt="" orgsrc="graphics/Gstamp.gif"
      oversrc="graphics/Gltstamp.gif" width="112"
      height="91" />
  <img id="Th_8" src="graphics/red.gif"
      style="position:absolute; left:9; width:4; top:31;
      height:64;" />
  <img id="I_8" style="position:absolute; left:23;
      top:9" alt="" width="80" height="55"
      src="GetShot.asp? 1931480" />
  <img id="tip" src="Graphics/1p-trans.gif"
      style="position:absolute; left:0; top:0"
      width="112" height="91" alt="Villagers who
      escaped avalanche, had to dig through two
      meters of snow to reach through friends and
      relatives who were trappe..., 0:52, 1-2-1999" />
</span>
```

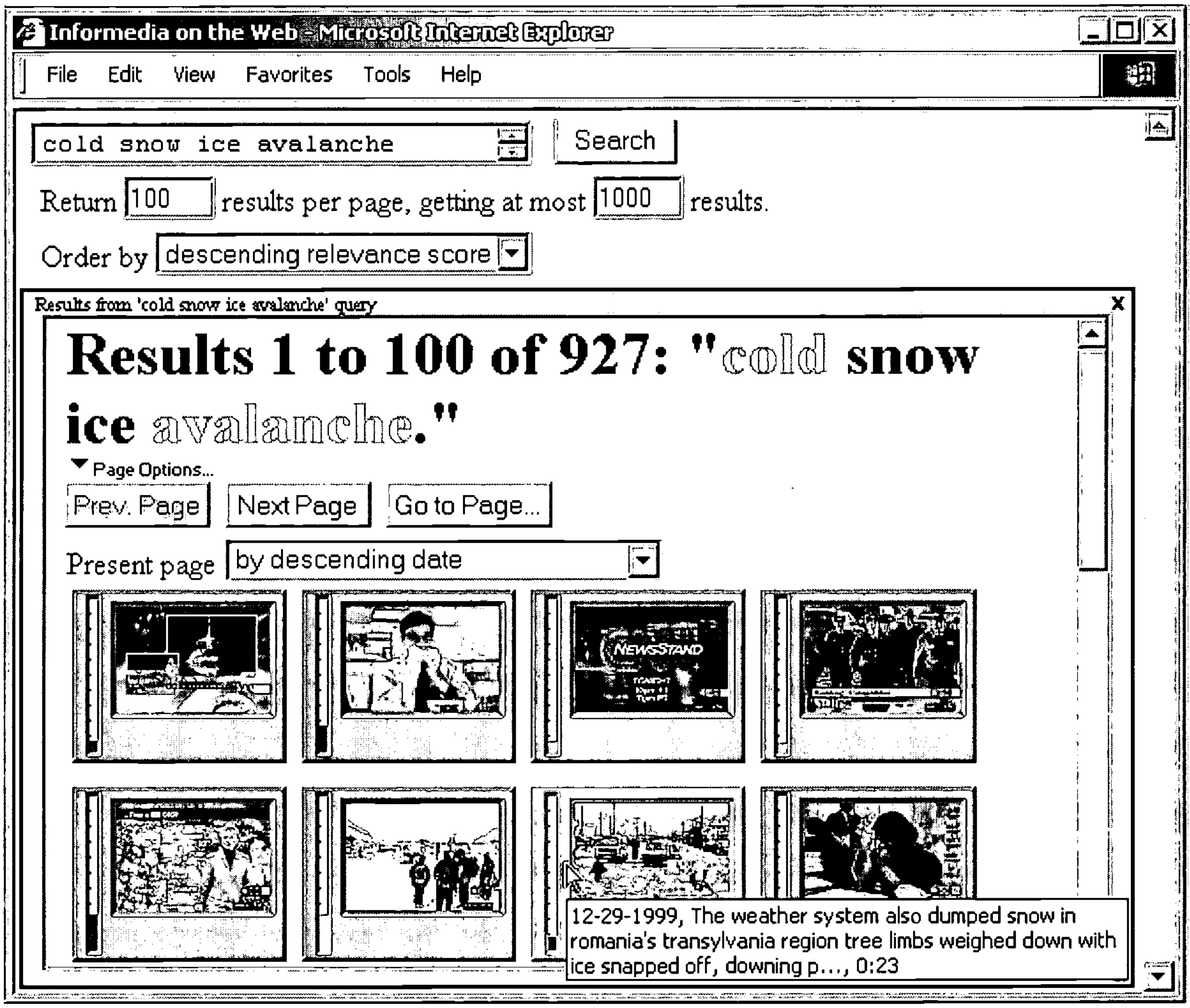
3.2 Enhancing Views with Match Data

By extending this simple opening example, match data information can be viewed by users in the same way as shown in Figure 1: through color coding of query terms and the vertical relevance score bar. The XML and schema definitions are extended to include information on which entities (in this case, words, but could be geographic regions, image features, etc.) match a video document, by how much and where:

```
<IDVSet xmlns:im="x-
      schema:idvResSchema.xml">
  <im:ScoreInfo>
    <im:ScoreEntity><im:mID>1</im:mID>
      <im:mLabel>cold</im:mLabel>
    </im:ScoreEntity>
    <im:ScoreEntity><im:mID>2</im:mID>
      <im:mLabel>snow</im:mLabel>
    </im:ScoreEntity>
    <im:ScoreEntity><im:mID>3</im:mID>
      <im:mLabel>ice</im:mLabel>
    </im:ScoreEntity>
    <im:ScoreEntity><im:mID>4</im:mID>
      <im:mLabel>avalanche</im:mLabel>
    </im:ScoreEntity>
  </im:ScoreInfo>
  <im:doc>
   {"doc" contents, e.g., im:id, im:pos as before}
    <im:m>
      <im:msrc>3</im:msrc>
      <im:mScore>386</im:mScore>
      <im:mOffset>528</im:mOffset>
    </im:m>
    <im:m>
      <im:msrc>3</im:msrc>
      <im:mScore>484</im:mScore>
    <im:mOffset>528</im:mOffset>
  </im:m>
  <im:m>
    <im:msrc>1</im:msrc>
    <im:mScore>333</im:mScore>
    <im:mOffset>249</im:mOffset>
  </im:m> {additional im:m continue here...}
</im:doc>
```

```
  {other "doc" contents likewise extended with
    match information via the im:m element}
</IDVSet>
```

The XSL style sheet is extended to make use of im:m match information, producing the view shown in Figure 9, which interactively changes the query word colors to indicate which words match in that document, and shows itemized scoring entity contributions in the vertical relevance bar (as done in Figure 1).



Informedia on the Web - Microsoft Internet Explorer

cold snow ice avalanche

Return 100 results per page, getting at most 1000 results.

Order by descending relevance score

Results from 'cold snow ice avalanche' query

Results 1 to 100 of 927: "cold snow ice avalanche."

Present page by descending date

12-29-1999, The weather system also dumped snow in romania's transylvania region tree limbs weighed down with ice snapped off, downing p..., 0:23

Figure 9. Display of HTML produced via XSL transformation of XML with match data.

3.3 Client-Side XSLT

The addition of XML data provides new interface functionality possibilities. By continuing with this strategy, the Informedia document XML description and its validating schema can be extended to that data necessary to generate all the interfaces described in Section 1, interfaces proven useful through prior investigations. The problem with such an approach is that perhaps the XML or XSLT-produced HTML would grow to huge sizes that take time to download in a Web browser, but never get viewed. Through XSLT in the client browser, however, users have the freedom to choose which views to use, with little or no need for communication back with the Web server.

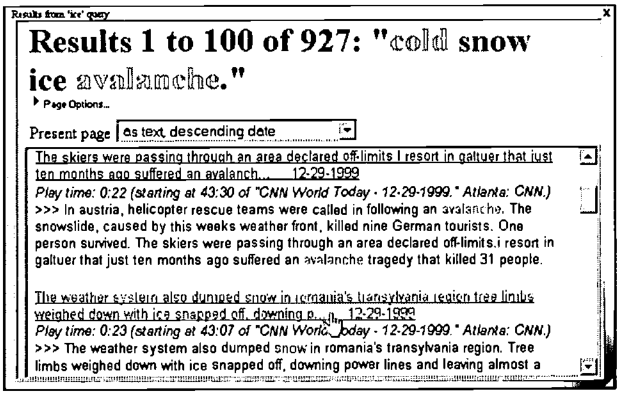
Figure 9 shows a "Present page" option where the user can select to order the page by relevance, date, or document size in ascending or descending order. The change in sort is accomplished through an XSL style sheet, e.g., the descending date is accomplished via the following:

```
<xsl:sort select="im:d_yr" order="descending"
  data-type="number" />
<xsl:sort select="im:d_mo" order="descending"
  data-type="number" />
```

```
<xsl:sort select="im:d_day" order="descending"
  data-type="number" />
```

The style sheet also reorders the pop-up information to give precedence to the date and lists that first, capitalizing on past experience that when sorting Informedia documents by date the user is more interested in that attribute and prefers such a reordering. Of course, the XSL style sheet could be altered to make the date information even more explicit. Client-side XSL transformations allow the user to sort and present the data to meet his or her specific browsing and information-seeking needs.

Other options available in "Present within page" include a text-centric view, shown in Figure 10, and a VIBE view, identical to Figure 4 and making use of the same XML with match information described in Section 3.2.


Results from 'ice' query

Results 1 to 100 of 927: "cold snow ice avalanche."

▸ Page Options...

Present page as text, descending date

The skiers were passing through an area declared off-limits I resort in galtuer that just ten months ago suffered an avalanch... 12-29-1999
Play time: 0:22 (starting at 43:30 of "CNN World Today - 12-29-1999." Atlanta: CNN.)
>>> In austria, helicopter rescue teams were called in following an avalanche. The snowslide, caused by this weeks weather front, killed nine German tourists. One person survived. The skiers were passing through an area declared off-limits.i resort in galtuer that just ten months ago suffered an avalanche tragedy that killed 31 people.

The weather system also dumped snow in romania's transylvania region tree limbs weighed down with ice snapped off, downing p... 12-29-1999
Play time: 0:23 (starting at 43:07 of "CNN World Today - 12-29-1999." Atlanta: CNN.)
>>> The weather system also dumped snow in romania's transylvania region. Tree limbs weighed down with ice snapped off, downing power lines and leaving almost a

Figure 10. Text view of same XML data presented with thumbnail grid in Figure 9.

The web architecture for the Informedia library is pushing dynamic interface selection to the user by sending XML data and XSL style sheets to the client-side browser. The first time a style sheet is referenced, e.g., to sort a page by date and emphasize that attribute, the style sheet is inserted into the browser's cache. Subsequent use of that transformation can then be applied without the need to contact the Web server. The user is free to explore multiple features of the document space through different views, from text-centric to image-centric, from linear lists to visualization strategies like VIBE.

The interface shown in Figure 9 also lets the user specify the size of the document set to be considered via multiple views, i.e., the "page size" indicating the number of documents described in XML for subsequent translation into HTML via XSLT. The user also sets the maximum number of documents cached at the server for potential future consideration. In this manner, users can control the flow of information to meet their bandwidth restrictions and patience thresholds. For example, a user on a T1 line may set a page size of 1000 and look through image-rich presentations such as multiple storyboards (Fig. 2), while a user with a PDA and 56 Kbps access may set the page size to 20 and make use of text-centric views.

While some transformations may require contacting the Web server to get additional data such as imagery, others are done completely at the client, making use of the original XML or previously cached information, as overviewed in Figure 11. For example, suppose the user initially defaulted to sorting documents by date, producing the html whose display is shown in Figure 9. The user now sorts and emphasizes by score, resulting in an ordering as shown in Figure 8. No new imagery is necessary, as the thumbnail image data has already been cached by the browser. Suppose the user now accesses "Present page by VIBE view" which requires the VIBE XSL style sheet to be downloaded the first time it is referenced. The style sheet is less than 2 KB in size, and is available in the browser's cache for quick reuse without the need to contact the web server the next time it is needed. Style sheets will generally be very small compared to the XML document. A vastly different VIBE presentation (Figure 4) of the document set utilizing match information is shown to the user with this style sheet, without needing to retrieve additional XML or data from the Informedia database.

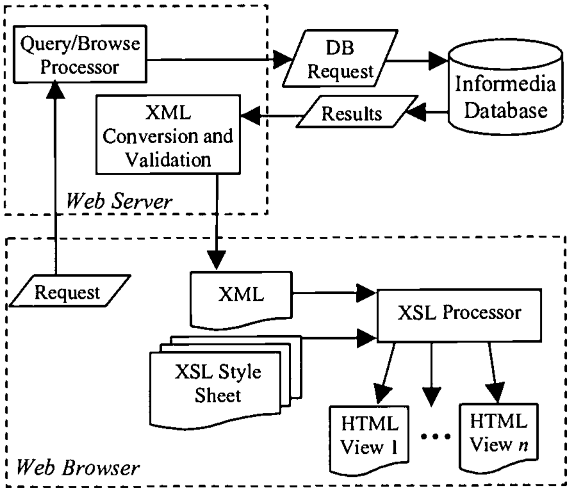


Figure 11. Overview of client-side XSL processing, where user interaction can produce multiple HTML views without Web server involvement.

3.4 Flexibility via XML and XSLT

Figure 7 shows already processed XML data being sent to clients. This architecture is useful for those clients with very focused or well-articulated needs. For example, another library service may need an Informedia document set expressed as Dublin Core elements, and the document set can be translated into that format by the Informedia Web server and sent to that service.

By contrast, Figure 11 shows XML data, along with XSL style sheets being communicated to clients. This allows clients to modify the views dynamically, offering flexibility to address the "informationitis" issues for multimedia libraries discussed in Jain's editorial [11]. Users can vary the views dynamically: those interested in image-rich overviews by date can be satisfied, as can users interested in query-specific set manipulation offered through VIBE. Given the numerous attributes and views into video collections, and the potential of each view to inform the user about specific characteristics like date, length, or geographic coverage, this architecture delays final rendering (in HTML or whatever form) of the semantic XML data until decisions made within the Web browser. In the examples used here, decisions are made through the "Present page" option.

4. CONCLUSIONS AND FUTURE WORK

Much work remains to be done in order to provide interoperable, tailored Web browser views into the Informedia library as expressive as those of the stand-alone Informedia library application. The W3C recommendations provide the ideal framework for creating these views, given the W3C's charter, broad industry support, and momentum from other National Science Foundation DLI-2 and NSDL projects also moving toward XML and XSLT; see for example DLESE [6] and the ACM SIGGRAPH Education Committee Digital Library [13]. XML, XSLT and related technologies XPATH and XML schemas allow semantics to be recorded, navigated, validated and translated in standard ways.

A necessary condition for widespread interoperability amongst digital video collections is agreement on a common metadata framework, as discussed in the usage guide for Dublin Core [8]. A common video metadata framework can be supported by Informedia and other video libraries through a default XSLT transforming the libraries' XML into this framework's XML. In all likelihood this framework would be an extension of Dublin Core, much as other groups such as metadata committees for learning, education and training are exploring use of Dublin Core as a foundation. A small subset of what such a minimal framework would look like for an Informedia document is as follows:

```
<?xml version="1.0" ?>
<!DOCTYPE rdf:RDF SYSTEM
 "http://purl.org/dc/schemas/dcmes-xml-
 20000714.dtd">
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
   syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description
   about="http://informedia.org/seg160814.mpg">
<dc:title>CNN World Today</dc:title>
<dc:description>On Monday that cold air in place over
   upper midwest and great lakes with showers
   over midwest and snow in great lakes
   ...</dc:description>
<dc:date>1999-1-14</dc:date>
<dc:format>video/mpeg</dc:format>
<dc:language>en</dc:language>
<dc:publisher>Cable News Network</dc:publisher>
<dc:contributor>Carnegie Mellon University
   Informedia Project</dc:contributor>
{Many more descriptors needed, e.g., coverage is from
   49:22 to 51:53 of the hour-long "World Today" show.}
</rdf:Description>
</rdf:RDF>
```

We will track closely the work of other digital libraries like DLESE that manage video resources, as well as the industry initiatives such as the work within the Association of Moving Image Archivists, as they address a common video metadata framework. In addition to providing a minimal but broadly applicable view (Figure 7), we also have the goal of migrating Informedia surrogates and visualizers to HTML-based expressions, so that they can be generated dynamically through XSL processing against XML within Web clients (see Figure 11). Hence, we will have a more detailed, "Informedia-rich" XML schema capable of supporting such enhanced views as those shown in Figures 1 through 6.

Work to date has addressed thumbnail grids, ordering, and query word-based views, including VIBE. Work is ongoing to provide interactive map interfaces, where zooming, panning, and map layer highlighting can be performed dynamically and efficiently. These features are required to provide a map visualization service like that shown in Figure 6, where countries highlight in different colors based on the user dragging a time period indicator across a scroll bar. We are currently investigating another W3C format, the Scalable Vector Graphics (SVG) format available as a Candidate Recommendation as of early 2001. SVG will allow quick map updating in the browser, as well as allow VIBE rendering to be more efficient so that greater numbers of documents can be shown simultaneously.

Improving summarization and visualization across video document sets is an ongoing research activity within the Informedia Project [10], and as new techniques become available, they will be added to the set of XSL style sheets available to the Informedia library patron. For example, work continues to identify faces within the video library, and name those faces with proper names. An interesting visualization along the lines of Figures 4 through 6 would be a key person/player view showing people's faces who dominate the news for particular time periods or for a specific text, image or geographic query.

We will continue implementing XSL style sheets and updating the Informedia-rich XML to allow users to have multiple views into the Informedia document sets. Future work includes usability tests on these views to investigate their utility and to determine the costs and benefits in supporting client-side XSL processing. Informedia metadata in particular is unusual compared to other libraries in that it is errorful, produced through automatic means without manual cataloging. Studies will need to be run to determine the effects of errorful metadata on subsequent XSL transformations and ultimately on the user's experience.

We will need to revisit the architecture of Figure 11 over time to see whether multiple style sheets operate on the same XML, or whether each style sheet has unique requirements for additional metadata from the Informedia database, and hence must contact the Web server anyway. If each XSL style sheet is essentially independent, requiring contacting the Web server, then there is no advantage to client-side XSLT. However, our first trials using XML with match information (Section 3.2) shows that the same XML supports diverse views, from thumbnails to plain text to VIBE. By adding match information to the "Informedia-rich" XML set, a match-specific view such as VIBE can be implemented through client-side XSLT. When a map view is added, metadata about geographic coverage for each Informedia document will need to be added to the XML. Should named faces be added, that metadata will need to be added to the XML as well. The same base XML can be grown to cover all the views, so that it is downloaded once and then operated on in the browser, an option that may be feasible given the expense of video data.

Video streaming is only now starting to reach a broader audience on the web. Video still requires comparatively large bandwidth and network integrity, and playback of web video beyond the tiny postage stamp window requires patience from even the well-connected university user on a T1 line. Users therefore may be willing to wait seconds to download lots of XML and associated XSL style sheets, so that they can then quickly browse through metadata representing hundreds of hours of video and megabytes or terabytes of actual video data. The views from XSLT allow a careful exploration of that material before investing in minutes or longer of video download time. Through the tailoring techniques described here, video library patrons can browse and explore video assets with minimal time commitments through surrogates and visualizations. These interfaces are rendered through W3C standards for increased potential to work within and across other digital video collections on the Web.

5. ACKNOWLEDGMENTS

Partial support for this work is provided by the National Science Foundation's National Science, Mathematics, Engineering, and Technology Education Digital Library Program under grant DUE-0085834. This work is also supported by the National Science Foundation under Cooperative Agreement No. IRI 9817496.

6. REFERENCES

[1] Christel, M., Winkler, D., and Taylor, C.R. Improving Access to a Digital Video Library. In *Human-Computer Interaction: INTERACT97, the 6th IFIP Conf. On Human-Computer Interaction*, Chapman & Hall, London, 1997, pp. 524-531.

[2] Christel, M., Smith, M., Taylor, C.R., and Winkler, D. Evolving Video Skims into Useful Multimedia Abstractions. In *Proceedings of CHI '98* (Los Angeles, CA, April 1998), ACM Press, 171-178.

[3] Christel, M.G., Hauptmann, A.G., Warmack, A.S., and Crosby, S.A. Adjustable Filmstrips and Skims as Abstractions for a Digital Video Library. In *Proc. ADL '99* (Baltimore MD, May 1999), IEEE Press, 98-104.

[4] Christel, M.G. Visual Digests for News Video Libraries. In *Proc. ACM Multimedia '99* (Orlando, FL, Nov. 1999), ACM Press, 303-311.

[5] Christel, M.G., Olligschlaeger, A.M, and Huang, C. Interactive Maps for a Digital Video Library. *IEEE MultiMedia* 7(1), 60-67.

[6] Ginger, K, web page maintainer. DLESE Metadata Working Group Homepage, Nov. 6, 2000, http://www.dlese.org/Metadata/index.htm.

[7] Green, D. Beyond Word and Image: Networking Moving Images: More Than Just the "Movies", *D-Lib Magazine*, http://www.dlib.org/dlib/july97/07green.html.

[8] Hillman, D. Using Dublin Core. July 16, 2000, http://purl.org/dc/documents/wd/usageguide-20000716.htm.

[9] Hunter, J. MPEG-7: Behind the Scenes. *D-Lib Magazine*, http://www.dlib.org/dlib/september99/hunter/09hunter.html.

[10] Informedia Project web site at Carnegie Mellon University, http://www.informedia.cs.cmu.edu.

[11] Jain, R. Informationitis. *IEEE MultiMedia* 7(4), 1, 5.

[12] Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B., and Williams, J. G. Visualization of a Document Collection: The VIBE System. *Information Processing & Management* 29(1), 69-81.

[13] Owen, G.S., Sunderraman, R., and Zhang, Y. Use of Database and XML Technology for Retrieval and Repurposing of DL Contents. Presentation at DLI-2 All Projects Meeting (Stratford-upon-Avon, U.K., June 2000), http://asec.cs.gsu.edu/asecdl-nsf-dli2/presentations/all-projects-meeting-6-2000-uk/asecdl-xml_files/frame.htm.

[14] Wactlar, H., Christel, M., Gong, Y., and Hauptmann, A. Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library. *IEEE Computer*, 32, 2 (Feb. 1999), 66-73.

[15] World Wide Web Consortium Synchronized Multimedia (SMIL) 1.0 Recommendation, June 1998, http://www.w3.org/AudioVideo/.

[16] World Wide Web Consortium Extensible Markup Language (XML) 1.0 Recommendation, Feb. 1998, http://www.w3.org/XML/.

[17] World Wide Web Consortium XML Path Language (XPATH) 1.0 Recommendation, Nov. 1999, http://www.w3.org/TR/xpath.html.

[18] World Wide Web Consortium XML Schema Candidate Recommendation, Oct. 2000, http://www.w3.org/XML/Schema.

[19] World Wide Web Consortium XSL Transformations (XSLT) 1.0 Recommendation, Nov. 1999, http://www.w3.org/TR/xslt.

Design of A Digital Library for Human Movement *

Jezekiel Ben-Arie
EECS Department M/C 154
University of Illinois at Chicago
851 S. Morgan St., SEO 1120
Chicago, IL 60607, USA.
benarie@eecs.uic.edu

Purvin Pandit
EECS Department
University of Illinois at Chicago
ppandit@eecs.uic.edu

ShyamSundar Rajaram
EECS Department
University of Illinois at Chicago
srajaram@eecs.uic.edu

ABSTRACT

This paper is focused on a central aspect in the design of our planned digital library for human movement, i.e. on the aspect of representation and recognition of human activity from video data. The method of representation is important since it has a major impact on the design of all the other building blocks of our system such as the user interface/query block or the activity recognition/storage block. In this paper we evaluate a representation method for human movement that is based on sequences of angular poses and angular velocities of the human skeletal joints, for storage and retrieval of human actions in video databases. The choice of a representation method plays an important role in the database structure, search methods, storage efficiency etc.. For this representation, we develop a novel approach for complex human activity recognition by employing multidimensional indexing combined with temporal or sequential correlation. This scheme is then evaluated with respect to its efficiency in storage and retrieval.

For the indexing we use postures of humans in videos that are decomposed into a set of multidimensional tuples which represent the poses/velocities of human body parts such as arms, legs and torso. Three novel methods for human activity recognition are theoretically and experimentally compared. The methods require only a few sparsely sampled human postures. We also achieve speed invariant recognition of activities by eliminating the time factor and replacing it with sequence information. The indexing approach also provides robust recognition and an efficient storage/retrieval of all the activities in a small set of hash tables.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Motion, Tracking*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Pattern Analysis*; E.2 [**Data Storage Representation**]: Hash-table representation

General Terms

Design,Algorithms

Keywords

Human Activity Recognition, Multi Dimensional Indexing, Temporal correlation, Sequence Recognition

1. INTRODUCTION

Human movement analysis is an important ingredient in many areas such as kinesiology, biomechanics, rehabilitative procedures, ergonomic evaluations of job tasks, anthropology, cultural studies, sign language, athletic analysis and sports medicine. In addition, research in artistic areas such as dancing, gymnastics, figure skating, ethnic studies and behavioral studies also require analysis, representation and classification of human motion. There are many libraries that include video and other motion data (like dance choreography notations [7]) that can be utilized as sources of raw data. However, such raw data do not accurately quantify and represent the actual three dimensional human motion, which can be quite complex. Furthermore, there is no universal method for accurate and detailed representation of human motion that can be employed to uniquely define search queries for generic or particular types of human actions/activities. Obviously, there is a need for a general method for representation of human action that can be employed to uniquely specify, store and retrieve such actions in digital libraries.

The overall architecture of our planned digital library which is called as HUman MOtion Retrieval System (HUMORS) is shown in Fig. 1. In this figure the system is divided into three parts, the user interface, the motion/action recognition module and the motion sequence and learning module. This paper is focused on the motion/action recognition module which enables to recognize human activity from video sequences using hash table representation. The other two modules are described briefly to outline the complete structure of the system. The user interface module handles the user's queries and provides a graphic feedback of humanoid animation for interactive querying.This module is designed to accept three modes for query input: spatio-temporal, symbolic and visual. In the symbolic query mode, the user can specify a query which is composed of a sequence

*This work was supported by the National Science Foundation (NSF) Grants No. IIS-9711925, IIS-9876904 and IIS-9979774.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

of basic actions from a menu. A basic action could be raising a hand or nodding the head. Such basic actions are similar to the actions that are defined in dance notations such as Labatonian Sutton Dance notation [7]. This input mode is quite limited and may not include all the possible human actions. Whenever necessary, it is proposed to add new actions to the actions menu by composing them with the spatio-temporal query module.

The user interface module also has a temporal sequencer which can combine these basic actions into more complete human activities. These sequences can be fed into a humanoid animation module which displays the resulting motion sequence and provides visual feedback to the user. The second mode of input query is the spatio-temporal mode. This mode is required for articulated and accurate specification of human motion. In this mode, the user quantitatively specifies angles and velocities of specific body parts which take part in the specified action. The third mode for motion query is query by visual example, where the query is selected from a video database or is presented from an external video source. One option for an external video source might be a video camera which captures the user himself who can directly demonstrate what he wants. The third module is used for storage and learning. This is a database of all the actions/activities known to the system. A new sequence may be added to the database, whenever it is sufficiently dissimilar to any of the existing sequences in the database. By this manner, the system is capable of learning new motion sequences.

The objective of this paper is to perform an initial assessment of the feasibility of a proposed representation method for human movement that is based on angular poses and angular velocities of the human skeletal joints, for analysis, querying and retrieval of human actions in video databases that describe human motion. The representation method plays a major role in determining the overall database structure, search methods, storage efficiency and other important facets and therefore any choice has to be evaluated very carefully.

Natural languages and symbolic temporal descriptions are not suited to **accurately** describe **articulated** human movements or actions. Natural language enables to describe human actions only in generic terms such as "walking", "running", "swimming", etc. Such actions constitute in reality, a sequence of complex motions of body parts that may differ from person to person. Other languages based on symbolic representations such as Spatio-Temporal Logic (STL) [5], Hierarchical Temporal Logic (HTL) [22], or Symbolic Projection based languages [3] [12] are also quite limited and can describe only gross spatial relations and motions between different rigid objects. Obviously, the human body is not a rigid object and its actions cannot be specified in such a limited representation. There are other notations that were developed to describe human motion in actions such as dancing [7] [8], figure-skating[9], athletic exercises[9] and natural gestures[24]. However, all of these notations are useful only for their specific domain and cannot universally describe articulated human motion. In addition, these methods also are based on symbolic representations and therefore quantize the motion of human body parts very coarsely.

Recently, the Virtual Reality Modeling Language (VRML) is being developed as a tool for animation of humanoids in networked applications (http://ece.uwaterloo.ca/~h-anim/) in conjunction with the development of MPEG-4 / SNHC [1] (http://www.es.com/mpeg4-snhc/)

The VRML humanoid provides a framework for accurate representation of human motion by enabling to define the angular poses and velocities of all the joints in the human body. In this work, we learn from the general structure of the VRML humanoid model and we propose a vectorial representation of human motion that is capable of accurate and detailed description by a sequence of multi-dimensional vectors. This proposed representation is elaborated in Section 2. As detailed in Section 4, we develop an efficient methods to store and index such vectorial sequences. This enables accurate analysis of articulated human activity/action as well as fast retrieval and articulated formulation of queries.

In order to evaluate the efficacy of our proposed angular pose/velocity representation , we develop and compare several approaches for the recognition of human activity based on indexing of sparsely sampled angular poses/velocities of the limbs and the torso. The sampled poses/velocities are obtained by tracking body parts in video sequences. We develop in this paper three indexing based methods for human activity recognition that differ in the pose and the temporal information used.

Human body's static posture frequently gives an indication of the action that takes place. This fact is evident from observing static images that reveal in many cases the actual activity. Based on this idea, we examine in this paper the possibility of recognizing human activities just from few sampled postures. A person's posture is composed of the poses of arms, legs, torso and head. Human activity can be described as a temporal sequence of pose vectors that represent sampled poses of body parts. Our principle of recognizing human activity from sparsely sampled postures is based on identifying these postures as samples of a complete, densely sampled model activity. To achieve this objective, we construct a database that includes all the model activities in the form of entries in multidimensional hash tables. The size of these tables is not too large since, body parts have limited angular motion and thus the number of bins that describe the full range of angular motion of each body part is quite limited.

An important feature of our approach is the separation of the multi-dimensional indexing into several hash tables, where each table corresponds to a different body part. This structure enables to index and recognize activities even when several body parts are occluded. Also, our approach of using multidimensional tuples proves to be very efficient in terms of storage since all the activities are stored in the same table. Experimental results described in Section 5 demonstrate robust recognition of activities.

Several approaches for activity recognition have been reported in the literature. However, none of these works aimed at a complete human activity recognition as is demonstrated in this paper. Schlenzig, Hunter and Jain [20] use Hidden Markov Model (HMM) and a rotation-invariant imaging representation to recognize visual gestures such as "hello" and "good-bye". HMMs are also utilized by Starner and Pentland [23] to recognize American Sign Languages (ASL). In

[1] The MPEG-4 standard is focused on content based coding of video objects such as animated humanoids whereas the new MPEG-7 standard which is designed for the far future is expected to address the issues of content based indexing and retrieval.

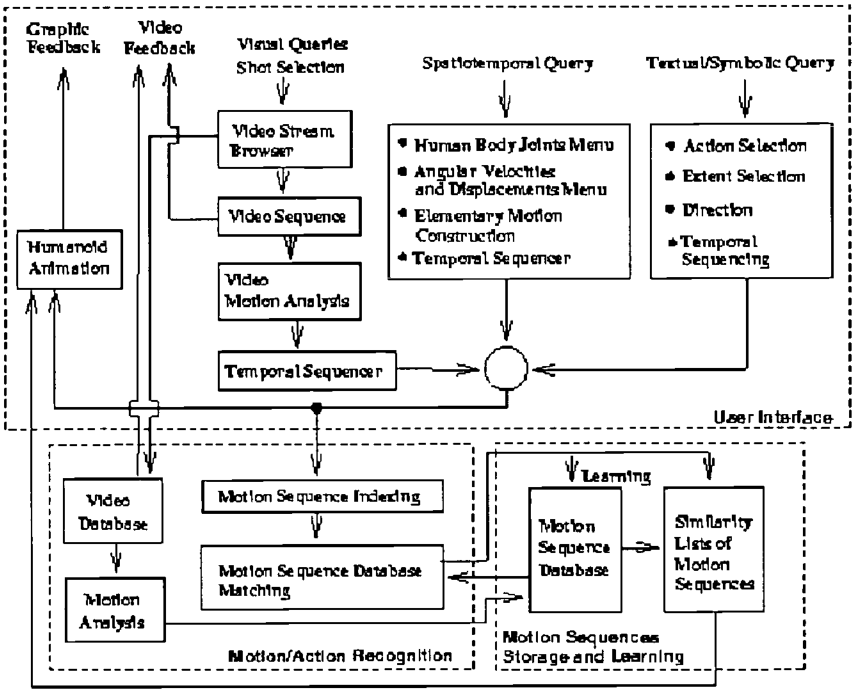


Figure 1: Architecture for the HUman MOtion Retrieval System (HUMOR-S)

these approaches, both the model and the test activities need to be densely sampled over time in order to maintain high recognition rate.

In this paper, we present a novel approach to complex human activity recognition by employing multidimensional indexing combined with temporal or sequential correlation. The representations of human activity models are actually sequences of body part poses. To be more specific, the postures of humans in each video frame are decomposed into a set of multidimensional tuples which represent the poses of human body parts such as arms, legs and torso. Whenever possible, the poses of the body parts are converted into a set of normalized angles to achieve size invariance. Hence, human activity is represented by a temporal or sequential arrangement of sets of multidimensional tuples that correspond to sampled angular poses of body parts over the entire time interval. The final outcome of this arrangement transforms human activity information into a set of hash tables each of which corresponds to an individual body part. The indices in these hash tables are the poses of the corresponding parts and the contents of those hash tables are the identities of the model activities and their time labels. At the recognition stage, a set of multidimensional indices are derived from the video sequence of the test activity.

As elaborated in the following sections, each video frame in the test activity sequence yields a 1D vote vector for each activity model in the database. The overall vote for each activity model is obtained by integrating the votes for all the test frames using temporal or sequential correlation. Details are provided in Section 4. One of the main advantages of this approach is the tremendous flexibility it provides in sampling the test activity sequence. There are no strict requirements either on the number of frames sampled or on the frame intervals of the test sequence. Users need only a set of sparsely sampled representative frames for activity recognition. This is especially useful for human activity retrieval from a large database. Our organization of the activity database also results in tremendous reduction of space requirement and significant simplification over other activity representations that usually require an explicit representation of all the activities. Experimental comparison of the three methods shows that the sequential method augmented with velocity data achieves the best results.

In Section 2 we introduce our proposed representation for human activity/action. The theoretical foundations of our approach is discussed in Section 3. In Section 4 we examine this representation in the major requirements of such a digital library, i.e. efficient storage and retrieval. The assessment is performed by developing a multidimensional indexing scheme for activity retrieval and storage. Section 5 discusses the application of our approach and demonstrates the method experimentally.

2. A PROPOSED SPECIFICATION OF HUMAN MOTION USING STATE VECTORS

In this section, we propose a representation of human motion that can be employed to accurately represent elementary motions which can be later composed to form complete human actions. This representation is conceptually similar to the approaches used for modeling human motion [2] [10] [18] [17] [6] [4] [11] [13]. Such representations are based on robot motions [15] [19]. The VRML humanoid specification provides us with a list of body joints and their relations

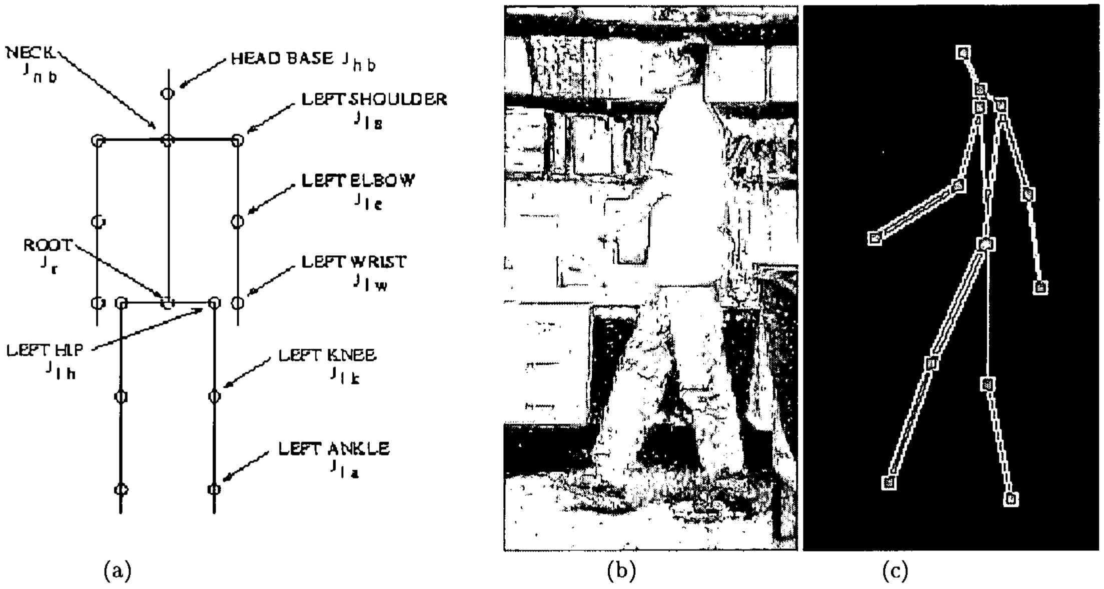


Figure 2: (a) Skeletal structure for humanoid animation.Joints are shown with the corresponding motion vector. (b) Human model in a walk posture. (c) Human Skeletal model for walk posture.

which can be expressed as a body tree. The body tree actually is displayed in Fig. 2(a). The tree nodes are the joints and the arcs are the rigid skeletal connections between the joints.

Since the skeleton is rigid, the human posture is uniquely specified by the angular displacements of all of the joints. Correspondingly, human motion will be uniquely represented by the angular velocities of all the joints[2]. Hence, we propose here to introduce vectorial representation of action/activity which uniquely specifies humanoid/human motion in terms of a sequence of vectors that represent the angular positions and velocities of all the joints. Since human motion may require complex sequence of varying joint angular velocities, we propose here to approximate the angular velocities by elementary motions where each segment has a constant velocity. Thus, each elementary motion may be specified by an angular joint velocity and initial and final angular displacements. A motion sequence is constructed from a temporal sequence of elementary motions. The elementary motion for the entire humanoid is actually determined by the skeletal body tree corresponding to Fig. 2(a). Each node in this skeleton tree has an associated motion sub-vector $\mathbf{J}_n$, which specifies the motion for the corresponding body joint n[3].

$$\mathbf{J}_n = [\alpha_n^i\ \phi_n^i\ \psi_n^i\ \alpha_n^f\ \phi_n^f\ \psi_n^f\ \dot{\alpha}_n\ \dot{\phi}_n\ \dot{\psi}_n]^T \ . \qquad (1)$$

$\mathbf{J}_n$ specifies the 3-D initial $(\alpha_n^i\ \phi_n^i\ \psi_n^i)$ and final $(\alpha_n^f\ \phi_n^f\ \psi_n^f)$ angular displacements and the angular velocities $(\dot{\alpha}_n\ \dot{\phi}_n\ \dot{\psi}_n)$ for the joint n along its three rotational axes. This motion data can be extracted using forward or inverse kinematics [15] [19], and corresponds to the relative motion of that joint. To specify the entire body motion, these $\mathbf{J}_n$ are arranged in a vector of motions specifying the complete set of joint motions. The motion state vector $\mathbf{M}$ is given by

$$\mathbf{M} = [\mathbf{J}_r, \mathbf{J}_{nb}, \cdots \mathbf{J}_{la}] \qquad (2)$$

where the subscripts like r, nb and la are the joint names as shown in Fig. 2(a). For example, la is the LeftAnkle joint.

3. THEORETICAL FOUNDATION OF OUR APPROACH

In this section, we describe the theoretical foundation to our approach in recognizing human activity using indexing. Our representation for human activity in video frames could be described as a concatenation of 18 dimensional sub-vectors $\mathbf{x}_i$ that describe the angles and angular velocities of 9 body parts[4]. Each sub-vector pertains to a video frame and thus the whole video sequence can be represented by a vector $\mathbf{Y}$ which is a concatenation of all the sub-vectors $\mathbf{x}_i$. Please note that in our representation the angles are only 2-D projections of the actual 3-D angles. Hence, our representation is limited to a given view of the activity and so our scheme is view based. However, we find that this representation is not very sensitive to changes in vantage point and the viewing direction can be changed in the range of ±30 degrees without seriously affecting the recognition rate. In the future, we plan to incorporate a method for recovery of the 3-D angles [1] that will enable us to make our recognition method view invariant.

To recognize an activity one has to compare the test video to a model activity. In other words, the test vector $\mathbf{Y}_t$ has to be compared with a set of model vectors $\{\mathbf{Y}_m; m \in [1, M]\}$ where M is the number of activity models in the database. A

[2] Given the translations and rotations of the central skeletal coordinate system as well.

[3] We model all the $\mathbf{J}_n$ with respect to the skeleton central coordinate system. Thus, all translations and rotations are entirely specified by $\mathbf{M}$.

[4] The nine body parts are torso and head, upper arms and legs (thighs) and lower arms (forearm plus hand) and legs (calf plus foot).

similar problem was dealt with using Hidden Markov Models(HMM) [23] [20]. We find that the solution can be significantly simplified if we make some assumptions that will be detailed later. The problem of activity recognition can be formulated as Maximum Likelihood Sequence Estimation(MLSE). The MLSE problem is to determine the most likely sequence $\mathbf{Y_m}$ given the observations $\mathbf{Y_t}$. The Viterbi algorithm [14] provides a computational approach to solve such a problem. We assume that the random differences between the sub-vectors $\mathbf{x_t}$ and $\mathbf{x_m}$ can be described as multivariate zero mean Gaussian distribution. Assuming that these variations are conditionally independent from sample to sample, then the likelihood function for the sequence $P(\mathbf{Y_t}|\mathbf{Y_m})$ can be written as

$$\begin{aligned} P(\mathbf{Y_t}|\mathbf{Y_m}) &= P(\mathbf{x_{t1}}, \mathbf{x_{t2}}, \cdots, \mathbf{x_{tk}}|\mathbf{x_{m1}}, \mathbf{x_{m2}}, \cdots, \mathbf{x_{mk}}) \\ &= \prod_{i=1}^{k} \frac{e^{[\frac{-1}{2}(\mathbf{x_{ti}}-\mathbf{x_{mi}})^T C_x^{-1}(\mathbf{x_{ti}}-\mathbf{x_{mi}})]}}{(2\pi)^{\frac{N}{2}}|C_x|^{\frac{1}{2}}} \end{aligned} \quad (3)$$

where C_x is the covariance matrix of the distribution of the training set for $\mathbf{x_m}$, N is the dimension of the sub-vectors $\mathbf{x_m}$ or $\mathbf{x_t}$(18 in our case) and k is the number of frames in the activity sequence. Using the log-likelihood function we get

$$\log P(\mathbf{Y_t}|\mathbf{Y_m}) = \sum_{i=1}^{k}[\frac{-1}{2}(\mathbf{x_{ti}} - \mathbf{x_{mi}})^T C_x^{-1}(\mathbf{x_{ti}} - \mathbf{x_{mi}})] - kG \quad (4)$$

where G is the logarithm of the denominator in Equation 1 given by

$$G = \log[(2\pi)^{\frac{N}{2}}|C_x|^{\frac{1}{2}}] \quad (5)$$

The most likely activity sequence Ω is found by the maximum likelihood,

$$\Omega = \arg\max_{m}(\sum_{i=1}^{k}[\frac{-1}{2}(\mathbf{x_{ti}} - \mathbf{x_{mi}})^T C_x^{-1}(\mathbf{x_{ti}} - \mathbf{x_{mi}})]) \quad (6)$$

3.1 Foundations of the Voting approach

Finding the most likely activity can now be solved by an indexing based voting approach. In this case, for each test sub-vector $\mathbf{x_{ti}}$ we accumulate votes for all the models. In such voting, a model m will accumulate an incremental vote of

$$\frac{-1}{2}(\mathbf{x_{ti}} - \mathbf{x_{mi}})^T C_x^{-1}(\mathbf{x_{ti}} - \mathbf{x_{mi}}) - G \quad (7)$$

for each test frame i. This process is repeated by voting all the frames i in the test sequence. In our method, we even simplify this voting further by voting only on a few representative frames which are sparsely sampled from the test video sequence. As demonstrated in Section 5 four sparse samples are sufficient to achieve quite robust recognition.

3.2 Dealing with time shifts and activity speed variations

In most test sequences, we encounter the problem that the activity is not synchronized with the model activity. Usually there is a time shift between the two sequences. This time shift denoted by a, is apriori unknown and has to be found along with the activity classification. We solve this problem by combining the votes with temporal correlation.

$$\begin{aligned} \Omega = \arg\max_{m}(\arg\max_{a_m}(\sum_{i=1}^{k}[\frac{-1}{2}(\mathbf{x_{ti}} - \mathbf{x_{m(i-a_m)}})^T \\ C_x^{-1}(\mathbf{x_{ti}} - \mathbf{x_{m(i-a_m)}})])) \end{aligned} \quad (8)$$

where a_m is the time shift between the test sequence and the m^{th} model sequence of the activity. We use this method in Section 4.1 in our temporal correlation scheme.

Another problem that arises in many activities is the problem of speed variations of the activity. The same activity could be performed with different speeds and the speed can even vary during the course of the activity. Variations of speed are actually equivalent to variations in time scale. This problem is quite difficult in general since it requires complex search for the optimum votes with various time scales and time shifts.

$$\begin{aligned} \Omega = \arg\max_{m}(arg\max_{s}(\arg\max_{a_m}(\sum_{i=1}^{k}[\frac{-1}{2}(\mathbf{x_{ti}} - \mathbf{x_{ms(i-a_m)}})^T \\ C_x^{-1}(\mathbf{x_{ti}} - \mathbf{x_{ms(i-a_m)}})]))) \end{aligned} \quad (9)$$

where s denotes the time scale.

In Section 4.2 we propose a method which provides an optimal and robust solution to speed invariant activity recognition. Our solution is based on Sequence matching of the sparse samples. The first underlying principle in the method is that the sequence of the samples of any activity do not change with any variations of speed. This is obvious. Thus, we can reduce the search space by first searching for the optimal vote for the first test frame $\mathbf{x_{t1}}$ and then search for the next optimal vote for the second test frame $\mathbf{x_{t2}}$ only in the reduced set of model frames which occur after the matched model frame with $\mathbf{x_{t1}}$. The same process repeats with the third test frame , the fourth test frame and so on. To avoid the problem that the first test frame is matched with a model frame which occurs towards the end of the sequence we extend all the model sequences by a full additional period.

4. MULTIDIMENSIONAL INDEXING AND VOTING

In this section we describe the three different schemes that we propose for the recognition of human activity. Subsection 4.1 discusses the approach based on temporal correlation. The subsequent two subsections use only the sequence information and disregard the temporal data. The method in subsection 4.3 differs from the one in subsection 4.2 by the additional consideration of the angular velocity of the body parts.

Our activity recognition scheme is based on the fact that the range of poses of different body parts is limited and the activity patterns of the parts are largely repetitive for most of human activities. By exploiting this fact, we can reduce the redundancy in the activity database drastically. This can be achieved by decomposing the activity pattern of the whole body into a combination of activity patterns of individual body parts and storing in the same bins similar postures with different time instants. For example, a person's running activity can be regarded as a combination of the moving sequences of the arms, legs and torso. In addition, the representation is compressed by quantizing all the possible poses of body parts and representing activity patterns by sequences of discrete symbols. In this paper, these symbols are represented by multi-dimensional tuples generated from the poses of the parts. Those are later used as indices of pose hashing tables.

4.1 Indexing with Temporal Correlation

In our approach, we use a human model similar to the one used in [4] with slight variations. The human body is represented by 9 cylinders with elliptic cross-sections for the torso, upper arms and legs (thighs), and lower arms (forearms + hands) and legs (calves + feet). Furthermore, it is assumed that the Cartesian coordinates of all the major joints connecting the above mentioned parts have been obtained using a tracking procedure for body parts [4] [17] [16]. The posture of the whole body any instant is composed of the poses of the arms, legs and torso. To achieve invariance to size, the Cartesian coordinates are transformed into angles. In this method, we use 2D tuples (θ_1, θ_2) to represent the angular poses of arms and legs, where θ_1 denotes the angle between the positive x-axis and the upper arm or the thigh and θ_2 represents the angle between the positive x-axis and the forearm or the calf. For the torso, a single angle θ_3 is used for pose representation, where θ_3 represents the angle between the positive x-axis and the major axis of the torso. All the angles are measured in counter-clockwise direction. We note that the absolute spatial position of the torso in the image does not bear much activity information since it largely depends on the relative position of the imaging system.

The next step is to quantize these multi-dimensional tuples into multidimensional bins to form indices into separate hash tables. In our indexing scheme, we have five hash tables: one (h_1) for the torso, two (h_2 and h_3) for legs and two (h_4 and h_5) for the arms. Depending on the context, $\{h_i; i \in [1,5]\}$ are used to denote both the body part and the corresponding hash table. h_1 has one dimensional bins $b_1 = (b_{11})$ and $\{h_i : i \in [2,5]\}$ have two dimensional bins $b_i = (b_{i1}, b_{i2}), i = 2, \cdots, 5$. Each bin in the hash table contains a pair of values which denote the model number $\{m; m \in [1, M]\}$ and the time instant $\{t; t \in [0, T_m - 1]\}$ of the model activity in the database, where M is the number of activity models in the database and T_m represents the number of image frames for model m. Each hash table is updated using the angular position of the body parts obtained from each activity model. Thus, the poses of body part h_i of model m at instant t are quantized into bin b_i. The complete activity models are scattered into five hash tables (four tables for the limbs and one for the torso). In the hash table, every entry may include a set of different activity models which pertains to the same body part pose. This arrangement of the hash tables is quite efficient for storage and also enables robust recognition. Similar general principles were used in other voting schemes such as the geometric hashing [21], but our method employs several hash tables in parallel.

Our recognition scheme consists of three stages. The first stage involves voting for the individual body parts. The second stage combines the votes of the individual body parts for each test frame. The third stage obtains the final activity vote by integrating the votes of individual test frames based on the temporal information contained in the test sequence.

In the first stage, we decompose the body posture in each frame into angular poses of body parts as described above and index into the hash tables of the corresponding parts. The voting scheme for each part h_i employs M 1D arrays $V^{h_i}_{mk}(t), m \in [1, M]$, where each array corresponds to a different activity model and k is the frame number of the test activity. If we have several items in the table entry that correspond to the same pose index, most likely these items correspond to different activity models and may pertain to different time instants. Thus, the sizes of these 1D arrays should be large enough to accommodate for all time instants (i.e. time bins) of the respective activity models.

In order to tolerate slight pose variations that may occur in the same activity, it is necessary to consider also the neighboring pose bins of the indices derived from the poses of the test activity. Thus, for a given test pose, votes are accumulated from all the neighboring pose bins. The indices for pose bins are two dimensional for the hash tables of the legs and the arms, one dimension relates to the pose of upper parts (upper arms or thighs) and the other dimension denotes the pose of the lower parts (forearms or calves). Let $b^k_i = (q^k_1, q^k_2)$ denote the quantized bin of one of the limbs ($h_i, i \in [2,5]$) for a test pose in test frame k, and let $b'_i = (q'_1, q'_2)$ denote a neighboring bin in the corresponding hash table. We define $f(d,e)$ as a mapping function from a bin's offset d, e to the f range $[0,1]$. Here, we have chosen the mapping function to be a 2D Gaussian,

$$f(d,e) = e^{\frac{-1}{2}[(\frac{d-d_0}{\sigma_d})^2 + (\frac{e-e_0}{\sigma_e})^2]} \qquad (10)$$

where σ_d, σ_e denote the scale of the Gaussian along the respective axes, (d_0, e_0) represent the center of the function. In such a case, a model m with time instant t in the entry $h_i(b'_i), i \in [2,5]$ receives a vote from the test pose according to

$$V^{h_i}_{mk}(t) \mathrel{+}= (f(|q^k_1 - q'_1|, |q^k_2 - q'_2|)) \qquad (11)$$

where += represents incrementing the value of the left-hand side by the value of the right-hand side. $V^{h_i}_{mk}(t)$ is initialized to zero before the voting begins. This voting mechanism is illustrated in Fig. 3. For voting on the poses of the torso, we use

$$V_{mk}^{h_1}(t) \mathrel{+}= f(|q_3^k - q_3'|) \tag{12}$$

where q_3^k denotes the quantized bin of the angular pose of the torso in test frame k, q_3' denotes a neighboring bin and the mapping function f is a 1D Gaussian.

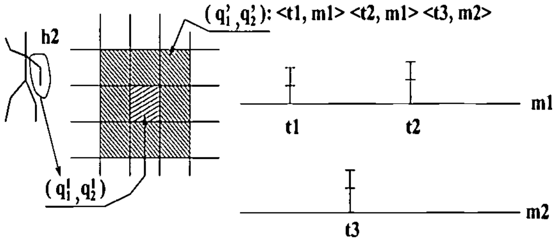


Figure 3: A voting example of the left arm. On the left, the center square (q_1^1, q_2^1) of the grid represents the bin from the pose of the left arm and the surrounding squares are neighboring bins. The upper-left bin (q_1', q_2') contains three entries from models m1 and m2. These votes are described by the bars on the right diagram. This diagram describes two 1D voting arrays for activity models m1 and m2.

In the second stage, the votes that correspond to a particular test image frame k is denoted as $V_{mk}(t)$ and are obtained by combining the votes for the torso and the votes for other body parts. The votes from the limbs and torso are combined by addition. Hence, the votes for a test image frame are given by:

$$V_{mk}(t) = V_{mk}^{h_1}(t) \ + \ \sum_{i=2}^{5} V_{mk}^{h_i}(t) \tag{13}$$

Further, if there are K number of test frames, we need to allocate one set of voting arrays ($V_{mk}^{h_i}$ and V_{mk}) for each image frame k in the test sequence and perform the same procedure as described above to gather votes from all the individual frames. The final result of the first two stages is a set of M 1D voting arrays $\{V_{mk}(t); m = 1, \cdots, M\}$ where m is the model number and k represents a test frame $k = 1, \cdots, K$.

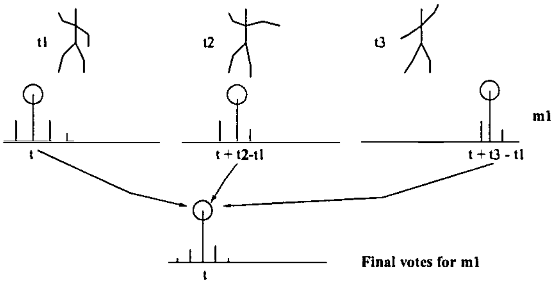


Figure 4: Using temporal correlation for integrating votes from individual test frames which provides a final voting array for one model activity.

After obtaining the votes for all the models from every distinct frame in the test sequence in the second stage, it is necessary to combine the votes from each test frame into a final vote for each activity model. This task is performed by the third stage. Let $t_1, \cdots, t_K$ represent the time instants of the frames in the test sequence, then the final vote for mth model can be obtained by the following discrete correlation

$$V_m(t) = \sum_{k=1}^{K} \sum_{\tau=-a}^{a} g(\tau)\, V_{mk}(\tau + t + t_k - t_1), t \in [0, T_m - 1] \tag{14}$$

where $g(\tau)$ is a symmetric weighting function and $[-a, a]$ is its support. An example of this stage is given in Fig. 4. The idea of temporal correlation is illustrated in Fig. 5.

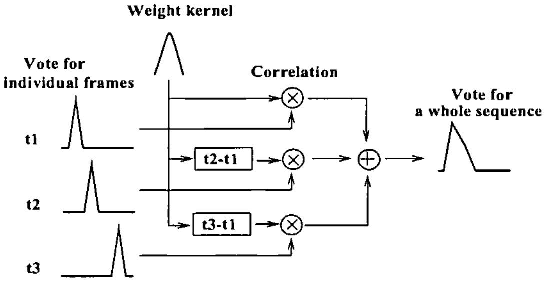


Figure 5: Combining the voting from 3 test frames in the third stage by discrete correlation.

A final scalar vote for each activity model can be obtained by

$$V_m = \max_t V_m(t) \tag{15}$$

and can be used to select one or several models which are the most similar to the test activity. The final selection also yields the exact timing of the matched activity.

4.2 Indexing with Sequence Correlation

The above temporal correlation has the shortcoming of not being able to recognize the same activity when the action is performed at different speeds. For example, fast walking may be recognized as running.

To overcome this problem, we eliminate in the second method the time component and keep only the sequence information when calculating the final vote in the third stage. When a test activity is performed at a different speed from that of the model activity, the time instants for each body pose are different. Hence the temporal correlation approach may not yield a strong response for the recognized activity which may lead to false alarms in the recognition system. In the second approach, we attempt to create speed-invariant activity recognition by eliminating the temporal information and replacing it by the sequence of the activity.

This method uses the same indexing scheme discussed above. The difference arises in the third stage. Here, we extend the vote table obtained in the second stage so that it is equivalent to the activity with two cycles instead of one. In the third stage, the final vote for the mth model can be obtained by the following equation

$$V_m = \sum_{k=1}^{K} V_{m,k}(L_k) \tag{16}$$

In the above equation the following conditions have to be satisfied: $L_i < L_j$; $i < j$ and $V_{m,k}(L_k)$ is the maximum vote for the activity m.

4.3 Sequencing with Angular Velocity

We develop this method to improve the discrimination between activities as compared to the method in 2.2. In this method, the multidimensional hash table is extended to 4 dimensions instead of two for the limbs and 2 dimensions instead of one for the torso. The additional two dimensions for the limbs are used to represent their angular velocities. We use 4D tuples $(\theta_1, \theta_2, \dot{\theta}_1\, \dot{\theta}_2)$, where θ_1 and θ_2 are same as for method one, $\dot{\theta}_1$ denotes the angular velocity of the upper arm or thigh and $\dot{\theta}_2$ denotes the angular velocity of the forearm or the calf. For the torso, the 2D tuples are, $(\theta_3, \dot{\theta}_3)$, where θ_3 is the same as method one and $\dot{\theta}_3$ is the angular velocity of the torso. The angular velocities are calculated as the difference of the angular positions of two successive frames. For the last frame, the angular velocity of the last but one frame is retained. The angular velocities are then quantized into 5 bins.

In this new indexing scheme, the four hash tables for the limbs are now indexed using four dimensions and the hash table for the torso is indexed using two dimensions. The bins in each table contain a pair of model number and frame number as before.

The first stage of voting is similar as in the previous methods where the body posture of each frame is decomposed into the poses of the five body parts and indexed in to the hash tables of the corresponding parts. In order to tolerate slight pose variations that may occur in the same activity, it is necessary to consider also the neighboring pose bins of the indices derived from the poses of the test activity. Let $b_i^k = (q_1^k, q_2^k, q_3^k, q_4^k)$ denote the quantized bin of one of the limbs ($h_i, i \in [2,5]$) for a test pose in test frame k, and let $b_i' = (q_1', q_2', q_3', q_4')$ denote a neighboring bin in the corresponding hash table. We define $f(b,c,d,e)$ as a mapping function from a bin's offset d, e to the f range $[0,1]$. Here, we choose this mapping function to be a 4D Gaussian,

$$f(b,c,d,e) = e^{\frac{-1}{2}[(\frac{b-b_0}{\sigma_b})^2+(\frac{c-c_0}{\sigma_c})^2+(\frac{d-d_0}{\sigma_d})^2+(\frac{e-e_0}{\sigma_e})^2]} \qquad (17)$$

where $\sigma_b, \sigma_c, \sigma d, \sigma e$ denote the scale of the Gaussian along the respective axes, (b_0, c_0, d_0, e_0) represent the center of the function. In such a case, a model m with time instant t in the entry $h_i(b_i'), i \in [2,5]$ receives a vote from the test pose according to

$$V_{mk}^{h_i}(t) += \alpha(f(|q_1^k - q_1'|, |q_2^k - q_2'|, |q_3^k - q_3'|, |q_4^k - q_4'|)) \qquad (18)$$

where += represents incrementing the value of the left-hand side by the value of the right-hand side. $\alpha = 1, if\ |q_3^k - q_3'| = 0\ and\ |q_4^k - q_4'| = 0\ \alpha = 0.5,\ if\ |q_3^k - q_3'| = 1\ and\ |q_4^k - q_4'| = 0$ or $if\ |q_3^k - q_3'| = 0\ and\ |q_4^k - q_4'| = 1$

$V_{mk}^{h_i}(t)$ is initialized to zero before the voting begins. This voting mechanism is illustrated in Fig. 3. For voting on the poses of the torso, we use

$$V_{mk}^{h_1}(t) += \alpha\ f(|q_5^k - q_5'|, |q_6^k - q_6'|) \qquad (19)$$

where q_5^k denotes the quantized bin of the angular pose of the torso in test frame k, q_6^k denotes the quantized bin of the angular velocity of the torso in test frame k, q_5' , q_6' denote a neighboring bin and the mapping function f is a 2D Gaussian. $\alpha = 1, if|q_6^k - q_6'| = 0\ \alpha = 0, if|q_6^k - q_6'| \neq 0$

The second and third stages are identical to the corresponding stages in method 2.

5. EXPERIMENTAL RESULTS

We apply our three methods to a database of eight different human activities. These activities are jumping, kneeling, picking, putting, running, sitting, standing and walking. A total of 26 activity sequences are stored in the database. For each activity we have three or four different sequences performed by different persons. Fig. 6 and Fig. 7 shows a few sample frames of such activity sequences.

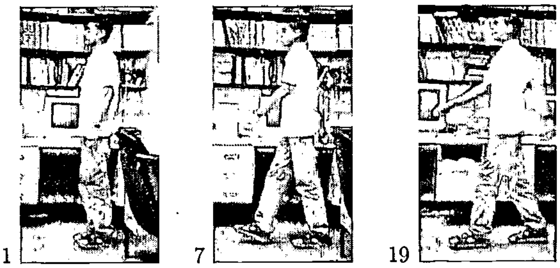


Figure 6: 3 frames of the 'walking' activity. The numbers on the lower-left corners indicate frame numbers.

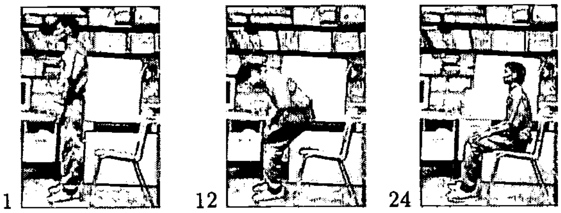


Figure 7: 3 frames of the 'sitting' activity. The numbers on the lower-left corners indicate frame numbers.

The test sequences are generated by taking three or four frames from the model activity sequences sampled at uneven time intervals and adding random perturbations to the positions of the body parts in the frames. We generate four test sequences for each kind of activity. These test sequences are matched against all the activity sequences in the database except the one from which the test sequence is extracted.

Table 1 displays the average votes and the standard deviation for each possible activity pairs for the first method.

This method may discriminate between activities performed at different speeds because the time instants for each body pose are different. This may result in an increase in the false alarm rate. We improve this method by eliminating the time factor and considering only the sequence information and hence make it invariant to speed. The results for this sequence based voting are shown in Table 2. We observe that this method helps in recognizing activities from the database that are performed at different speeds.

In order to discriminate between the activities with more accuracy we introduce the angular velocity of the body parts. The results of this method are shown in Table 3. It is observed that the discrimination between the activities improves. The drawback of this method is that the recognition is slower than the other two methods due to the increased size of the hash tables.

	Jump	Kneel	Pick	Put	Run	Sit	Stand	Walk
Jump	**12.35** (1.26)	8.24 (1.30)	6.51 (0.89)	6.06 (0.74)	6.00 (1.00)	5.85 (1.19)	4.60 (1.07)	7.79 (1.00)
Kneel	9.67 (0.97)	**13.3** (0.60)	7.32 (1.85)	5.91 (0.35)	5.78 (0.74)	7.16 (1.00)	6.02 (1.44)	8.72 (0.70)
Pick	3.26 (0.89)	4.90 (0.58)	**9.57** (0.24)	6.23 (0.38)	3.64 (0.92)	6.14 (0.67)	8.19 (0.40)	4.43 (0.46)
Put	5.57 (0.64)	5.77 (0.62)	8.15 (1.50)	**13.0** (1.04)	6.06 (1.25)	7.56 (1.11)	6.27 (0.90)	6.13 (0.81)
Run	5.58 (0.74)	5.52 (0.91)	5.98 (1.12)	6.44 (1.73)	**7.46** (1.28)	4.97 (0.98)	5.02 (1.42)	6.70 (1.37)
Sit	3.97 (1.32)	5.14 (1.35)	7.58 (1.85)	7.82 (1.22)	4.76 (1.68)	**12.8** (0.86)	9.11 (1.46)	5.69 (1.55)
Stand	1.12 (0.47)	1.66 (0.48)	3.56 (0.93)	5.87 (0.94)	3.67 (0.98)	7.48 (1.06)	**12.3** (0.26)	2.19 (0.59)
Walk	7.64 (0.56)	7.13 (0.67)	6.31 (1.56)	5.09 (1.12)	6.60 (1.55)	6.30 (1.33)	6.00 (1.18)	**9.60** (1.96)

Table 1: Average votes (Standard deviation) of activity sequences for the temporal correlation based voting. The rows correspond to test activity while the columns correspond to the model activities. The best score in each row is in boldface numerals. The method yields correct recognition since the scores along the diagonal are the highest in each row.

	Jump	Kneel	Pick	Put	Run	Sit	Stand	Walk
Jump	**15.37** (0.71)	7.00 (0.88)	6.00 (0.42)	6.38 (0.24)	5.38 (1.12)	6.54 (0.42)	3.90 (0.84)	8.10 (1.41)
Kneel	10.9 (2.42)	**16.0** (0.81)	7.60 (2.47)	6.91 (0.41)	5.80 (2.07)	9.60 (0.97)	5.63 (1.63)	9.58 (1.73)
Pick	3.45 (2.12)	4.77 (2.10)	**11.7** (0.57)	8.00 (2.30)	5.14 (2.30)	6.00 (2.10)	8.35 (1.15)	5.31 (2.31)
Put	5.30 (1.34)	5.57 (1.57)	10.3 (2.24)	**15.0** (1.57)	6.19 (2.21)	8.90 (2.19)	6.80 (1.73)	7.75 (1.31)
Run	4.06 (0.71)	4.25 (0.73)	3.96 (1.48)	5.20 (1.68)	**7.21** (1.89)	4.60 (1.26)	4.85 (1.62)	5.84 (1.41)
Sit	4.41 (1.00)	5.01 (1.16)	7.97 (1.99)	8.04 (2.14)	6.22 (2.45)	**15.0** (0.71)	10.8 (1.61)	6.28 (1.70)
Stand	2.42 (3.10)	3.08 (2.24)	6.16 (3.16)	6.75 (2.50)	5.51 (2.00)	10.1 (1.41)	**15.1** (0.63)	4.15 (4.00)
Walk	7.60 (2.17)	7.66 (1.14)	4.80 (0.88)	5.82 (0.31)	7.27 (2.00)	7.47 (1.40)	5.00 (1.67)	**12.3** (1.18)

Table 2: Average votes (Standard deviation) of activity sequences for the sequence based voting. The rows correspond to test activity while the columns correspond to the model activities. The best score in each row is in boldface numerals. The method yields correct recognition since the scores along the diagonal are the highest in each row.

	Jump	Kneel	Pick	Put	Run	Sit	Stand	Walk
Jump	**12.31** (1.21)	3.91 (0.51)	1.97 (0.54)	2.00 (0.45)	2.18 (0.60)	2.00 (0.55)	1.20 (0.95)	3.55 (0.78)
Kneel	4.90 (2.16)	**9.99** (2.18)	3.20 (0.88)	2.77 (0.63)	2.20 (0.77)	2.18 (0.71)	2.40 (1.58)	3.80 (0.92)
Pick	0.67 (0.72)	2.00 (0.60)	**8.00** (0.71)	2.40 (0.85)	1.97 (1.00)	1.90 (0.86)	3.80 (0.96)	1.36 (0.95)
Put	1.95 (0.73)	2.58 (0.71)	3.10 (0.51)	**8.37** (2.95)	1.58 (1.13)	4.71 (1.00)	2.50 (0.81)	1.74 (0.90)
Run	2.00 (0.75)	2.25 (0.83)	1.40 (0.37)	1.70 (0.32)	**3.23** (1.28)	1.40 (0.28)	1.50 (0.39)	2.90 (0.94)
Sit	1.36 (0.47)	1.73 (0.45)	3.00 (0.67)	4.20 (0.94)	0.90 (0.95)	**8.60** (0.83)	3.40 (1.48)	1.60 (0.63)
Stand	0.00 (0.00)	0.55 (1.04)	2.34 (1.83)	1.86 (1.00)	1.23 (1.17)	3.50 (0.67)	**9.90** (0.37)	0.63 (1.11)
Walk	3.40 (0.99)	3.18 (1.07)	1.97 (0.87)	1.60 (1.05)	2.61 (1.09)	1.50 (0.44)	1.16 (0.69)	**5.75** (2.20)

Table 3: Average votes (Standard deviation) of activity sequences for the sequence with angular velocity based voting. The rows correspond to test activity while the columns correspond to the model activities. The best score in each row is in boldface numerals. The method yields correct recognition since the scores along the diagonal are the highest in each row.

6. CONCLUSIONS

In this paper, we propose and evaluate a representation for human activity/action that is based on sequences of angular poses/velocities of the human skeletal joints. The evaluation is implemented by developing a multidimensional indexing scheme for activity retrieval and storage. For this purpose, we develop three different approaches to human activity recognition/retrieval which are based on this representation. The sequence based voting approach in the second and third methods, is introduced since the temporal correlation approach in the first method , is not invariant to speed and incorrectly recognizes running activity as walking in one case. The second method solves this problem, but at the expense of incorrectly recognizing the stand activity as sitting in one case. This happens because we take two cycles during the voting process. These two activities differ only in the sequence with which they occur. Hence the misclassification. When we introduce the angular velocity in the third method this misclassification is no longer present and in fact it gives better discrimination between the activities as it is expected, due to the increased dimensionality. To evaluate the effectiveness of the methods quantitatively we define the Average Discrimination Ratio (ADR) as the average of the ratios of the first maximum vote to the second maximum vote for each activity. The ADR for the three methods are 1.38, 1.49 and 2.15 respectively. This shows that the third method has the best discrimination power.

In summation, we propose here a representation for human action/activity which can describe accurately any complex human activity/action and develop a robust method for activity recognition/retrieval. The indexing approach also provides an efficient storage/retrieval of all the activities in a small set of hash tables.

7. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) Grants No. IIS-9711925, IIS-9876904 and IIS-9979774.

8. REFERENCES

[1] C. Barron and I. A. Kakadiaris. Estimation anthropometry and pose from a single image. *Proc. Conf. Computer Vision and Pattern Recognition*, pages 669–676, June 2000.

[2] C. Bregler and J. Mallik. Tracking people with twists and exponential maps. *Proc. IEEE 1998 Int'l Conf. Computer Vision and Pattern Recognition (CVPR'98)*, pages 8–15, June 1998.

[3] S. K. Chang, Q. Y. Shi, and C. W. Yan. Iconic indexing by 2-d strings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(3):413–427, July 1987.

[4] L. Concalves, E. Bernardo, E. Ursella, and P. Perona. Monocular tracking of human arm in 3d. *Proc. 1995 International Conference on Computer Vision*, 1995.

[5] A. Del-Bimbo, E. Vicario, and D. Zingoni. Symbolic description and visual querying of image sequences using spatiotemporal logic. *IEEE Transactions on Knowledge and Data Engineering*, 7(4):609–622, August 1995.

[6] D. Gavrila and L. Davis. Towards 3-d model-based tracking and recognition of human movements. *Proc. of the 1995 Int. Workshop on Automatic Face and Gesture Recognition*, 1995.

[7] A. H. Guest. Dance notation: The process of recording movement on paper. *London:Dance Books*, 1984.

[8] A. H. Guest. Chore-graphics: A comparison of dance notation systems from the fifteenth century to the present. *Systems from the Fifteenth Century to the Present, Gordon and Breach Science Publishers S. A.*, 1989.

[9] D. Herbison-Evans. Dance, video, notation and computers. *Leonardo*, 1988.

[10] D. Hogg. A program to see a walking person. *Image and Vision Computing*, 5(20), 1983.

[11] S. JU, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. *2nd Int. Conf. on Automatic Face and Gesture Recognition*, pages 38–44, 1996.

[12] E. Jungert. The observer's point of view, an extension of sympolic projections. *Prof.In. Conf. of Theories and Methods of Spatio-Teporal Reasoning in Geopraphic Space*, pages 179–195, September 1992.

[13] I. A. Kakadiaris and D. Metaxas. Model-based estimation of 3d human motion with occlusion based on active multi-viewpoint selection. *Proc. 1996 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'96)*, 1996.

[14] T. K. Moon and W. C. Stirling. *Mathematical Methods and Algorithms for Signal Processing.* Prentice Hall Inc., 2000.

[15] R. M. Murray, Z. Li, and S. S. Sastry. *A Mathematical Approach to Robotic Manipulation.* CRC Press Inc., 1994.

[16] T. M. Naoyuki Sawasaki and T. Uchiyama. Design and implementation of high-speed visual tracking system for real-time motion analysis. *Proc. of ICPR 1996*, pages 478–484, 1996.

[17] J. Regh and T. Kanade. Model-based tracking of self-occluding articulated objects. *Proc. 1995 International Conference on Computer Vision*, 1995.

[18] K. Rohr. Incremental recognition of pedestrians from image sequences. *Proc. 1995 Comp. Soc. Conference on Computer Vision and Pattern Recognition*, pages 8–13, June 1993.

[19] R. J. Schilling. *Fundamentals of Robotic Analysis & Control.* Prentice Hall Inc., 1990.

[20] J. Schlenzig, E. Hunter, and R. Jain. Vision based hand gesture interpretation using recursive estimation. *Proceedings of the 28th Asilmoar Conference on Signals, Systems and Computers*, 1994.

[21] J. Schwartz, Y. Lamdan, and H. Wolfson. Geometric hashing : A general and efficient model-based recognition scheme. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 335–344, 1988.

[22] A. P. Sistla and O. Wolfson. Temporal triggers in active database systems. *IEEE Transactions on Knowledge and Data Engineering*, 7(3), June 1995.

[23] T. Starner and A. Pentland. Real time american sign language recognition from video using hidden markov models. *Proceedings of the International Symposium on Computer Vision*, 1996.

[24] A. D. Wilson, A. F. Bobick, and J. Cassell. Temporal classification of natural gesture and application to video coding. *IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recogni tion*, pages 948–854, June 1997.

A Bucket Architecture for the Open Video Project

Michael L. Nelson
NASA Langley Research Center
MS 158
Hampton, VA 23681
+1 757 864 8511
m.l.nelson@larc.nasa.gov

Gary Marchionini, Gary Geisler, Meng Yang
University of North Carolina
School of Information and Library Science
Chapel Hill, NC 27599
+1 919 966 3611
{march, geisg, yangm}@ils.unc.edu

ABSTRACT

The Open Video project is a collection of public domain digital video available for research and other purposes. The Open Video collection currently consists of approximately 350 video segments, ranging in duration from 10 seconds to 1 hour. Rapid growth for the collection is planned through agreements with other video repository projects and provision for user contribution of video. To handle the increased accession, we are experimenting with "buckets," aggregative intelligent publishing constructs for use in digital libraries.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: - *collection, dissemination, systems issues.*

General Terms

Management, Documentation, Design, Experimentation.

Keywords

Digital Video, Digital Objects, Open Source, Buckets.

1. INTRODUCTION

The Open Video project provides a World Wide Web (WWW) accessible standard corpus of public domain video segments suitable for use by the research and education communities [1]. The Interaction Design Laboratory at UNC is the first research group to use this collection of video for research in alternative searching and browsing interfaces for digital video. Other prominent digital video projects, such as Carnegie Mellon University's Informedia Project [2], are not publicly available due to copyright limitations. Another digital video collection, the Internet Moving Images Archive [3], but only some is available at the time of this writing and it is only in MPEG-2 format.

In addition, we are experimenting with storing the video in "buckets": aggregative, intelligent publishing constructs [4]. Buckets are aggregative in that they combine data, metadata and the methods defined on them. This is designed to prevent the information "drift" we have observed in digital libraries (DLs) when there is a multiplicity of data and metadata types and formats which become "unlinked" and "lost" over time. Buckets are intelligent in that they are entirely self-contained and self-sufficient. They do not depend on the existence (or absence) of any particular DL, repository, database or search engine. As a result, they manage their own contents, enforce their own terms and conditions, and internally transport their source code.

2. ORIGINAL IMPLEMENTATION

The original architecture for the Open Video project was centered around the relational database management system (RDBMS) that maintained the descriptive and structural metadata. The video segments were only in MPEG-1 format, and some of the video segments also had textual output from the "mpeg_stat" program associated with them. The MPEGs and output files were stored as files on a Unix filesystem, and the RDBMS maintained the URLs to these files in its metadata.

However, as the collection grew this approach would have became increasingly unwieldy. Open Video is planning to add multiple video encodings and formats: MPEG-2, AVI, QuickTime, and possibly others. Similarly, we are experimenting with alternate methods for browsing video segments, including overviews, previews, and AgileViews [5]. We are also using software to extract the keyframes from a video and store them as JPEGs, GIFs or PPMs. In short, each logical video segment represents an increasing number of physical data objects, many of them derived from the original segment, and some of them transient and experimental. Rather than expose this level of complexity to the RDBMS, we encapsulate the varying and dynamic nature of video segments into buckets.

3. BUCKET ARCHITECTURE

Buckets provide an aggregative container mechanism to hold all data items that comprise the logical unit of a video segment. Buckets are currently implemented as individual Perl CGI scripts, and their API is accessible through http encoded messages. Although bucket tools make use of the API, buckets appear as normal web sites to the casual user.

There is a subtle shift in responsibility in the bucket architecture. Where the RDBMS used to contain the canonical metadata for the collection, the buckets are now canonical. The metadata is still stored in the RDBMS, but it is harvested from the buckets themselves. We believe this makes it easier for the video segment buckets to be included in other DLs, since all data and metadata content is available directly from the buckets.

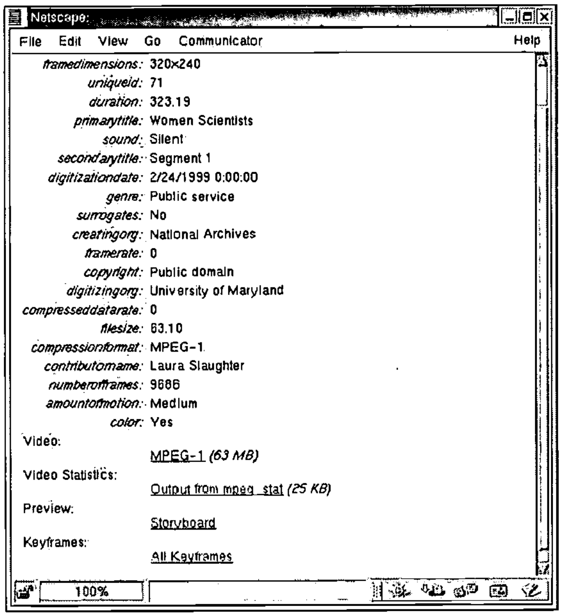
An example of a video bucket can be seen in Figure 1. This is the result of direct access to the URL:

http://buckets.dsi.internet2.edu/openvideo/buckets/ov-71/

which is the same as invoking the default "display" method:

http://buckets.dsi.internet2.edu/openvideo/buckets/ov-71/
?method=display

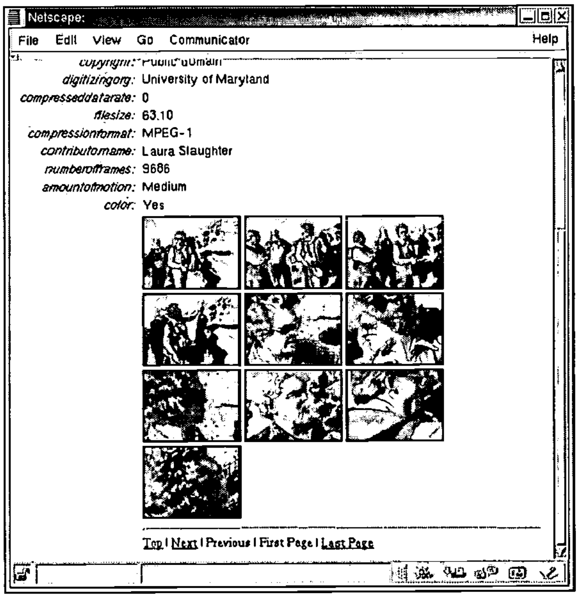
JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.



framedimensions: 320x240
uniqueid: 71
duration: 323.19
primarytitle: Women Scientists
sound: Silent
secondarytitle: Segment 1
digitizationdate: 2/24/1999 0:00:00
genre: Public service
surrogates: No
creatingorg: National Archives
framerate: 0
copyright: Public domain
digitizingorg: University of Maryland
compresseddatarate: 0
filesize: 63.10
compressionformat: MPEG-1
contributorname: Laura Slaughter
numberofframes: 9686
amountofmotion: Medium
color: Yes

Video: MPEG-1 *(63 MB)*
Video Statistics: Output from mpeg_stat *(25 KB)*
Preview: Storyboard
Keyframes: All Keyframes

Figure 1. Default Display Method for a Bucket.



copyright: Public domain
digitizingorg: University of Maryland
compresseddatarate: 0
filesize: 63.10
compressionformat: MPEG-1
contributorname: Laura Slaughter
numberofframes: 9686
amountofmotion: Medium
color: Yes

Top | Next | Previous | First Page | Last Page

Figure 2. Video Storyboard With Selected Keyframes.

Figure 1 shows all collected information about this one video segment, including the MPEG-1, the textual analysis of the MPEG-1 file, all keyframes, and a video storyboard of 20 keyframes, with pagination control (see Figure 2).

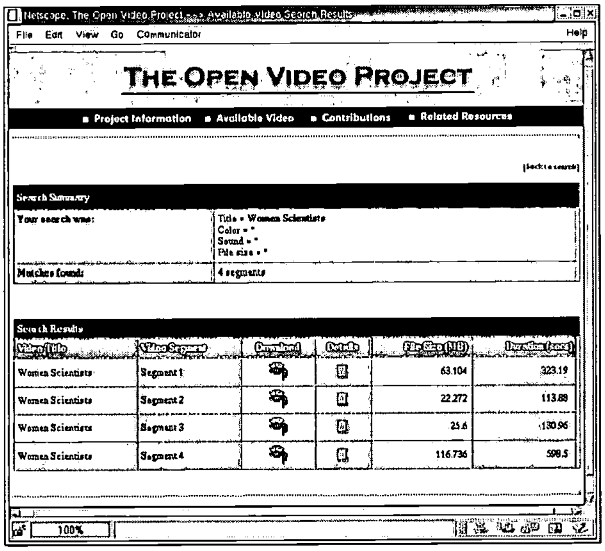
Direct use of buckets exposes the user to the granularity of storage. However, this is not strictly required. It is possible for the interface to present a horizontal slice of the video contents. In Figure 3, the user has requested all video segments that comprise an entire video. Since the MPEGs are individually accessible, each video icon in Figure 3 can link to URLs similar to:

http://buckets.dsi.internet2.edu/openvideo/buckets/ov-71/?method=display&pkg_name=video.pkg&element_name=video.mpg

Thus, the presentation granularity (e.g., all MPEG-1 files that comprise an entire video) does not have to be equivalent to the storage granularity (e.g., all files that comprise or are derived from a single video segment).

4. CONCLUSIONS

We have built a test implementation of the Open Video collection using buckets and are evaluating its performance in anticipation of adopting this architecture as the project evolves. Successful deployment of buckets will not impact the search, browse and preview experience of Open Video users, but should allow for greater flexibility in sharing the Open Video buckets with other projects and long-term management of the project's physical resources. Buckets provide a significant shift from RDBMS as canonical to stored objects as canonical in a DL. This shields the RDBMS (and DL as a whole) from the fluid nature of individual file formats and derived data types.



Search Summary

Your search was:	Title = Women Scientists
	Color = *
	Sound = *
	File size = *

Matches found:	4 segments

Search Results

Video Title	Video Segment	Download	Details	File Size (MB)	Duration (secs)
Women Scientists	Segment 1			63.104	323.19
Women Scientists	Segment 2			22.272	113.88
Women Scientists	Segment 3			25.6	130.96
Women Scientists	Segment 4			116.736	598.5

Figure 3. The Open Video Search Results Page.

5. REFERENCES

[1]Slaughter, L., Marchionini, G. and Geisler, G. Open Video: A framework for a test collection. *Journal of Network and Computer Applications*, 23(3): 219-245, July 2000.

[2]Informedia Digital Video Library, http://www.informedia.cs.cmu.edu/

[3]Internet Moving Images Archive, http://www.archive.org/movies/

[4]Nelson, M. L. Buckets: Smart objects for digtial libaries. Ph.D. Dissertation, Old Dominion University, 2000. http://home.larc.nasa.gov/~mln/phd/

[5]Marchionini, G., Geisler, G. and Brunk, B. AgileViews: A human-centered framework for interfaces to information spaces. In *Proceedings of ASIS 2000* (Chicago IL, November 2000), pp. 271-280.

The Físchlár Digital Video System: A Digital Library of Broadcast TV Programmes

A. F. Smeaton, N. Murphy, N. E. O'Connor, S. Marlow,
H. Lee, K. McDonald, P. Browne, and J. Ye
Centre for Digital Video Processing
Dublin City University, Glasnevin, Dublin 9, Ireland
+353 – 1 - 7005262
Alan.Smeaton@dcu.ie

ABSTRACT

Físchlár is a system for recording, indexing, browsing and playback of broadcast TV programmes which has been operational on our University campus for almost 18 months. In this paper we give a brief overview of how the system operates, how TV programmes are organised for browse/playback and a short report on the system usage by over 900 users in our University.

1. INTRODUCTION

The Físchlár digital video system is a web-based system for recording, analysis, browsing and playback of TV programmes which is used within the campus environment at Dublin City University. It allows users to initiate the recording of programmes from any of the 8 terrestrial TV stations for our area. Once digitised, programmes are analysed for shot boundaries and shot-based representative frames are selected [1]. Shots are then clustered into scenes using a variety of techniques. Físchlár is accessed through a conventional web browser on a desktop machine although we have developed a WAP interface to allow users to reserve programme recordings through a mobile phone and we developed a version of the browser/player for a PDA in a mobile environment. The Físchlár system is described in [2].

In its current configuration, Físchlár allows users to record, play (stream) and browse programmes, and it is the way we allow browsing of TV programme content which is one of the reasons that Físchlár is novel. Commercially available TV recording devices such as TiVo [3] allow recording and playback of broadcast TV content but the browsing function is little more than fast forward and rewind. In Físchlár we have developed 8 different browser interfaces, described and evaluated elsewhere [4], each of which is tailored to a user's task, context and preferences. For example, there are different keyframe browsers for users depending upon whether they have seen the particular programme being browsed before and are interested in locating a particular scene they know, or they are watching for the first time. There are browsers for users who prefer linear vs. structured browsing or for users who prefer a static vs. a dynamic interaction style. Altogether this means that there is sound support for a user's specific task and the context for their need.

Another aspect of Físchlár which is novel, and which is of interest here, is that it has been integrated with a large-scale TV recommender system called PTV [5] and in the remainder of this paper we give a brief overview of what that provides and how it is being used.

2. THE COMBINED FÍSCHLÁR-PTV SYSTEM

Físchlár has been in continuous use for almost 18 months and during that time we have developed, refined and enhanced its functionality. Digital video, in MPEG-1 format, is stored on a SUN Enterprise video server and our archive has over 300 hours of TV (about 400 broadcast programmes) content at any one time. Our server is capable of streaming to over 200 clients concurrently via a web browser plug-in and the system is available from student residences, undergraduate and postgraduate laboratories, and from the main library on campus.

All users must register before using Físchlár and through a logging on process, we are able to track usage and offer personalised services. This includes remembering the user's favoured browser interface as well as programmes. Users use the system mostly for entertainment but also for study-related activities such as browsing/playback of news or specialist programmes (broadcast documentaries, etc.).

Físchlár has two modes of operation, one for recording and another for browse/playback. In recording mode, users are presented with TV listings for the 8 major terrestrial TV channels within our area, for today and for tomorrow. We provide, for each programme, some text details on what the programme is about, who it stars etc., taken from an online entertainment guide. Programmes are also automatically assigned to one or more of a dozen genres including sport, documentary, soap, movies, music, kids, home and garden, etc. Users can view the programme listings by TV station, or across the broadcast channels by genre.

Each Físchlár user is also an indirect user of the PTV system. PTV generates recommendations of TV programmes for users to watch based on their past preferences (positive and negative), the past preferences of others who share or have differing preferences, and the descriptions of new, unviewed programmes. PTV uses

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

case based reasoning as part of its underlying processing and this is described in [5].

A transparent link between Físchlár and PTV allows each Físchlár user's TV viewing recommendations, from PTV, to be presented alongside the TV listings by channel and by genre. In this way we provide not only the standard and genre-organised TV listings for the next 2 days but also a personalised view of what programmes PTV thinks a user should explicitly request recording of. The function of the recording mode in Físchlár is to have the user explicitly select TV programmes which s/he wants to be recorded and to invite the user to grade each programme on a scale of 1 to 5 in terms of their interest in viewing it. These ratings are then passed back to PTV leading to higher quality recommendations. The alternative to having users explicitly request program recording is recording 24/7 on all 8 TV channels but this would allow us to maintain an archive of only the recently broadcast programmes. We feel that users may prefer to access not just materials from within the last week but further back in time, and selective recording rather than taking a shotgun approach, allows us to do just that.

In browse/playback mode, each user is presented with the full library of recorded programmes among which to browse, as well as our within-programme browsing facilities based on keyframe navigation. Recorded programmes (at any one time over 300 hours) can be viewed by TV station, by genre, or by examining recordings from the most recent 7 days only. In addition, we use the connection with PTV to allow personalised recommendations of programmes from the archive to be viewed as well as a category called "favourites", corresponding to subsequent episodes of a user's previous viewings. From a user's perspective this means that when using Físchlár, a user is presented with a TV schedule for the next 2 days from which s/he can request specific programmes to be recorded, with personalised recommendations built in, and a user can browse an archive or library of already broadcast and recorded programmes, again with personalised recommendations built in.

On selecting a specific programme, a user is immediately presented with the set of keyframes drawn from that program which can be a large number of images. For example, a recent 25 minute episode of "The Simpsons" generated 313 keyframes, a 50 minute episode of "Little House on the Prairie" generated 326 and the movie "Crimson Tide", which is 2 hours and 5 minutes, generated 1755 keyframes. Our different browser interfaces described in [4] are used to provide efficient navigation through these keyframes. As a user is browsing the keyframes, he/she can switch to streaming the video of that programme from that keyframe onwards, by clicking the keyframe.

3. USAGE OF THE COMBINED FÍSCHLÁR-PTV SYSTEM

At the time of writing there are over 900 registered users of the Físchlár system. Some of these users are using old PCs in residences, donated by the University to the project and others are using their own desktop machines from within the University intranet. Almost 3000 recording requests have been received in the last 12 months, with 1034 of these requests by 105 users in the last 2 months alone. In fact with our video server limited to storing 300 hours only, we have to remove programmes older than about 1 month in order to provide space for incoming material. The programmes most frequently recorded are "The Simpsons" and "Friends". Other popular programmes are Star Trek (SciFi), Top of the Pops (music), Coronation Street (soap) and 100 years (documentary).

Almost 30% of our users have logged into the system 5 times or more but this statistic is a bit misleading since a single login persists over the whole of a browser's session, so if a user "logs on" to the system then that session remains until the browser application on the PC is shut down. Feedback from users has been hugely positive, especially from those using it in residences. Using the system from labs is less comfortable for users and has been likened to watching TV in public.

In coupling Físchlár with the PTV system we have extended the usefulness of both systems and the combined system presents the user with personalised access to a library of digital video materials. We will shortly introduce other functionality to the system such as text-searching and content-based alerting based on teletext capture. One application which has been requested by staff and students is what we call "buddy clipping", the ability to scope out and define a clip of video from the library whose address can be emailed as an embedded link to others, with text annotation. We have also extended the Físchlár interface to operate on a Compaq iPAQ, a mobile PDA which accesses the system over a wireless LAN.

4. REFERENCES

[1] Evaluating and Combining Digital Video Shot Boundary Detection Algorithms. Browne P, Smeaton A, Murphy N, O'Connor N, Marlow S and Berrut C. In Proceedings of IMVIP 2000, Belfast, Northern Ireland, September 2000.

[2] O'Connor, N., Marlow, S., Murphy, N., Smeaton, A., Browne, P., Deasy, S., Lee, H. and Mc Donald, K. Físchlár: an On-line System for Indexing and Browsing of Broadcast Television Content. In Proceedings of *ICASSP 2001* (Salt Lake City, UT, May, 2001).

[3] TiVo. http://www.tivo.com.

[4] User Interface Design for Keyframe-Based Content Browsing of Digital Video. Lee., H. PhD thesis, Dublin City University, 2001.

[5] Smyth, B. and Cotter, P. A Personalized Television Listings Service. *Communications of the ACM*, **43**(8), 2000, 107-111.

Design Principles for the Information Architecture of a SMET Education Digital Library

Andy Dong
Department of Mechanical Engineering
University of California, Berkeley
5138 Etcheverry Hall
Berkeley, CA 94720-1740
+1 510 643 1819
adong@me.berkeley.edu

Alice M Agogino
Department of Mechanical Engineering
University of California, Berkeley
5136 Etcheverry Hall
Berkeley, CA 94720-1740
+1 510 642 6450
aagogino@me.berkeley.edu

ABSTRACT

This implementation paper introduces principles for the information architecture of an educational digital library, principles that address the distinction between designing digital libraries for education and designing digital libraries for information retrieval in general. Design is a key element of any successful product. Good designers and their designs put technology into the hands of the user, making the product's focus comprehensible and tangible through design. As straightforward as this may appear, the design of learning technologies is often masked by the enabling technology. In fact, they often lack an explicitly stated instructional design methodology. While the technologies are important hurdles to overcome, we advocate learning systems that empower education-driven experiences rather than technology-driven experiences. This work describes a concept for a digital library for science, mathematics, engineering and technology education (SMETE), a library with an information architecture designed to meet learners' and educators' needs. Utilizing a constructivist model of learning, the authors present practical approaches to implementing the information architecture and its technology underpinnings. The authors propose the specifications for the information architecture and a visual design of a digital library for communicating learning to the audience. The design methodology indicates that a scenario-driven design technique sensitive to the contextual nature of learning offers a useful framework for tailoring technologies that help empower, not hinder, the educational sector.

Categories and Subject Descriptors

Computing Milieux - Computers and Education – Computer Users in Education

General Terms

Design, Human Factors

Keywords

Science, mathematics, engineering, technology, education, learning technology

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

1. INTRODUCTION

Instructional technology and 'e-learning' gain momentum daily as educators and educational policy-makers strive to incorporate Web-based learning strategies to improve education and achievement. In the United States, this momentum undoubtedly benefits from former President Clinton's initiatives on closing the digital divide advocated by the President's Information Technology Advisory Committee[1]. Congress has commissioned a Web-based education commission[2] to set policy for leveraging the Internet as a vehicle for education. The IEEE Learning Standards Technology Committee[3] works to develop standards and specifications to facilitate computer implementations of education and training components and systems.

However, the focus, whether in government, industry or academia, seems placed squarely on advancing the "technology" part of instructional technology rather than "instruction." Learning technologies could be characterized as technology-focused, that is, focused on the enabling technologies rather than tailoring their design to local educational practices. The bulk of the current set of learning technologies deliver tools needed to create, deliver, and manage on-line courses. More emphasis and effort is placed on full-featured learning management systems and ever more complex instructional technology systems incorporating more content and more capabilities.

Anecdotal evidence suggests that teachers and students are often frustrated by these systems. The need exists for systems that place instruction at the core of design and technical functionality, to distinguish them from systems that merely contain educational content but whose design and technical functionality is indistinguishable from a wide-array of others. We advocate a contextual design approach focused particularly on applying current learning research to instructional technology.

A prototype platform and portal for a new instructional system based on this approach is being built for the science, mathematics, engineering and technology education (SMETE) community. [11] The mission of this instructional system is to create an environment in which educators and students work together as active learners – a public space for the teaching and learning of science, mathematics, engineering and technology as an integrative study.

The foundation of this instructional system is a digital library of learning object resources. The digital library offers direct access

[1] http://www.ccic.gov/ac/

[2] http://www.hpcnet.org/webcommission

[3] http://ltsc.ieee.org/

to and delivery of instructional resources through the establishment of a federation of representative SMETE digital libraries. The digital library promotes learning through personal ownership and management of the learning process while connecting the leamer with the content and communities. of learners and educators. Content and services provided through the digital library will generally include multimedia courseware, digital problem sets and exercises, educational software applications, related articles and joumals, and instructional technology services for educators and students, both commercial and non-commercial – all organized and labeled for the purpose of education and instruction.

Layered on top of the content are various tools, such as search tools, learning object management tools, and community-building tools, which leverage the educational content. These tools will permit users to leam, connect and manage their personal educational portfolio. The digital library seeks to connect communities of educators and leamers to a rich set of pedagogical resources for SMET education. Educators engage intelligently in public discourse and debate about matters of technical importance. This combination of SMETE contents, services, and tools positions the digital library as a 'learning space' for learning science, mathematics, engineering and technology.

Direct access to a broad collection of SMET educational content and services presents simultaneously an opportunity and a source of frustration for the leamer and educator. In order to conceptualize and design a new instructional system, instructional considerations, not technology, must take center stage. Utilizing a constructivist view of leaming, we outline the specifications for an information architecture for the presentation of educational resources in a manner that communicates education and instruction to the educational user. Under this viewpoint, the information design of the digital library should increase students' responsibility for their own leaming; they take control of their actions and interactions and organize their own time for learning. At the same time, the design should empower the educator to guide the student through this educational path. Whoever the user may be, the objective is the same: to assemble educational resources for the purpose of leaming. The essence of the information architecture is to devise interactions that let users achieve this goal.

Information architecture [14] deals with the design of organization, labeling, navigation, and searching systems to help people find and manage information. This implementation paper introduces the grounding principles for the information architecture based on instructional considerations for each of the objectives of information architecture, that is, organization, labeling, navigation and searching. Based on the principles, we formulate an information architecture that supports these principles. The principles are based primarily on constructivist theories of learning with reference to information processing [1] as a model of mental cognitive tasks [7]. A visual concept and reference technical implementation that exemplifies the information architecture is presented. The paper concludes with plans for an assessment of the effectiveness of the architecture in meeting the design principles.

2. Learning: A Constructivist Viewpoint

A prevailing theory of leaming, constructivism, asserts that leaming is based on students "constructing" their own knowledge by testing ideas and approaches based on their prior knowledge and experience, applying these to a new situation, and integrating the new knowledge gained with pre-existing intellectual constructs. [8] Information access environments, such as an educational digital library, offer students and educators opportunity to gain access to these "pre-existing intellectual constructs." Unfortunately, giving them unstructured access to the digital library is similar to an unguided visit to a "bricks and mortar" library. Pedagogical structure is necessary for leaming and education to happen. Deciding which resources to use, and what information to extract, that is, altering, rearranging or recomposing information, are among the numerous information processing tasks [13] associated with constructing mental models [8] that could be better left up to the educational user. The tasks of the digital library are to find the leaming resources, supply useful tips on applying them to current leaming goals, and surface information that would aid in the decision to incorporate the learning elements. The following sections propose principles for applying pedagogical structure to the information architecture of an educational digital library in support of this constructivist view of leaming.

2.1 Information Organization

Separating content and context, or content and leaming processes, affords leamers the flexibility of applying leaming objects towards different instructional strategies to teach the same or related subject matter. One of the primary challenges of information organization for education is the uncertainty in knowing the cognitive models of the leamer or educator. Studies in generative leaming theory for science and technology [11] posit that leaming is not necessarily limited to the manipulation of existing cognitive structures but rather begets new associations for the leamer. We are led then to the following principle.

Principle 1: Organize information to provide opportunities for students and educators to create, synthesize, manipulate or debate content rather than merely to passively receive instruction.

2.2 Information Labeling

Using appropriate instructional strategies for a particular level of leaming and incorporating necessary conditions for leaming presentation are significant components of success associated with a specific instructional delivery mechanism. The effectiveness of the leaming resource hinges on the type of learning undertaken, whether the pedagogical style is inquiry-based, project-based, peer-based, or model-based, among others. This leads to the second principle specification.

Principle 2: Label resources with pedagogical identifiers such as age group, teaching method, and academic standards to indicate educational uses.

2.3 Information Navigation

Few would argue that the best possible instruction involves individualized treatments that differ in structure and completeness depending on the leaming goals and ability of the leamer. Research has shown that curriculum with highly structured treatments seem to help students with low ability but hinder those with high abilities. [3] The navigation scheme of the digital library acts as a proxy for curriculum in so far as how the digital library guides users through the task of finding leaming elements. The implication for the information architecture then is to balance prescriptive navigation while allowing users ability to explore. Two individuals using the same navigation should be guided to only the specific information that a particular learner wants or needs in the appropriate manner and at the appropriate time.

Principle 3: Guide the collection and adaptation of learning elements towards individual leaming goals.

2.4 Information Search

A popular approach to implement digital library search services is to utilize an existing full-text information retrieval system such as Google. The primary deficiency, though, is that these systems lack sufficient context for leaming. The question that needs to be asked is how to formulate a search that aids in personal leaming. The difference between finding a leaming resource and doing research for developing course curriculum versus merely conducting an information retrieval task lies in the interpretation of relevance. Relevance in the sense of leaming is much more complex than as treated by standard information retrieval methods. A search for "solar system" could retum numerous relevant documents with high precision, defined in information retrieval terms, but not contain information addressable by a class of high school physics students performing an inquiry-based study of orbital pattems nor would it be useful for providing information relevant to the specific leaming context.

Educational objectives should be searchable and listed in the search results. The extent to which a leaming element is relevant correlates with how the leaming element achieves a learning goal. This leads to the following principle.

Principle 4: Optimize search to meet the interests, knowledge, understanding, abilities, and experiences of the users in their roles as educators or students.

3. Methodology

This information design methodology follows along the principles of "contextual design." [2] Contextual design suggests that systems development should follow a deep understanding of the users' work, thereby explicitly defining the interaction of the users with the system. We utilized several tactics for obtaining this task information: 1) a review of user needs provided by case studies and user scenarios; 2) a simple benchmark of two existing, prototypical educational digital libraries; and 3) user personas and task modeling.

3.1 Analysis of Existing User Studies

An important first step in this process is gaining knowledge of the educational objectives of leamers and educators and what tasks they would undertake in the digital library to achieve their objectives. This knowledge was gathered from a review of use scenarios from an evaluative case study of users [9] in the National Engineering Education Delivery System (NEEDS) [10] and user scenarios[4] developed by the Digital Library for Earth Systems Education (DLESE). Both of these studies centered on identifying and clarifying users' (end-users such as leamers and educators as opposed to catalogers and authors) needs regarding the functionality and behavior of an educational digital library and prototypical tasks in utilizing the digital library's resources.

The major findings from the above studies were:

- When users search for instructional materials (leaming objects), they want to find materials that are useful and meet their needs. For educators, "useful" materials include labs, exercises, lecture notes and primary materials. Educators also noted the desire for quality standards and peer review as a filtering mechanism. For students, useful materials were resources for papers or research.
- Faculty noted the importance of instructional and pedagogical labeling to understand successful courses and learning experiences.
- Given the broad array of disciplines represented by the resources of the digital library, both students and educators noted the need for cross-references to information that cuts across the disciplines. Tools such as glossaries and thesauri could ease the referencing of such material.
- Both students and educators emphasized community as an integral component in the leaming process.

Based on these findings, we identified the following high-level user needs:

- Information Organization
  - Methods to organize the materials around personal context(s) rather than a prescriptive context.
  - Community resources for social network building in particular to assist in discussion and synthesis of leaming resources.
- Information Labeling
  - Educational labels to describe leaming resources.
- Information Navigation
  - Leamer information profile to improve adaptation of leaming elements towards educational goals.
- Information Search
  - Leaming descriptors as search filters.

3.2 Benchmarks of Existing Educational Digital Libraries

For the second phase, we compared the information organization, labeling, navigation and search of NEEDS and MERLOT[5], two widely used educational digital libraries, on the high-level needs shown in Table 1.

Table 1 Comparison of educational digital libraries

	NEEDS	MERLOT
Information Organization		
Methods to organize the materials around personal context(s) rather than a prescriptive context.	No	No
Community resources for social network building in particular to assist in discussion and synthesis of leaming resources.	Yes	No
Information Labeling		
Educational labels to describe learning resources.	Yes	Yes
Information Navigation		
Leamer information profile to improve adaptation of leaming elements towards educational goals.	No	No
Information Search		
Leaming descriptors as search filters.	Yes	Yes

3.3 User Personas and Task Scenarios

Finally, we developed user personas and task scenarios these users might potentially undertake to reach their leaming objectives. We noted that finding learning resources using this digital library represents but one of many tasks that a user might undertake to satisfy their leaming objectives. Further, we thought through how

[4] Available at http://www.dlese.org/usecases/forum.html.

[5] http://www.merlot.org/

the digital library might complement the user's existing learning environment rather than supplant it with a virtual learning environment. That is, the digital library should enhance overall learner and educator productivity rather than enhance just the ability to search and retrieve learning resources.

Examples of the user personas are:

1) Sally is a 10th-grade student at an inner-city high school. Sally's favorite course is biology and would like to become a doctor or veterinarian.
2) Tom has taught general science in junior high for over 10 years and would like to be rejuvenated. He has started to attend professional development workshops where he has been introduced to new pedagogical techniques for teaching science.

Tasks these personas might undertake include:

Student:

The student has a homework assignment on "electrons." To complete the assignment, the student begins by browsing through the digital library collection of materials appropriate for K-12 general science and mathematics education on "electron configuration."

A student has just gone on a field trip to an observatory. The student learns about solar winds and their effect on radio transmission. The student is not able to find materials on solar winds and thus posts a question to a "science buddy" about solar winds.

Teacher:

A teacher is preparing a new lesson plan for a general science class. The teacher finds curriculum material. Instead of printing them out, he directs his students to the material on-line where they can work together on some experiments. An example of this on-line personal collection is shown in Figure 1.

A teacher wants to share her experience using a "mathlet" for plotting equations. The teacher uses the Comments and Reviews feature, shown in Figure 3, to attach a teaching use comment with the learning resource. The comment is sent to the author.

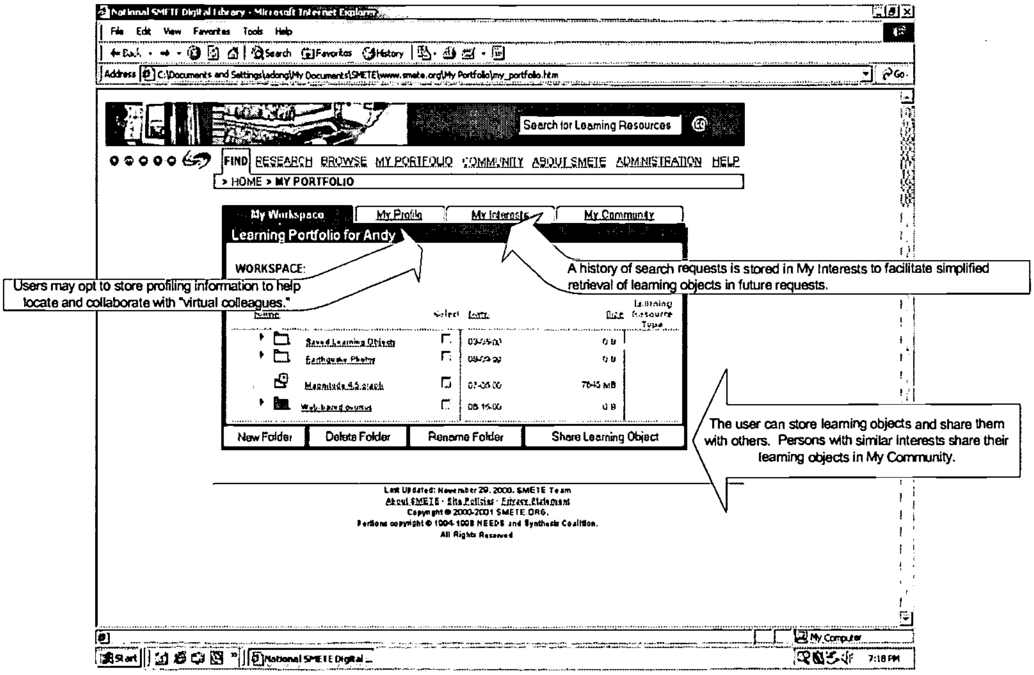
4. The Architecture

The primary driver for the information architecture of an education digital library is to facilitate a better way to accommodate the task of retrieving learning objects that can be re-used for learning. A reference information architecture for a SMETE digital library, available for preview at http://www.smete.org/, is presented.

4.1 Information Organization

The challenge here is to develop a personal collection or learning portfolio to keep track of materials found, not necessarily for convenience, but rather to assist the learner or educator in conceptualizing the material in a manner conducive to individual learning. Further, the collection can be shared to engage discussion between the user and the author(s) or contributor(s) of the resource as well as with others interested in learning about the same subject matter using the same learning object. The learning portfolio shown in Figure 1 is both a radiant experience and focal point for organizing material. As a secondary purpose, it allows the user to navigate through the space of digital library resources in a concept space that is defined by the user.

The folder metaphor is just one interface currently under development to assist users in organizing learning resources. Other interfaces, including concept maps [3] in which users create their own concept maps and organize materials according to the map, are currently under consideration.


National SMETE Digital Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address C:\Documents and Settings\adong\My Documents\SMETE\www.smete.org\My Portfolio\my_portfolio.htm
Go
Search for Learning Resources
FIND RESEARCH BROWSE MY PORTFOLIO COMMUNITY ABOUT SMETE ADMINISTRATION HELP
> HOME > MY PORTFOLIO
My Workspace
My Profile
My Interests
My Community
Learning Portfolio for Andy
WORKSPACE:
New Folder
Delete Folder
Rename Folder
Share Learning Object
My Computer
Start
National SMETE Digital ...
7:18 PM

Name	Select	Date	Size	Learning Resource Type
Saved Learning Objects	☐	03-25-00	0 B	
Earthquake Photos	☐	09-29-99	0 B	
Magnitude 4.5 graph	☐	07-26-00	76.45 MB	
Web-based courses	☐	08-15-00	0 B	

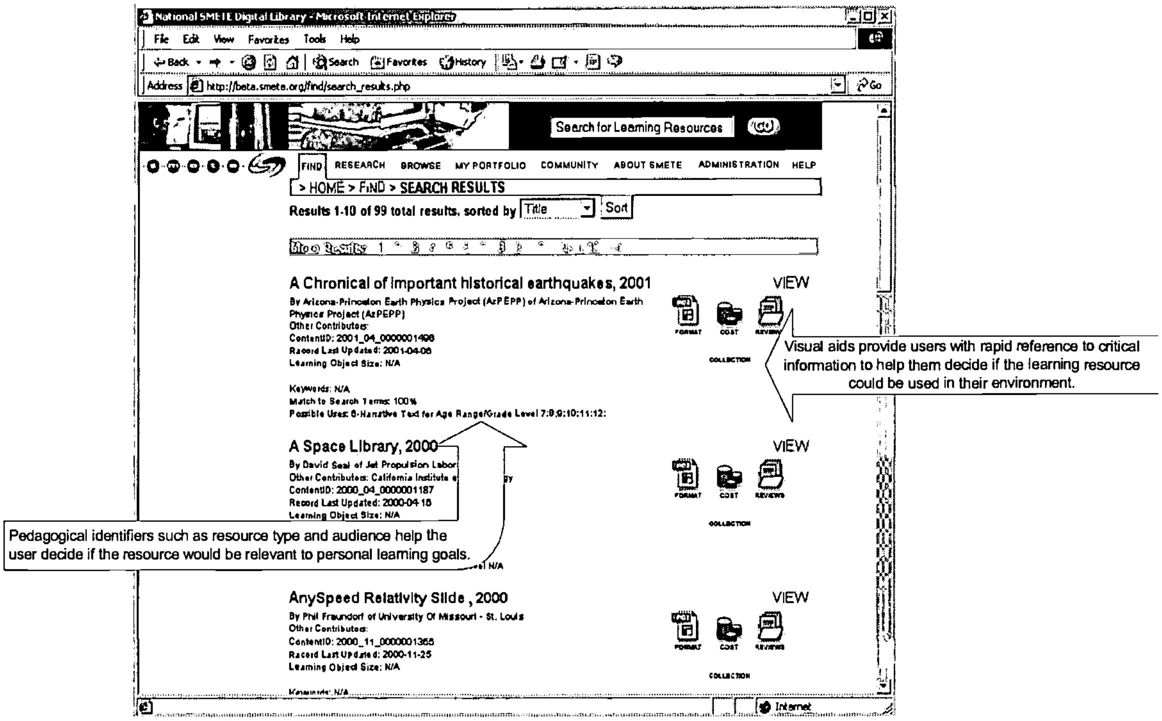
Last Updated: November 29, 2000, SMETE Team
About SMETE · Site Policies · Privacy Statement
Copyright © 2000-2001 SMETE.ORG.
Portions copyright © 1994-1998 NEEDS and Synthesis Coalition.
All Rights Reserved

Users may opt to store profiling information to help locate and collaborate with "virtual colleagues."

A history of search requests is stored in My Interests to facilitate simplified retrieval of learning objects in future requests.

The user can store learning objects and share them with others. Persons with similar interests share their learning objects in My Community.

Figure 1 The My Portfolio service is a radiant experience and focal point for organizing material. The user is able to create a personal conceptual model of learning resources and share that model with others.


National SMETE Digital Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://beta.smete.org/find/search_results.php
Go
Search for Learning Resources
FIND RESEARCH BROWSE MY PORTFOLIO COMMUNITY ABOUT SMETE ADMINISTRATION HELP
FORMAT COST REVIEWS
COLLECTION
VIEW
Internet

> HOME > FIND > SEARCH RESULTS

Results 1-10 of 99 total results, sorted by Title Sort

A Chronical of important historical earthquakes, 2001
By Arizona-Princeton Earth Physics Project (AzPEPP) of Arizona-Princeton Earth Physics Project (AzPEPP)
Other Contributors:
ContentID: 2001_04_0000001498
Record Last Updated: 2001-04-06
Learning Object Size: N/A

Keywords: N/A
Match to Search Terms: 100%
Possible Uses: 6-Narrative Text for Age Range/Grade Level 7:8:9:10:11:12:

A Space Library, 2000
By David Seal of Jet Propulsion Labor
Other Contributors: California Institute o gy
ContentID: 2000_04_0000001187
Record Last Updated: 2000-04-18
Learning Object Size: N/A
el N/A

AnySpeed Relativity Slide, 2000
By Phil Fraundorf of University Of Missouri - St. Louis
Other Contributors:
ContentID: 2000_11_0000001365
Record Last Updated: 2000-11-25
Learning Object Size: N/A

Keywords: N/A

Visual aids provide users with rapid reference to critical information to help them decide if the learning resource could be used in their environment.

Pedagogical identifiers such as resource type and audience help the user decide if the resource would be relevant to personal learning goals.

Figure 2 The search results indicate information to help decide if the learning resource would be relevant and useful to learning goals by surfacing educational identifiers. Equally important is Cost, which can restrict access to the resource.

4.2 Information Labeling

The key goal in information labeling is supporting the discovery of education resources rather than merely supporting the discovery of resources in general. The search results (Figure 2) surface educational use indicators such as age group, learning resource type and learning context through the use of the IEEE Learning Object Metadata (LOM). [6] While the LOM, and similarly the Dublin Core metadata in the description of educational resources [5], do not currently make semantically identifiable statements about the pedagogical aspects of a learning resource, information indicating how the learning resource can be integrated into a curriculum can be made available through the use of domain-specific 'best practices' vocabulary. In the LOM, the tags

Educational.LearningResourceType,
Educational.TypicalAgeRange and
Educational.TypicalLearningTime

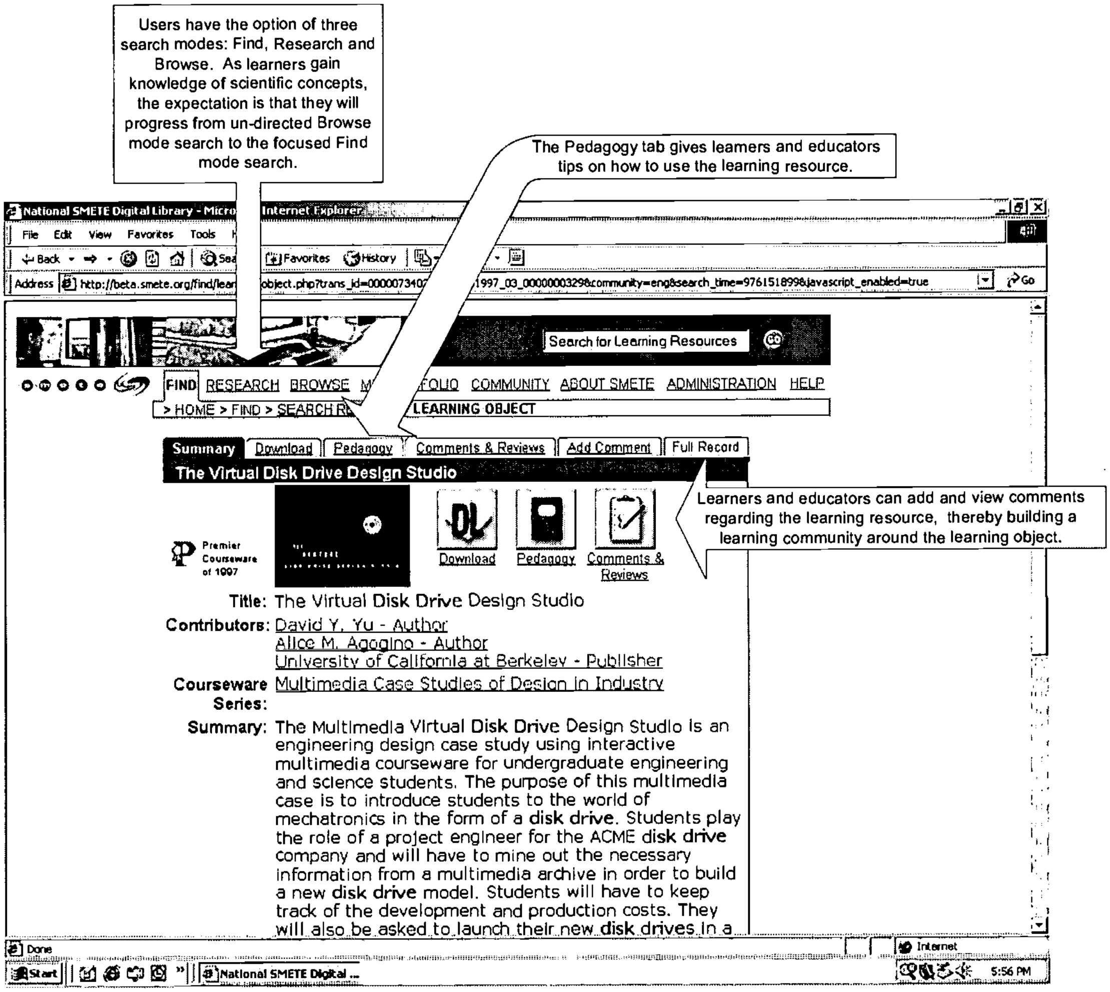
offer teaching and learning information that may aid in the discovery, retrieval and eventual use of the learning resource when populated with appropriate controlled vocabulary. That is, by labeling learning resources with information about how they might be used, the labeling supports better learning through better instructional design.

These domain-specific vocabularies for describing the educational use of learning resources are being developed in various communities. The Mathematics Association of America has completed a subject thesaurus for mathematical concepts and the Eisenhower National Clearinghouse has completed a taxonomy of mathematical concepts for K-12 education. These thesauri are being integrated into the information labeling of this educational digital library.

4.3 Information Navigation

Since a learning object normally has several elements and requires instruction about the pedagogy, navigation (Figure 3) is required to guide users towards the adaptation and collection of learning objects associated with different learning goals. Navigation cues to "learning elements" lists the components of the learning object with a brief description of the elements associated with the learning object. Cues to "pedagogy" contain information on how the user might integrate the material into curriculum including references to past successful implementations.

Ad-hoc on-line communities for review and assessment integrated into the search results and learning objects, cross-linked by subject and description keyword(s), guide users towards debating and synthesizing learning materials in collaboration with others. By bringing in these discussions, learners and educators can capture issues and others' conceptual models to develop a shared view of the subject.



Users have the option of three search modes: Find, Research and Browse. As learners gain knowledge of scientific concepts, the expectation is that they will progress from un-directed Browse mode search to the focused Find mode search.

The Pedagogy tab gives learners and educators tips on how to use the learning resource.

Learners and educators can add and view comments regarding the learning resource, thereby building a learning community around the learning object.

National SMETE Digital Library - Micro Internet Explorer
File Edit View Favorites Tools
Back Favorites History
Address http://beta.smete.org/find/lear object.php?trans_id=000007340 1997_03_00000003298community=eng&search_time=976151899&javascript_enabled=true Go
Search for Learning Resources
FIND RESEARCH BROWSE M FOLIO COMMUNITY ABOUT SMETE ADMINISTRATION HELP
> HOME > FIND > SEARCH R LEARNING OBJECT
Summary Download Pedagogy Comments & Reviews Add Comment Full Record
Premier Courseware of 1997
Download Pedagogy Comments & Reviews
Done Internet
Start National SMETE Digital ... 5:56 PM

The Virtual Disk Drive Design Studio

Title: The Virtual Disk Drive Design Studio

Contributors: David Y. Yu - Author
Alice M. Agogino - Author
University of California at Berkeley - Publisher

Courseware Series: Multimedia Case Studies of Design in Industry

Summary: The Multimedia Virtual Disk Drive Design Studio is an engineering design case study using interactive multimedia courseware for undergraduate engineering and science students. The purpose of this multimedia case is to introduce students to the world of mechatronics in the form of a disk drive. Students play the role of a project engineer for the ACME disk drive company and will have to mine out the necessary information from a multimedia archive in order to build a new disk drive model. Students will have to keep track of the development and production costs. They will also be asked to launch their new disk drives in a

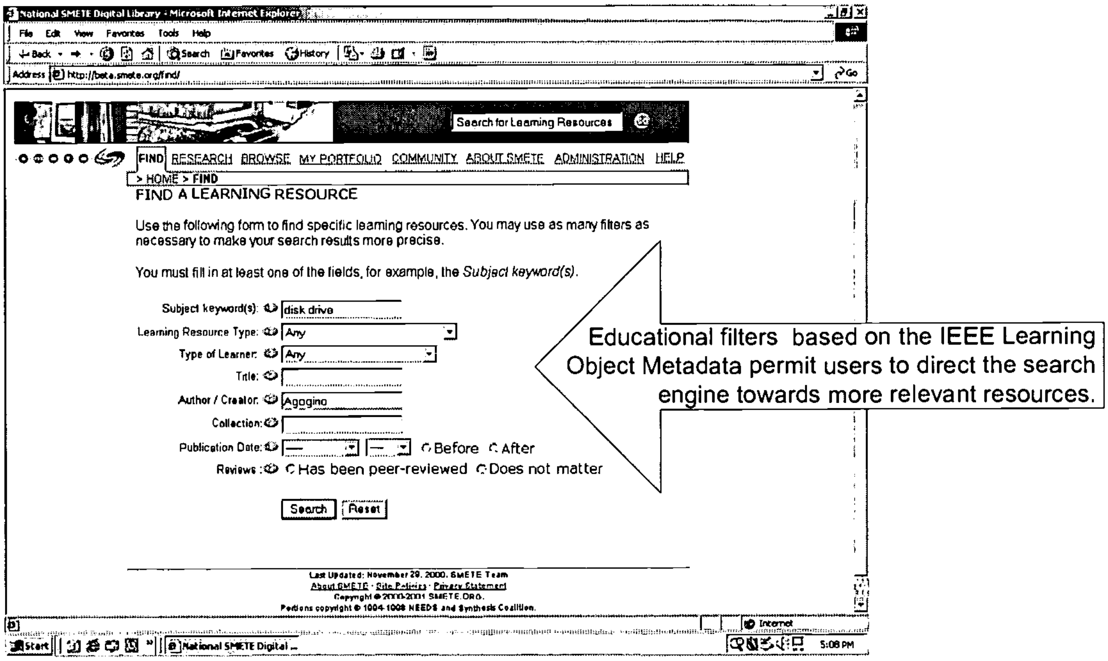
Figure 3 The learning object forms the nucleus of information from which users can learn how to incorporate the learning object, and add comments on their experience with the learning object.

4.4 Information Search

One of the important pieces of user feedback regarding the search mechanism of the NEEDS digital library was the confusion between "Search" and "Advanced Search," the former connoting usefulness to novice users of the digital library and the latter to advanced users. Instead, the difference between the two modes of search is really the level of filtering, with "Advanced Search" containing more learning descriptors as search filters. As such, we differentiate the search mechanisms in terms of assisting users find materials with potentially more relevance to their education levels, experience and knowledge.

A more relevant distinction to searching for learning resources is the level of knowledge the user has of the SMET concepts and how those concepts would typically be addressed. That is, an experienced person in a specific subject discipline might use the Find feature to locate a particular learning object whereas a beginning student in the field might utilize the Browse feature to locate learning objects within a more general subject category familiar to the student.

In the "Find" mode (Figure 4), users apply numerous learning descriptors to direct the search engine towards particular learning object(s). A typical "Find" search might be, "Find the courseware titled 'Interactive Frog Dissection.'" In "Research" mode, users conduct a broad search using subject headings, with the option of filtering the results by a set of learning descriptors. The "Research" mode also offers the capability to view courseware within a collection by subject headings, such as all "civil engineering" courseware from the "NEEDS" collection. In "Browse" mode, users conduct high-level subject searches unconstrained, necessarily, by any learning descriptors. Hierarchical browsing through subject keywords, visual browsing through an image database of multimedia learning elements, and browsing by teaching method assists the user in conceptualizing at a high level the type of resources available in the digital library. It is expected that users unfamiliar with a subject area would begin by browsing, eventually using some of the navigational aids and information labeling to inform more directed "Find" or "Research" to locate specific learning objects.



National SMETE Digital Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://beta.smete.org/find/
Go
Search for Learning Resources
FIND RESEARCH BROWSE MY PORTFOLIO COMMUNITY ABOUT SMETE ADMINISTRATION HELP
> HOME > FIND
Search
Reset
Internet
Start
National SMETE Digital ...
5:08 PM

FIND A LEARNING RESOURCE

Use the following form to find specific learning resources. You may use as many filters as necessary to make your search results more precise.

You must fill in at least one of the fields, for example, the *Subject keyword(s)*.

Subject keyword(s): disk drive
Learning Resource Type: Any
Type of Learner: Any
Title:
Author / Creator: Aggogino
Collection:
Publication Date: — — ○ Before ○ After
Reviews: ○ Has been peer-reviewed ○ Does not matter

Last Updated: November 29, 2000. SMETE Team
About SMETE · Site Policies · Privacy Statement
Copyright © 2000-2001 SMETE.ORG.
Portions copyright © 1994-1999 NEEDS and Synthesis Coalition.

Educational filters based on the IEEE Learning Object Metadata permit users to direct the search engine towards more relevant resources.

Figure 4 The Find page incorporates educational filters based on the IEEE Learning Object Metadata standard. These filters permit users to direct the search engine towards more relevant resources depending upon their educational needs.

5. Conclusions

The objective of this implementation paper was to investigate the support of learning through the resources of a digital library and then translate those needs to the design of an information architecture for an educational digital library. By focusing on the learning objectives of educational users, rather than how to enable a particular user interface or learning technology, the study establishes principles on how to introduce features that focus the digital library on its educational goals. By understanding the motivations that lead individual educators and learners to utilize the resources of an education digital library, we were able to obtain knowledge of how a larger population of learners and educators would use the digital library. The information architecture principles present guidelines on how to design an education digital library that better satisfies the information processing tasks associated with learning, particularly inquiry-based learning. These principles enabled the design of a digital library for networked information discovery and retrieval with education being the objective, a primary distinction from information retrieval in general.

Currently under parallel development is an Instructional Architect[6], an integrated Web-based development environment in which learning objects can be assembled into instruction with a method for enabling the spontaneous formation of new communities of users based on shared interests. Users may search for learning objects using the digital library presented in this paper and download them into the Instructional Architect system.

Assessment studies are currently being conducted on a lead-user group. The evaluation plan for the information architecture is organized loosely around three types of evaluation methods:

- Assessment of user needs
- Formative assessment focused on improving services and features
- Short and long term impacts of the digital library on the SMET educational community including students

In this early stage of development, we are focusing our evaluation efforts on identifying user needs in order to ensure that the library design reflects those needs. The needs assessments have been instrumental in the design of our site and are essential to developing an accessible library that does not exacerbate the growing digital divide. To better understand the needs of our diverse community of users we conducted in person focus groups and, this year, will implement on-line focus groups. We also continue to survey our registered users on a regular basis as well as specified samples within the larger population. Our next steps will center on implementing user studies of various users (e.g., students, faculty, K-12 teachers) interacting with the site. Here we will implement observation studies of users as well as coordinated surveys of users.

Questions guiding the user studies include:

- Do the users interact with the site as designed?
- Are the users satisfied with the site? Do they find it easy to navigate? Do users find useful materials? What hinders their use of the site?
- Do users prefer particular aspects of the interface design?

To examine the short and long term impacts, we will conduct more in-depth studies that focus on the expected outcomes associated with our goals. These studies will include tracking usage statistics and patterns across, between and among specific user communities as well as for the library as a whole. Metrics

[6] http://ia.usu.edu/

associated with expected outcomes for teachers, students and in some cases courses, curricula and schools may be examined. Particular attention will be focused on the impact on student learning. Studying the immediate and long-term impact of the library is a complex project requiring expertise from various different disciplines. Collection of use data is currently underway, and planning has begun with regards to long-term studies.

The primary impact the digital library may have on teaching and learning is to actively engage the participants in the creation of shared conceptual understanding of science, mathematics, engineering and technology as an integrated study. By giving teachers and students access to learning resources across disciplines with the flexibility to incorporate them into their own personal educational goals, the digital library creates new learning opportunities for students and new ways to present materials for teachers.

6. References

[1] Atkinson, R. C., and Shiffrin, R. M., "Human memory: A proposed system and its control processes," In Spence, K. W., and Spence, J. T. (Eds.), *The Psychology of Learning and Motivation*, 1968, pp. 90-191, Academic Press.

[2] Beyer, Hugh and Holtzblatt, Karen, "Contextual Design," *Interactions*, February 1999, 30-42.

[3] Chen, H.C., Houston, A.L., Sewell, R.R., and Schatz, B.R., "Internet browsing and searching: User evaluations of category map and concept space techniques," *Journal Of The American Society For Information Science*, **49**(7), 1998, 582–603.

[4] Chu, Larry F., and Chan, Bryan K., "Evolution of web site design: implications for medical education on the internet," *Computers in Biology and Medicine*, **28**, 1998, 459-472.

[5] Dublin Core Metadata Initiative Education Working Group, http://dublincore.org/groups/education/.

[6] IEEE Learning Technologies Standards Committee Learning Object Metadata base document available at http://ltsc.ieee.org/doc/wg12/LOMdoc2_4.doc.

[7] Klahr, David and Kotovsky, Kenneth, (Eds.), *Complex Information Processing: The Impact of Herbert A. Simon*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989.

[8] Larochelle, Nadine Bednarz, Bednarz, Nadine and Garrison, Jim, (Eds.), *Constructivism and education*, Cambridge, U.K., Cambridge University Press, 1998.

[9] McMartin, Flora, "Preliminary Findings from 'Science, Mathematics, Engineering, and Technology Education' User Study Focus Groups." http://www.needs.org/engineering/info/papers/focusgroup599/focusgroup599.pdf.

[10] Muramatsu, Brandon, and Agogino, Alice, "The National Engineering Education Delivery System: A Digital Library for Engineering Education," *D-Lib Magazine*, April 1999, Vol. 5, Issue 4, ISSN: 1082-9873.

[11] *Report of the SMETE Library Workshop*, National Science Foundation, July 21-23, 1998, http://www.dlib.org/smete/public/report.html.

[12] Schaverien, L and Cosgrove, M., "A biological basis for generative learning in technology-and-science Part I: A theory of learning," *International Journal of Science Education*, December 1999, **21**(12), 1223-1235.

[13] Simon, H. A. and Feigenbaum, E. A., "An information-processing theory of some effects of similarity, familiarization, meaningfulness in verbal learning," *Journal of Verbal Learning and Verbal Behavior*, **3**, 1964, 385-396.

[14] Wurman, Richard Saul, *Information Anxiety*, New York, Doubleday, 1989.

Acknowledgements

The National Science Foundation funded this research work under Grant #DUE-0085878 and program solicitation DUE #00-44 for the National SMETE Digital Library Program. The authors would like to acknowledge the members of the SMETE.ORG Alliance for their invaluable feedback during the development of the information architecture. This research is part of an ongoing collaboration to build a National SMETE Digital Library at www.smete.org.

Toward A Model of Self-administering Data

ByungHoon Kang
Division of Computer Science
UC Berkeley
Berkeley, CA 94720
510 642-8468
hoon@cs.berkeley.edu

Robert Wilensky
Division of Computer Science
UC Berkeley
Berkeley, CA 94720
510 642-7034
wilensky@cs.berkeley.edu

ABSTRACT

We describe a model of *self-administering data*. In this model, a declarative description of how a data object should behave is attached to the object, either by a user or by a data input device. A widespread infrastructure of *self-administering data handlers* is presumed to exist; these handlers are responsible for carrying out the specifications attached to the data. Typically, the specifications express how and to whom the data should be transferred, how it should be incorporated when it is received, what rights recipients of the data will have with respect to it, and the kind of relation that should exist between distributed copies of the object. Functions such as distributed version control can be implemented on top of the basic handler functions.

We suggest that this model can provide superior support for common cooperative functions. Because the model is declarative, users need only express their intentions once in creating a self-administering description, and need not be concerned with manually performing subsequent repetitious operations. Because the model is peer-to-peer, users are less dependent on additional, perhaps costly resources, at least when these are not critical.

An initial implementation of the model has been created. We are experimenting with the model both as a tool to aid in digital library functions, and as a possible replacement for some server oriented functions.

Keywords: Self-administering data, data access model, data management, peer to peer, distributed file system, asynchronous collaboration, file sharing, scalable update propagation.

1. Introduction

Central to the digital libraries enterprise are issues of creating, managing and facilitating access to collections of digital objects. At one extreme, emulating traditional libraries, a digital library may serve as finding aid, collection manager, repository, and distributor for a set of digital objects. Alternatively, these services may be disaggregated, with affiliated or independent services providing collection management service, repository service, and so forth. Disaggregation, we suggest elsewhere ([20]), can enable more "democratic" management of resources, with individuals and groups using library-like services to manage their own collections, and to make use of associated services.

As digital libraries become more democratic, their interaction with other aspects of the data management process will become more important. For example, a working group may want to house the latest draft of a document in a repository for general access. While they are working on the document, versions need to get passed around or shared, and, as work progresses, the draft in the repository will need to be updated. In addition to incorporation, repositories need to be able to disseminate their data to interested users, who may be able to use related services to comment upon or annotate the document. For example, a user reading the draft may appreciate receiving an updated copy when the draft is changed, or receive notifications of various sorts. Such scenarios represent a need for end-to-end data management, in which collection and repository management form only one piece, and whose boundary with personal and group information management become indistinct.

Here we propose a data management strategy that is largely concerned with personal information management issues as they relate to the larger picture of digital object management. While we are developing this strategy to support digital library tasks, such as incorporation and dissemination of digital objects into conventional digital library structures, the data model we present may have additional implications. In effect, we provide a peer-to-peer digital object management tool. This tool can support interaction with services, but it may also serve as replacement for them. Thus, we propose for consideration a more radical notion of a peer-to-peer version of digital libraries, with no services at all.

2. Scenarios and Design Goals

In this section, we consider a number of scenarios that motivate our design. Mostly, we express our frustration with current tools available for simple processes, and suggest what, to us, seem like more attractive scenarios. Below we present the data processing model we designed to enable these scenarios. Then we describe our initial implementation.

2.1 Co-Authoring across Administering Domains

Example Problem: Suppose a web-page designer is commissioned to create some web pages from a customer. The customer somehow communicates what is desired to the designer, who then creates an initial version of these pages. Perhaps the designer sends these page drafts to the customer by email as attachments, or has the customer download the web-pages from the designer's web site, or uses some other protocol, like ftp, to move copies about. Later, the customer makes some modifications and returns the pages to the designer, and the process iterates.

As a result, both users' email boxes, or file spaces, etc, get filled with email attachments of versions. These versions are often hard to

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

manage because there is no built-in version management support for email attachments, HTTP, or FTP. Our collaborators could instead try to use some collaboration tool designed for this purpose. Heavy weight document management system like Lotus Notes [10] or even Xerox's Docushare [11] are probably overkill for this purpose; moreover, they may require administrative commitments neither user can make. Web-based file sharing system such as www.desktop.com [17], www.hotoffice.com [18], WebEdit [6], i-drive [12] and BSCW [4], and synchronization services, such as FusionOne [13], provide an interesting alternative. However, such services don't provide control over important aspects of data management, such as back-up, conversion, and merging. Moreover, the users are at the mercy of a potentially overloaded server, perhaps at a precariously financed dot com. Also, adding a third party to the interaction introduces increased vulnerability: Users are not able to perform their sharing operation when the central server is down even though their local machines and services are functioning, and have introduced a new security concern. In addition, they are subject to various, and, we think, avoidable, human errors, such as forgetting to transmit the shared copy to the web repository after every change.

Desired Properties: The above scenario suggests to us the following properties of an ideal system for this task:

P1: No repeat user involvement in routine data management

P2: No unnecessary dependence on shared resources, such as shared data repositories or file servers

P3: No prior administrative set up costs

P4: Ability to exploit minimal use of central server as only required

P5: Undo/Redo capability within user's domain

P6: Secure and safe incorporation of updates at user's domain

P7: Lightweight enough to be widely deployed

A Proposed Solution: We propose a way of accessing and managing data to achieve the above desiderata. We introduce an infrastructure of *Self-administering Data Handlers*, which are deployed wherever users wish to take advantage of their services. These Self-administering Data Handlers (SD Handlers) administer data according to an attached *Self-administering Data Description*. The Self-administering Data Description (SDD) is metadata describing how, where, and to whom the data are to be copied, updated and otherwise administered. In other words, the SD Handlers are daemon processes that administer data by honoring attached self-administering descriptions.

Consider how the task above might be performed if SD Handlers were available to the collaborating parties. When the web-page designer creates web-pages, she saves them into a directory or folder somewhere on her local disk, as is her standard practice. Her SD Handler detects this action and pops up a UI with a self-administering description for the saved web pages, probably representing her defaults. She examines the default preferences, checks a couple of choices and adds a new destination, in this case, a location specifier provided by the client. Then the SD Handler attaches to the data objects their respective self-administering description.

Suppose the designer specified that these pages should be delivered to client's public web folder whenever she updates one of those. When a page is updated, the SD Handler will automatically sign it with the designer's private key and encrypt the result with the client's public key. The signed and labeled data object is deposited into the network of SD Handler infrastructure.

The client's SD Handler receives and verifies the authenticity of the self-administering description. In this case, it interprets the description as instructing incorporation of the data object into client's web folder. The client's SD Handler logs this event of data incorporation. If the designer's name is not found in the client's *trustee list*, the incorporation is denied. If the recipient's SD Handler is not available, the SD Handler could retry or delegate the retrying of delivery to a pre-negotiated server.

If the designer prefers strong update serialization, the SD Handler might be configured to first contact a pre-negotiated central serialization service, (say, a CVS[8] server or the Oceanstore [2] service) and have her changes merged according to the arrival order at the central serialization server. The merged data are then delivered back to the designer's SD Handler, which forwards the merged data to the destinations specified in the self-administering description.

Such a network of SD Handlers provides a lightweight asynchronous collaboration infrastructure for sharing data in a secure way. Centralized servers may be exploited in this process, but only when there is some particular need that justifies the cost, such as a desire for strong serialization of updates.

2.2 Data-Collection

Example Problem: Let us briefly consider a less desktop-centric scenario. Suppose a botanist takes pictures of plants in a field with her digital camera. She wants to transfer these to multiple remote designations, including her own web page, her research group's database, and her collaborator's disk. To do so, she must go to her office desktop and download the image from the digital camera into some buffer space, and then copy it into her own web page folder. She must then open up a database connection, authenticates herself, and then upload the data into database. She would also pop up an email client, create a new email message and upload the image data as the message's attachment. Then she sends the email to her collaborators, asking them to download the attached image onto their disk space. She repeats these procedures whenever she takes a picture or pictures she wishes to so incorporate.

This scenario provides comparable desiderata to the initial one, except that one would like to deploy our proposal as close to the data as possible. Thus, we must modularize SD Handler functionality so we can implement its services within a device's limited resources. For example, the camera might be enabled with a simple interface for using some pre-downloaded self-administering descriptions. Services that the camera couldn't perform locally could be performed by an affiliated proxy server. The camera need only reach the proxy server for the rest of the tasks to be automated as above.

We envision data collection involving SD Handlers from a wide variety of a simple special purpose devices, include scanners, smart cards, smart mobile phones, PDAs, and lightweight widely distributed sensors. These devices, perhaps together with a helpful proxy, simply deposit their tagged data into the infrastructure, which takes care of all routine data management and transport issues.

3. Self-administering Data Model

As suggested above, we envision a network of SD Handlers, each "close" to a user or device that it serves. To a first approximation, there would be one SD Handler per networked device, perhaps more. Some would be associated primarily with users, some with data collection in devices, others with services, such as digital object repositories, each supporting basic SD Handler functionality, but perhaps implementing services associated with the particular characteristics of its application. Such a network is illustrated in Figure 1.

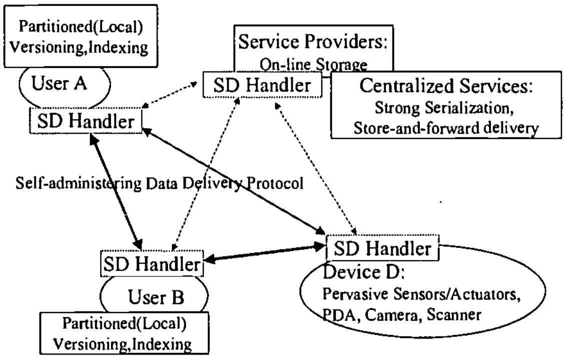


Figure 1: Network of SD Handlers

SD Handlers form a network, within which data is moved in accordance with the SD Handler's discipline. In addition, each SD Handler may provide an interface to a local collection or stream of data. The data may be a user's file system, web space, database, or other collection, administered by some mechanism other than the SD Handler. While these may be administered by a wide variety of mechanisms, the data looks the same once it is with the SD Handler network. We refer to each diverse collection of data as a data *realm*. In effect, the SD Handler bridges a realm into the SD Handler infrastructure.

Figure 2 presents an overview of the architecture of each SD Handler. SD Handlers are required to implement the bottom tier of the architecture, we define its basic functions. These are named bottling, floating, popping, and logging, and are described further below. To exploit capabilities fully, however, it is recommended that SD Handlers also implement an additional tier of functions on top of the basic services. These are called notifying, diff-ing, and versioning. Applications of various sorts may be built on top of these functions. In addition, GUIs and API need to be provided, to communicate with the user, and to form a bridge between the SD Handler and the user's data realms.

3.1 Basic Functionality

Here we present the basic building blocks of SD handling. These are *bottling*, *floating*, *popping*, and *logging*. We then describe how other functions can be built on top of these basic functions.

Prior to this process, a self-administering data description is attached to the data object, akin to creating a packing slip for a shipment. This description contains the shipper's preferences for handling the data, as well as the lists of recipients and/or destinations. The data preference can include high-availability, strong-serialization and default archival support. The destinations can be an on-line storage of collaborator's, a database, a PDA, a smart phone, a speaker/media device and pervasive sensors/actuators. In accordance with this packing slip, at some point, the data object is *bottled*, i.e., prepared for shipping. To make a data bottle, the labeled data object is signed and encrypted. Then the bottle is *floated* across a sea of data. Finally, the bottle is *popped* at its destination(s), and the data extracted. All events are logged, so that support for other services, e.g., version control, and be readily accommodated.

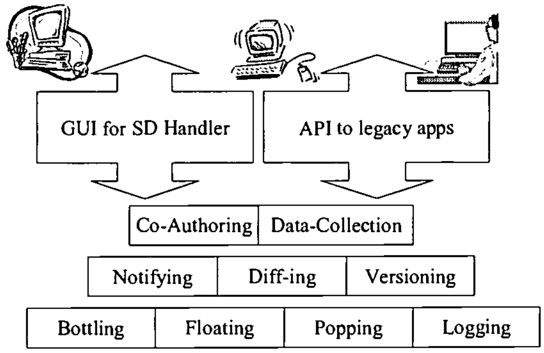


Figure 2: The Tiered SD Handler Architecture

3.1.1 Preparation

Prior to a SD Handler performing any operation on a data object, the object must be bridged into the infrastructure. I.e., a SD Handler has to be made aware of the object, and of the user's specifications for it. This is done by attaching a Self-administering Data Description (SDD) to the data. Since a self-administering data description can have many options and get quite complicated, we assume that most users never work with one directly. Instead, users interact with a UI. We have implemented a standard UI for a SD Handler running on desktop, which we discuss below. We assume that a different UI would be suitable for different devices, and that there would be default description templates for each user and each device, perhaps inherited or cascaded together as a function of the user and device environment.

Once an object has a SDD attached to it, the SD Handler aware of it will begin monitoring the object and attempting to enforce the specification of the SDD. Doing so typically results in sending a copy of the object to one or more recipients.

3.1.2 Bottling

When a SD Handler decides it must send a data object to a recipient, it first prepares a data *bottle*. It does so by signing the self-administering description and its data with its user's private key, for authenticating the sender at the recipient's SD Handler. The result is then encrypted with the destination's public key so that only the real destination can access the description and the data. The sender's credentials are checked against receiver's trustee list to allow appropriate access in incorporating the data at the destinations. Then the bottle is floated, i.e., dropped into the SD Handler network. We describe floating below, but first examine the inverse operation of bottling, *popping*, which occurs when a SD Handler receives a bottle destined for a known user.

3.1.3 Popping

A delivered bottle is inspected for its integrity and the sender is authenticated for appropriate access right. Then the bottle is uncapped with matching encryption keys to be incorporated into the destination realm according to the packing slip. For safe incorporation, every incorporation is logged for undoing or redoing operations.

The trustee-list maintained by SD Handler is used for giving or denying the delivery action from the sender. When SD Handler daemon process receives the bottled data, it authenticates the sender with trustee-list and decrypts the self-administering description to guide the incorporation activities.

Incorporation is based on appending; SD handling never overwrites data, but may shadow it. Since every incorporation is logged, it is always possible to revert the changes back to a specific version.

The bottled data is incorporated through SD Handler into any number of places, and in any number of different manners: onto a user's desktop, PDA, collaborator's domain, online-storage (NFS, Web), database entry, and even subdocument elements, such as anchor points in HTML page. The SD Handler running on a desktop computer maintains the history for the versioned content, and the incorporation activities. If the destination is database, the incorporation could comprise adding new entry; if the destination is a collaborator's online storage, the incorporation may create a newly updated file in a sandboxed location.

The followings are the examples of incorporations at various destinations.

A bottled data delivered

- onto UNIX file system, creates an i-noded file.
- onto a database, creates an updated (appended) database entry.
- onto a repository, creates a new index entry and is moved into repository space.
- onto a speaker device, creates voice data at the speaker
- onto another trusted user's file system, creates an i-noded file in a sandboxed location.
- onto a calendar/address book in a personal information managing application, creates anchor contains new data or new hyperlink pointing to a file in a sandboxed location.
- onto an anchor in a HTML document owned by another trusted user's, creates an anchor contains new data or new hyperlink pointing to a newly updated data in a sandboxed location.
- onto a writable CD, creates a newly added data on the writable CD

3.1.4 Floating

A bottled data object is dropped into the SD Handler network infrastructure. The infrastructure provides the delivery of the bottled data to the destination, as illustrated in Figure 3. SD Handler has its own delivery protocol (SDDP: Self-administering data Delivery Protocol) but SD Handlers can also uses legacy protocols such as HTTP, FTP, and SMTP by tunneling SDDP.

The floating architecture of SD Handler provides store-and-forward service for delivering the data to a recipient who is not available at the time of delivery. It also provides the update serialization service, where the updates from the multiple participants are serially ordered according to the arrival time at the server. The bottled data has to flow into the serialization server and flow out to its destinations.

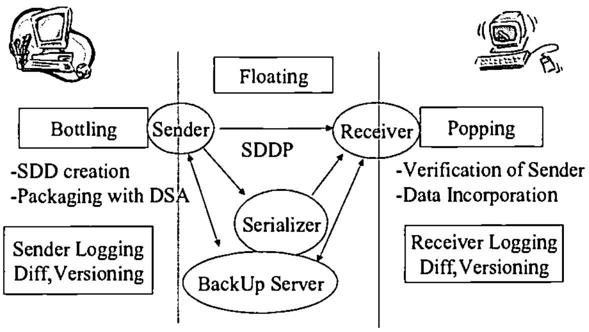


Figure 3: From Sender to Receiver

Finally, the infrastructure of network of SD Handler provides a naming service to map the user's SD Handler's location into its current IP address. Each SD Handler that does not have static IP address, register its current IP address to the name resolution server in the infrastructure. And the SD Sender would cache the latest mapping and use it until the host becomes unreachable, and then it contacts the name resolution server for the current IP address of the participant's SD Handler.

3.1.5 Logging

The data and packing slip and its bottling/popping activities are logged for undoing/redoing and auditing purposes. The logging history can be flushed to a designated archival repository from the local space of the bottling and popping point.

The log can be incremental in that the delta of changes is recorded. Doing so, of course, increases the dependency between logged objects, although it saves the disk space.

3.2 Advanced Functions

There are a set of functions that are useful, but not required, of compliant SD Handlers. We describe these here.

3.2.1 Versioning

Each SD Handler maintains its own version tree at its own realm by enhancing the logging feature. Decentralization is achieved by naming the same resource uniquely along with its version number across different administrative domain. This is done by prefixing the owner-name to a local name of resource, as each SD Handler has its own name space per given owner-name. We use the owner name to uniquely locate its public key. The typical owner name could be email address where the uniqueness is being maintained at its organization or email service provider. In CVS[8] and WebDAV[7] there is one version tree that is maintained at the central server with its single name space. In contrast, in SD Handler, different version trees are maintained at each realm. They share only a naming convention that uniquely addresses identifiable resources across different version trees.

3.2.2 Diff-ing

Given two unique resource names (which includes version numbers), the SD Handler shows the differences of the two data objects. If one or both of data objects are not available, SD Handler looks at the self administering description and contacts the SD Handler which maintains the logs of requested data. Users should be able to diff the changes before and after the transaction so that one

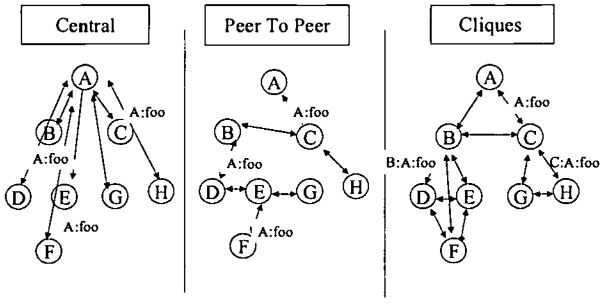


Figure 4: Update Propagation Models

can verify that the automated updates from trusted source are meaningful.

3.2.3 *Notifying*

The SD Handler provides a notification of the basic activities to the end user. This can be achieved by simply showing the log entry that has been added most recently. A user's latest activity involving the same data that is under the established self-administering control such as saving after modification, deleting, would be monitored by SD Handler and notified to the user for further actions.

3.3 Scalable Update Propagation Model

We have developed a scalable update propagation model based on cliques. We define a clique as a strongly connected group of users who share the same SDD. As shown in Figure 4, the update for the data, A:foo, is reported back to the central coordinating user/server, A in the diagram, and then propagated back to the rest of the participants in the central model. In a peer-to-peer model, as shown in Bayou's anti-entropy algorithm [3], the updates are reconciled among peers in an arbitrary order whenever they are connected to each other; then the update is propagated to other members. In our model, changes are immediately propagated within cliques; secondary propagation to other cliques is through the junction point, as shown in B, C in the "cliques" diagram. We can imagine that B and C can filter and aggregate the changes made by members in the clique, so that the tiny changes made within B's clique and C's clique are not visible to A. We believe this approach will fit naturally to group interaction.

Both central and the peer-to-peer propagation models use the same name for all the propagation. However, in our model, the naming carries more clique interaction information. For example, given C:A:foo, one can deduce that this object originated from data A:foo and that user C created a new clique by creating a new SDD associated with C:A:foo. The SDD for A:foo would have a different member's policy and serialization preference. For the data, A:foo, the clique composed of A,B,C would specify strong serialization and back-up support, while, the clique composed of B,D,E,F might not need to use such strong serialization and back-up support among them.

3.4 Self-administering Data Delivery Protocol

The Self-administering Data Delivery Protocol is illustrated in Figure 5.

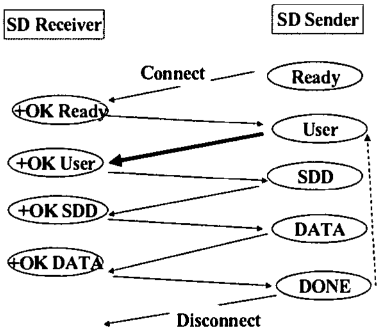


Figure 5: Self-administering Data Delivery Protocol

The SD Sender parses the SDD file written in SDDL (Self-administering Data Description Language) and makes a connection to each Sharer's SD Receiver. When it receives the "+OK Ready" message from the SD Receiver, it sends the SD Sender's user id, (generally an email address). The SD Receiver then checks the membership of this user's id in its SD Receiver's trustee list. If it is not on the list, the SD Receiver refuses the connection and the SD Sender sends an email notification to the user of the SD Receiver. If the user id is found on the trustee list then the SD Receiver tries to verify whether the user of the SD Sender is who she claims to be. For user authentication, we use a public-private key encryption system, where the user of the SD Sender's identity are verified using a digital signature. To avoid the man-in-the-middle-attack, all the links between the SD Sender and the SD Receiver can be encrypted.

After the authentication is finished, the SD Sender sends the SDD to the SD Receiver, followed by the DATA command. The SD Receiver then parses the received SDD and follows its description, which may involve synchronizing the shared document with latest copy given by the DATA command.

If the more files are coming from the SD Sender, it repeats its protocol from the authentication point onwards, until all files have been received; then the connection is closed. The SDDP thus provides a selective delivery acceptance mechanism using a trustee list. Only the owner of the trustee list can add a new user id into its own trustee list.

3.5 Self-administering Data Description (SDD)

The basic responsibility of a SD Handler is to interpret SDDs, i.e., a Self-administering Data Description attached to a data object. SDDs are written in SDDL that is based on XML. Figure 6 provides a typical example.

We stipulate that every SDD has one owner, but may contain multiple users as sharers. Each user in the sharer element can have its own multiple self-administering data server locations. For example, the user "Robert Wilensky" in the following example contains four different SD server locations, the last of the four being his own archival server.

The archival server is an instance of SD server where the SD Handler governs the archival repository for its subscribed users. The owner "B. Hoon Kang" provides its own archival server to be accessible by the sharer. The central server element provides the central server location for SD Handler's "floating" operations (to be described below), such as store-and-forward data delivery, serializing the updates at a central location, and dynamically mapping the user's SD server name into its current IP address.

```
<SELFDATA
 ownername="B.Hoon Kang"
 UAN ="dlib2001 selfdata paper"
 archivalsupport ="yes"
 secureaccesscontrol="yes"
 consistency="yes"
 availability="high"
 changenotification="always">
<UAN name = "dlib2001 selfdata paper">
 <ITEM location = "selfdata01.doc"/>
 <ITEM location = "diagram-image01.gif"/>
 <ITEM location = "diagram-image02.gif"/>
</UAN>
<SHARER coherency = "SERIALIZE" >
 <USER name="B. Hoon Kang"
      id="hoon@cs.berkeley.edu"
      initial="B.H.Kang">
  <SELFDATASERVER
  name="alpine.cs.berkeley.edu" port="7070"
  location="Soda Rm 493" />
  <SELFDATASERVER
  name="sb.index.berkeley.edu" port="7070"
  location="145 wilson st" />
  <ARCHIVALSERVER
   name="dlibarchiver.cs.berkeley.edu" port="7070" />
  <ALTERNATEEMAIL
   name="hoon@now.cs.berkeley.edu" />
 </USER>
 <USER
  name="Robert Wilensky"
  id="wilensky@cs.berkeley.edu" initial="RW">
  <SELFDATASERVER name="bonsai.cs.berkeley.edu"
   port="7070" />
  <SELFDATASERVER
   name="mobile-ip.cs.berkeley.edu" port="7070" />
  <SELFDATASERVER
  name="home-ip.eecs.berkeley.edu" port="7070" />
  <ARCHIVALSERVER
  name="myarchiver.eecs.berkeley.edu" port="7070" />
 </USER>
</SHARER>
<CENTRAL>
 <CENTRALSERVER
  name="galaxy.cs.berkeley.edu" port="7070" />
</CENTRAL>
<ARCHIVE>
 <ARCHIVALSERVER
  name="galaxy.cs.berkeley.edu"
  port="7070" />
</ARCHIVE>
</SELFDATA>
```

Figure 6: A Typical Self-administrating Data Description

A UAN (Unique Activity Name) is used to uniquely specify a SDD. I.e., there is one to one mapping between UAN and SDD. The Activity Name is unique within the owner's "realm", or unit of computing administration, as known to the SD infrastructure. In the SD Handler's network, the GUAN (Global UAN), which is composed of owner-name and UAN, is used to uniquely specify the data of interest. The activity is defined in 4.1.2

The UAN tag contains one or more items, each of which maps into a data handle. There are a variety of different types of handles, depending on the nature of the realm. For example, when data item is from a file system structure, then the data handle would be a file or directory name.

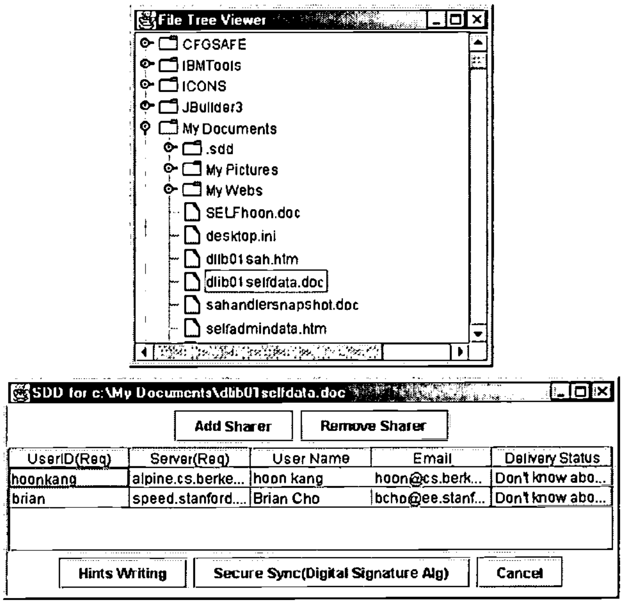
4. Implementation: Self-administering Data Handler (SD Handler)

We have built a prototype of the SD Handler in Java. We have modularized the basic tier functionalities and have built the GUI for desktop environment. We have used the Java's built-in security library for public/private key management and its encryption/decryption. In our experimentation, we have focused on co-authoring as our initial target application.

4.1 SDD Viewer/Editor (Bottling)

4.1.1 From Legacy File System

In order to bring data into SD Handler's world, we need to create its SDD first. As shown in the top of Figure 7, the user can browse the file or directory through a File Tree Viewer. If the user places a cursor on a file or directory that has previous SDD, the SDD Viewer/Editor shot will pop up, as shown in the bottom of Figure 7. If this is a new SDD, an empty SDD Viewer/Editor pops up. The user can add or delete sharers and can specify the preferences. (In the current prototype, the preference fields for strong serializing server, store-and-forward server, back-up repository server are associated per SD Handler daemon process, not per SDD instance.) Then, the user can synch out the associated data, using DSA (Digital Signature Algorithm), to every participant: The SD Handler's bottling functionality wraps the content with SDD and encrypt with the sender (owner)'s private key and the recipient's public key. The owner is asked to type in her pass-phrase to authenticate the use of her own private key. The delivery status is recorded.



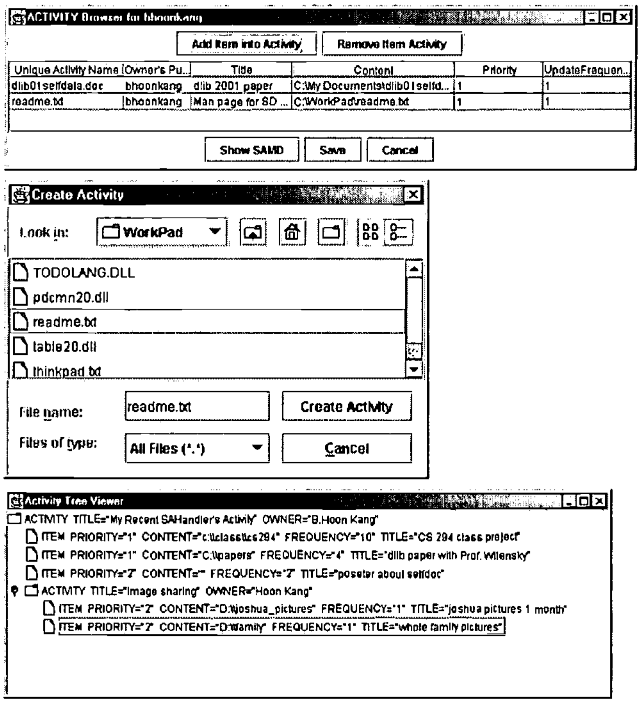
UserID(Req)	Server(Req)	User Name	Email	Delivery Status
hoonkang	alpine.cs.berke...	hoon kang	hoon@cs.berk...	Don't know abo...
brian	speed.stanford....	Brian Cho	bcho@ee.stanf...	Don't know abo...

Figure 7: SDD Viewer/Editor from File Tree Browser

4.1.2 From Activity Browser

We have found that it is cumbersome to remember to go to the file or directory every time we want to use SD Handler's bottling service. Therefore, we have introduced the *activity* as a mnemonic reminder for the collection of items that share the same SDD. Each item is a representative of data unit such as a file and an image. Thus, the activity is a unit of SDD association. A new activity can be created explicitly from an activity browser and implicitly from file tree browser. A new activity named as the data handle is created implicitly when a SDD is associated with data file from the file tree browser. Later, the user can find the SDD using either the activity browser or the file tree browser. In order to provide the most active activity item in the first page of the activity browser window, the activities are sorted according to their number of accesses and by the priority that it is given by the user.

We also prototyped the activity browser tree to provide hierarchical activity management, however, the complexity of tree management to the user seems to outweigh the benefit of SDD inheritance in a hierarchical activity management. The activity interactions are illustrated in Figure 8.



Unique Activity Name	Owner's Pu...	Title	Content	Priority	UpdateFrequen...
dlib01selfdata.doc	bhoonkang	dlib 2001 paper	C:\My Documents\dlib01selfd...	1	1
readme.txt	bhoonkang	Man page for SD ...	C:\WorkPad\readme.txt	1	1

TODOLANG.DLL
pdcmn20.dll
readme.txt
table20.dll
thinkpad.txt

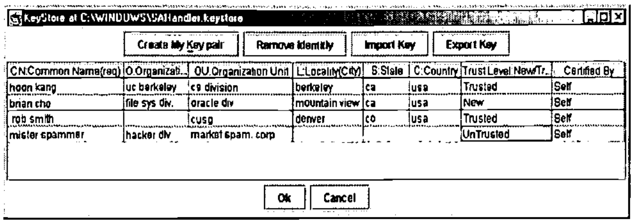
File name: readme.txt

```
ACTIVITY TITLE="My Recent SAHandler's Activity" OWNER="B.Hoon Kang"
    ITEM PRIORITY="1" CONTENT="c:\\class\\cs294" FREQUENCY="10" TITLE="CS 294 class project"
    ITEM PRIORITY="1" CONTENT="C:\\papers" FREQUENCY="4" TITLE="dlib paper with Prof. Wilensky"
    ITEM PRIORITY="2" CONTENT="" FREQUENCY="2" TITLE="poseter about selfdoc"
ACTIVITY TITLE="Image sharing" OWNER="Hoon Kang"
        ITEM PRIORITY="2" CONTENT="D:\\joshua_pictures" FREQUENCY="1" TITLE="joshua pictures 1 month"
        ITEM PRIORITY="2" CONTENT="D:\\family" FREQUENCY="1" TITLE="whole family pictures"
```

Figure 8: Activity Browsers: flat(top), hierarchical(bottom)

4.2 Public/Private Key-Store Manager

The Java framework provides a keystore architecture and a command line interface, keytool. We provide our own public/private keystore manager to store other collaborator's public key and trust level. (See Figure 9.) The trust level "new" means that the associated public key has never been added to the owner's keystore before and the owner of keystore has to decide whether to accept the key or not. If the owner accepts the public key with new status, the status is changed to "trusted". If the owner denies the public key of a collaborator, the status is set to "untrusted" and further contact from this source is denied at the earliest stage in the SDDP protocol stack.



CN:Common Name(req)	O:Organizati...	OU:Organization Unit	L:Locality(City)	S:State	C:Country	Trust Level New/Tr...	Certified By
hoon kang	uc berkeley	cs division	berkeley	ca	usa	Trusted	Self
brian cho	file sys div.	oracle div	mountain view	ca	usa	New	Self
rob smith		cusg	denver	co	usa	Trusted	Self
mister spammer	hacker div	market spam. corp				UnTrusted	Self

Figure 9: Key Store Manager

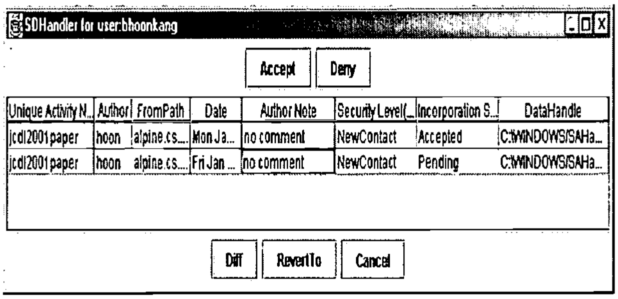
4.3 SD Sender/Receiver (Floating)

Our current prototype uses SDDP to deliver the bottled data directly between SD Handlers. The desktop version of the SD Handler has both the SD Sender and the threaded SD Receiver. We designed the SD Handler to minimally use the centralized highly available servers for common P2P infrastructure services such as store-and-forward delivery, naming/tracking the current location (IP address) of SD Handler, and the strong serialization of updates. The infrastructure services are not being shared among SD Handlers across administrative domains. Rather, each SD Handler can subscribe to its own infrastructure services.

If they don't use common P2P infrastructure services, the SD Receiver has to be available to receive a data delivery from the sending SD Sender. However, we have come up with a simple solution where the SD Handler of the initial creator of the SDD is used as an infrastructure service point for the duration of the interactions unless one of the participant's SD Handlers specified in SDD provides the infrastructure services for their interactions. For example, if the recipient's SD Receiver does not have store-and-forward delivery service, it can contact the creator of its SDD for the latest logged copy that it might have missed. As a same token, each participant can register its current IP location to the creator of its SDD.

4.4 Receiver Log Viewer (Popping/Logging)

When self-administered data is sent to a receiving SD Handler, the receiver log viewer (See Figure 10.) records the receiving activity, noting its UAN (Unique Activity Name), author (sender), date, and author's note. If the sender is a new/trusted contact, the receiver first creates a versioned copy and then records the data handle to it, then the incorporation status is set to "Pending". The owner can accept or deny the incorporation of this versioned copy into his own realm. The incorporation status shows whether the current logged activity has been accepted or denied. If the incorporation is successful, the status will be set to "Accepted", if not it will be set to "Error". If the user distrusts the sender, then the sender's public key is registered into the owner's keystore as "Untrusted". If the user denies the incorporation entry, then only the associated update is denied and the incorporation status is set to "Denied".



Unique Activity N...	Author	FromPath	Date	Author Note	Security Level(...	Incorporation S...	DataHandle
jcdl2001paper	hoon	alpine.cs....	Mon Ja...	no comment	NewContact	Accepted	C:\WINDOWS/SAHa...
jcdl2001paper	hoon	alpine.cs....	Fri Jan ...	no comment	NewContact	Pending	C:\WINDOWS/SAHa...

Figure 10: Receiver Log Viewer

Table 1: Comparison of SD Handler with Legacy Applications

Desired properties	Email (SMTP)	CVS/ Web-DAV	FTP	ICQ/ AIM	Groove	SD Handler
P1	No	No	No	No	No	Yes
P2	Yes	No	Yes	Yes	Yes	Yes
P3	Yes	No	No	Yes	Yes	Yes
P4	Yes	No	Yes	Yes	Yes	Yes
P5	No	Yes	No	No	No	Yes
P6	No	No	No	No	No	Yes
P7	Yes	Yes	Yes	Yes	Yes	Yes
P to P	Yes	No	No	Yes	Yes	Yes
Goal	Message exchange	Central version	File transfer	Instant message	P2P services	Self-admin

5. Comparisons

5.1 Comparison to "Legacy" Applications

We compared SD Handler with Email, CVS, and FTP, according to the seven desiderata listed in 2.1. A summary of the comparisons is given in Table 1.

P1. No repeat user involvement in routine data management

All of the applications except SD Handler require repeated user involvement in copying, moving, and sending the data. SD Handler requires SDD creation once for repeated usages.

P2. No unnecessary dependence on shared resources, such as shared data repositories or file servers

Email does not require shared resources for collaboration; CVS require a shared server location.

P3. No prior administrative set up costs

Both CVS and FTP provide the password-controlled access to the data that is being shared among collaborators. Either group account or individual account needs to be set up by an administrator, and need to be distributed to each collaborator to access the data prior to the collaboration. In Email or ICQ [14] or AIM [16], however, the password is not required to send or receive the message and its attached data. The access is purely controlled by the user's discretion whether to accept or refuse the attachment. An orthogonal end-to-end security method, for example, PGP[9] email, could be added. Both the SD Handler and Groove[15] provide public/private key based access control to the data without requiring prior administrative account set up. The user's discretion is guided by the key issuer's certificate or web-of-certificates.

P4. Ability to exploit minimal use of central server as only required

ICQ provides this property, so as to be scalable when their central server is contacted for name resolution of recipient's current IP address and store-and-forward data delivery to the unavailable recipient. By this measure, web-based file sharing systems over-utilize their central server in terms of the network bandwidth, processing power and disk space.

P5. Undo/Redo capability within user's domain

Only CVS support this.

P6. Secure and safe incorporation of updates at user's domain

Email could use DSA (Digital Signature Algorithm) for end-to-end security but the incorporation of email attachment is not sandboxed. ICQ and FTP do not provide safe-guarded incorporation either. CVS's undo capability could provide a safe incorporation since one can go back to the previous change in the case of an incorporation error.

P7. Lightweight enough to be widely deployed

All the applications above are considered to be lightweight since they do not require a heavyweight server infrastructure.

5.2 Declarative vs. Session-Based Data Management

FTP, NFS, HTTP, and derivative applications (e.g. WebDAV[7]) require a session with a resource controlling a data object in order to create, update, move, delete, or otherwise manage that object. Moreover, during this session, the data are managed by procedural commands. Network file systems, e.g., AFS and NFS, basically provide file semantics in sharing data, so, once again, intentions are expressed procedurally. Ficus [5] , Bayou (peer-to-peer optimistic file replication [16]), and Rumor (user-level replication system [19]) use file sharing semantics, and hence are fundamentally procedural as well. Also the overwriting semantics of file systems does not provide the knowledge about who made which changes. Hence one would have to use versioning software like CVS[8] in an explicit way, requiring the user's involvement in setting up check-in, check-out and copying.

In contrast, the SD Handler model provides a declarative way of managing data across administrative domains in a wide area scale. The SD Handler model also enables the user to specify that the data needs to be versioned at the different administrative domains. We believe this model can simplify data management, achieving our goals of minimizing the user's participation in routine tasks.

5.3 Scripted Email Attachment

A SD Handler can perform incorporation of received data into the recipient's internal data storage. A similar effect can be achieved by running a script (VBScript, UNIX shell script) with an email attachment. However, as is well-known, doing so is dangerous since the script can run any arbitrary command. However, the SD Handler's incorporation operation is sandboxed within SD Handler's address boundary where the access is limited only through the sanitized SD Handler's incorporation functionality.

5.4 P2P (Peer to Peer) Systems

ICQ [14], AIM [16] provide a peer to peer instant messaging with infrastructure services such as identity (user account) management, store-and-forward delivery and dynamic mapping of user's current IP address. We have found that these infrastructure services are common to most P2P systems. For example, Groove [15] provides collaborative P2P software tools with just these infrastructure services. The "shared space" in Groove provides an interactive collaborative environment where various applications (tools) can be built upon such as instant messaging, file sharing, free form drawing, and real-time conversation. However, unlike SD Handler, the management of data still requires repeated user interactions. The delivered files (attachments) have to be manually downloaded and saved. Versioning and logging are not provided since the incorporation of data is not automated but depends on manual end-user commands. Moreover, Groove and ICQ/AIM assume each peer to be an end user; in SD Handler the peer could be a personal repository server, a back up server, and a device in addition to other desktop users. Finally, Groove is focused on building a collaboratively shared space (or workspace) in a P2P way. SD Handler is focused on providing new semantics and controls for managing data with minimal user interactions. The co-authoring application is an example of using self-administering data model in a collaborative scenario.

6. Discussion

We have not yet had enough experience with our implementation to draw any forceful conclusions. However, we keep discovering new applications for self-administering data as we progress. For example, with an appropriate SDD, mirrored copies of documents can easily be made, in effect implementing a peer-to-peer RAID, or eliminating the need for tape backups. Doing so makes sense, as the loaded cost of tape backup is probably about one order of magnitude greater than the cost of low-end disk on a PC.

Self-administering data seem especially useful for data incorporation. For example, as in our Personal Library service [20], users may want to place a document in a scanner, and have it incorporated in their repository and added to a collection they maintain. Providing a SD Handler at the scanner provides a simple means to accomplish this goal.

Self-administering data can be used to update data managed by particular remote programs. Consider contact information inside a remote program. One can have one's contact information in an object whose SDD instructs that it be sent to one's contact list upon modification, and that the remote "contact update" program be executed. The remote user would need to specify what such a program is. The result, though, is that everyone's contact information for an individual can in effect be edited by the individual to whom it refers.

Self-administering data can provide an alternative to traditional digital library functions. Documents can be made available to a community of users by copying, rather than via repositories. Again, doing so may make sense, depending upon the cost of servers versus the cost of low-end machines, and the kind of availability one desires.

On the other hand, one-to-many communication is costly in this model, as we currently encrypt each communication on a per user basis. One possibility is to provide an option for users to trade off security for communication efficiency, in the case of one-to-many transactions.

7. Acknowledgements

This research was supported by the Digital Libraries Initiative, under grant NSF CA98-17353.

8. References

[1] Randy Katz, et al. The Endeavour Expedition: Charting the Fluid Information Utility, http://endeavour.cs.berkeley.edu/proposal/

[2] John Kubiatowicz, et al, OceanStore: An Architecture for Global-Scale Persistent Storage, Proceeedings of the Ninth international Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2000), 2000. http://oceanstore.cs.berkeley.edu/

[3] K. Petersen, M. J. Spreitzer, D. B. Terry, M. M. Theimer, and A. J. Demers. Proceedings of the 16th ACM Symposium on Operating Systems Principles (SOSP-16), Saint Malo, France, 1997, http://www.parc.xerox.com/csl/projects/bayou/.

[4] R. Bentley, W. Appelt, U. Busbach, E. Hinrichs, D. Kerr, S. Sikkel, J. Trevor, G. Woetzel, Basic Support for Cooperative Work on the World Wide Web. International Journal of Human-Computer Studies, 46(6), 1997. http://bscw.gmd.de/index.html

[5] T.W. Page, Jr et al, Perspectives on Optimistically Replicated, Peer-to-Peer Filing, Software Practice and Experience, v.28, n.2, February, 1998, http://ficus-www.cs.ucla.edu/travler/ficus_summary.html

[6] Ken Pier, Eric A. Bier, Ken Fishkin, Maureen Stone WebEdit: Shared Editing in a Web Browser. WWW4 Poster Proceedings, 1995. http://www.parc.xerox.com/istl/groups/gir/doc/webedit/webedext.htm.

[7] Jim Whitehead, Collaborative Authoring on the Web: Introducing WebDAV, Bulletin of the American Society for Information Science, Vol. 25, No.1,1998, http://www.webdav.org/papers/

[8] CVS (Concurrent Versions System), http://www.cvshome.org/

[9] PGP (Pretty Good Privacy), http://www.pgpi.org/

[10] Lotus Notes, http://www.lotus.com/

[11] Xerox Docushare, http://www.xerox.com/

[12] I-drive, http://www.idrive.com/

[13] FusionOne, http://www.fusionone.com/

[14] ICQ, http://www.icq.com/

[15] Groove Networks, http://www.groove.net

[16] AIM, http://www.aim.com/

[17] Desktop, http://www.desktop.com/

[18] Hotoffice, http://www.hotoffice.com/

[19] Peter Reiher, Michael Gunter, Gerald Popek, Rumor: A User-Level File Replication Middleware Service, http://fmg-www.cs.ucla.edu/

[20] Robert Wilensky, Personal Libraries: Collection Management as a Tool for Lightweight Personal and Group Document Management (forthcoming).

PERSIVAL, a System for Personalized Search and Summarization over Multimedia Healthcare Information

Kathleen R. McKeown*, Shih-Fu Chang†, James Cimino‡, Steven K. Feiner*, Carol Friedman‡, Luis Gravano*, Vasileios Hatzivassiloglou*, Steven Johnson‡, Desmond A. Jordan**, Judith L. Klavans††, André Kushniruk‡‡, Vimla Patel‡, and Simone Teufel*

*Department of Computer Science, Columbia University, New York, NY 10027, USA
†Department of Electrical Engineering, Columbia University, New York, NY 10027, USA
‡Department of Medical Informatics, Columbia University, New York, NY 10032, USA
**Department of Anesthesiology and Department of Medical Informatics, Columbia University, New York, NY 10032, USA
††Center for Research on Information Access, Columbia University, NY 10027, USA
‡‡Department of Mathematics and Statistics, York University, Toronto, Ontario, CANADA

Contact email: kathy@cs.columbia.edu

ABSTRACT

In healthcare settings, patients need access to online information that can help them understand their medical situation. Physicians need information that is clinically relevant to an individual patient. In this paper, we present our progress on developing a system, PERSIVAL, that is designed to provide personalized access to a distributed patient care digital library. Using the secure, online patient records at New York Presbyterian Hospital as a user model, PERSIVAL's components tailor search, presentation and summarization of online multimedia information to both patients and healthcare providers.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.5.2 [**HCI**]: User Interfaces; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval —*Information filtering, Query Formulation*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-base services*; H.3.5 [**Information Storage and Retrieval**]: Digital Libraries—*Dissemination, Systems issues, User issues*

Keywords

Medical digital library, personalization, search, query interface, multimedia, summarization, natural language

JCDL'01, June 24–28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

1. INTRODUCTION

In healthcare settings, both consumers and their providers need quick and easy access to a wide range of online resources. Patients and their family members need information that can educate them about their personal situations, while physicians need information that is clinically relevant to an individual patient. But unless online search and presentation of results is personalized to the individual patient, end users can be overloaded with far more information than is useful. Providing patient-specific information can have enormous benefits. When the latest medical information is provided at the point of patient care, it can help practicing clinicians and physicians in training to avoid missed diagnoses, choose only effective interventions and diagnostic tests, and minimize impending complications. The latest medical information, when expressed in understandable terms, can empower patients to take charge of their health, take appropriate preventive measures and make more informed choices regarding their treatment.

In this paper, we report on the ongoing development of PERSIVAL (PErsonalized Retrieval and Summarization of Image, Video And Language), a system designed to provide personalized access to a distributed digital library of medical literature and consumer health information. Our goal for PERSIVAL is to tailor search, presentation, and summarization of online multimedia information to the end user, whether patient or healthcare provider. PERSIVAL utilizes the online patient records at New York Presbyterian Hospital [7] as a sophisticated, pre-existing user model that can aid in predicting user interests. For the healthcare provider, our approach facilitates finding literature relevant to the specific patient under her care; for the healthcare consumer, our approach facilitates finding and understanding information relevant to his medical situation.

In the remainder of this paper, we first present the system's architecture and then discuss the components that we are developing. These include a user query component, which allows inference of meaningful questions given the

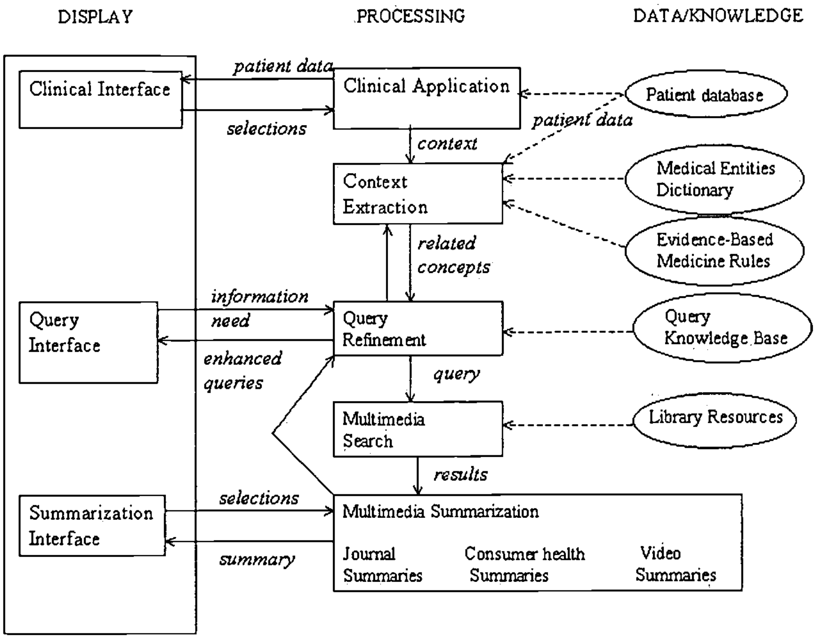


Figure 1: PERSIVAL's system architecture

clinical context, providing relevant information from the context for search; search over heterogeneous, distributed sources, re-ranking results by matching against the patient record; presentation of results to highlight information that is relevant to the patient, including summarization of textual and video resources along with definitions of unfamiliar terms; and interaction with the results through a highly interactive, multimodal, web-based thin-client architecture. Finally, we have begun preliminary experiments with end users that can provide specific information on how to personalize search and summarization.

2. SYSTEM ARCHITECTURE

The online clinical information infrastructure at New York Presbyterian Hospital provides many applications for viewing patient data. One of these, WebCIS [16], allows physicians to view patient records and another, PatCIS [3], allows patients to view information from their record. Patients may have questions about the meaning of different concepts (e.g., "What is unstable angina?") while physicians may have questions about possible treatments and diagnoses (e.g., "What are the risk factors for unstable angina?"). Thus, one way for users to access PERSIVAL is from within one of the above clinical applications, which trigger PERSIVAL to extract content from the patient record. Patients and physicians may also access PERSIVAL when questions arise at other times. For example, a physician, resident or medical student may have a question about a patient during the course of normal tasks (e.g., during rounds), in which case she must first provide the patient's identifying information and go through a secure clinical interface before accessing PERSIVAL, thus again giving the system the opportunity to associate a question with patient data.

The system architecture for PERSIVAL is shown in Figure 1. On the basis of data extracted from the patient record, PERSIVAL will generate questions that are meaningful in the clinical context and determine which concepts from the patient record are relevant for the search. The query, augmented with concepts from the patient record, is passed to the search component, triggering both textual and multimedia search. Textual search requires determining which of various distributed resources (some of which may only be accessible through a remote, online interface) may contain relevant documents. The results of this initial search are then fed through a component which more closely examines the retrieved documents by extracting and matching terms from the patient record, and by re-ranking the documents which are more relevant to this patient's case. Multimedia search may initially be triggered by a concept from the patient's record (e.g., "valve regurgitation") which matches diagnostic videos (e.g., echocardiograms), but PERSIVAL also allows follow-up search using image features representing important diagnostic information. The results of the search are passed on to a summarization component. For text documents, different approaches are used depending on whether the end user is a clinician or a patient. For the clinician, an informative summary indicates which results in the retrieved journal articles are relevant to the patient, while for the patient, a mixture of informative and indicative information is used to allow the patient to browse retrieved consumer health information. For video documents, PERSIVAL segments the video and creates a storyboard summary. A layout component will arrange and present the results to the user using a thin client server which allows for the intensive computation to be done at the server, incrementally passing results to the client as they become available. After results are presented, the user ultimately will have the option of refining his query or issuing a new search, though we do not report on this here. We are currently developing an XML interface to connect the components, which at this point in time have been developed separately.

For most of the above mentioned modules, we have identified major challenges, developed initial solutions to some of these challenges, and performed preliminary evaluations. We report in detail our approach to each of these subproblems and the status of current implementation in the sections that follow. Some of the modules, such as the layout planning module and the manager of interactions and feedback between modules are in an early prototype or design stage, and for those we report on our planned approach.

One contribution of our work is a remote application execution infrastructure that we call *Remote JAVA Foundation Classes*. Our infrastructure is built upon standard JAVA technologies (Java Foundation Classes (JFC) and Remote Method Invocation (RMI)) and supports the delivery of a highly interactive user interface over a low-bandwidth network connection. We take a hybrid approach between fat and thin clients that provides the benefits of a thin-client approach (e.g., low-cost terminals and centralized management and security), while offering functionality typically only found in fat clients (e.g., asynchronous server events and multimodal interaction). This is accomplished by delivering a user-interface-toolkit–aware client via HTTP to the web browser, which uses the RMI registry to obtain a remote reference to the application server. Once the client is registered, the server uses RMI to transmit JFC user interface toolkit commands to, and receive events from, the client. By manipulating user interface toolkit components and handling events remotely, our infrastructure provides functionality typically found in remote framebuffer systems (e.g., Tridia VNC [30], Symantec PCAnywhere (`http://www.symantec.com`), and Citrix Metaframe (`http://www.citrix.com`)), while consuming a fraction of the bandwidth.

3. ASKING QUESTIONS

Our approach features the ability to ask questions in context, where the context is drawn from the patient record. The patient record, however, is quite large, often consisting of hundreds of individual text, graph or video documents. It is critical that we extract information that represents information that the patient or physician is interested in and that is relevant to the query and the current clinical situation. Our approach will both allow the user to pose queries and will also automatically generate queries based on data that the end user is currently viewing, if the user indicates the need for information. The user also will ultimately have the option of entering a full question using a speech interface. In that case, we will use evidence-based medicine techniques [31] to determine which information from the record is relevant.

Previous work has examined technological solutions for linking medical records to online information resources. The MedLine Button [6] was the first system to use clinical data as input to real-time searches against a bibliographic database. In that system, patients' diagnoses and procedures were translated into Medical Subject Headings (MeSH) terms, assembled into search statements, and passed on to the MedLine search engine. That work has continued, in the form of *infobuttons* [4] that use a variety of clinical data and a variety of information resources to attempt to address the information needs of clinical information system users. A key part of providing automated searches is an understanding of the information needs that are most likely to arise in a given context. The MedLine Button generated questions that were drawn from a set of *generic queries* [5] which were, in turn, created from a database of actual questions posed to medical librarians. The questions were analyzed to determine recurring semantic patterns of concepts, such that specific concepts could be replaced by general semantic types. For example, "Do chest x-rays cause cancer?" becomes "Does <procedure> cause <disease>?" When a person reviewing a medical record selects, for example, a procedure and disease of interest, the system can then select the generic query that has such semantic types, and then instantiate it with the specific terms of interest. This prior work established a direct link between the data that the user was examining in the clinical application and potential questions that could be posed.

Our work on PERSIVAL makes significant enhancements in several areas: leveraging what is known about the individual patient whose record the clinician is reviewing; matching to an improved model of user information needs; and enabling the user to refine the query. The patient record contains a wealth of information that can be used to suggest and refine user information needs. For example, simply knowing the age or sex of the patient may help in focusing on pediatric or gynecologic searches. We are currently developing a set of rules founded on principles of evidence-based medicine is used to extract relevant facts from the patient record. These patient data are then matched against a new knowledge base of generic queries, which are based on questions asked by clinicians regarding patient care [11]. An enhanced interface enables the user to indicate the focus of the search, select an appropriate query, and refine the query as needed, by replacing or adding terms. A prototype of this enhanced interface and the component for matching patient data has been developed, while we are currently working on evidence-based medicine rules. A prototype link between the New York Presbyterian Hospital's clinical information system (WebCIS) is being created, specifically for a set of cardiology patients whose records have been sanitized. If the user is looking at a cardiology procedure, such as the 12-Lead Electrocardiogram, (which includes a description of abnormalities, such as lateral ischemia, unstable angina or left ventricular hypertrophy) and clicks on a link to PERSIVAL, the system extracts all of the findings from the report and displays them, in tabular fashion, to the user. The user can then select any number of these concepts for submission to PERSIVAL. For example, the user may select *lateral ischemia* and *unstable angina*. PERSIVAL also examines the patient's records for concepts relevant to lateral ischemia and if it finds, for example, that the patient is on the medication *aminophylline*, it will generate questions such as "What is unstable angina?", "What are the risk factors for lateral ischemia?", and "Can aminophylline cause unstable angina?". If the user selects the last of these questions, the system returns and issues the search query

```
User Question Analysis Summary
Type: pharmacologic-agent causes pathologic-finding
Clinical concepts: aminophylline, unstable angina
Search: (((aminophylline[mh]+OR+albuterol[tw])+AND+
    (unstable angina[mh]+OR+unstable angina[tw]))
  +AND+(adverse+[tw]+AND+effect[tw]))
```

When the user is examining a narrative report in the patient's record, information is extracted using a natural language information extraction system called MedLEE [12] that was initially developed for the domain of radiographic reports, and was subsequently extended to other domains. Primary findings are extracted as well as modifier relations. Modifier relations, such as negation, time, and uncertainty are frequently expressed in the reports, and are critical to obtain because they change the underlying meaning of the information. MedLEE contains several processing components: i) the preprocessor uses a lexicon containing semantic categories and target forms to segment the report into sections, sentences, words, and atomic phrases; ii) the parser structures the sentences by using a semantics-based phrase structure grammar to specify well-formed structures and their target forms; iii) a phrase regularization component composes phrases when appropriate; and iv) an encoding component maps the target terms to codes associated with a standard vocabulary. MedLEE was independently evaluated a number of times, and was shown to perform effectively for realistic clinical applications; in fact, it was not significantly different from medical experts in detecting specific conditions [17].

For PERSIVAL, MedLEE was extended to handle electrocardiogram and echocardiogram reports; we plan on further extensions in the cardiology domain. The extensions consisted of adding new entries to the lexicon, and of adjusting certain grammar rules. The encoding portion of MedLEE was also fine-tuned in order to generate output encoded into UMLS (Unified Medical Language System [18]) codes. These codes are critical because they form the basis for interoperability of two heterogeneous Natural Language Processing systems in this project: the patient record extraction system and the reranking of search results based on patient matching (Section 4.2). Following extraction of concepts by MedLEE, evidence-based rules are used to determine which of the many concepts are most important.

4. SEARCH

After a patient's or physician's query has been augmented with important information from the patient record, the augmented query is processed by a multimedia search component. PERSIVAL provides search tools over distributed collections of both text and echocardiogram video. For textual documents, search occurs in two stages. In the first stage, the system searches over multiple distributed text collections using a uniform metasearcher. In the second stage, results from the first are reranked by examining characteristics of patients studied in the article and matching those characteristics against concepts found in the patient record. For videos, the system searches relevant video segments from large collections by using automatically extracted annotation labels as well as clinically important multimedia features.

4.1 Searching and Browsing over Distributed Text Collections

Distributed resources available on the Internet do not present uniform searching capabilities, which complicates query processing. Furthermore, these resources differ widely in their topic and along every conceivable dimension. As a central component of PERSIVAL, we are deploying infrastructure for searching over distributed collections. Also, we have developed techniques for automatically classifying document collections into a topic categorization scheme that users can then browse to find the collections of interest.

PERSIVAL's query interface will offer the illusion of a single collection; the *metasearcher* is the component that enables a virtual integrated view of heterogeneous sources. To build our metasearcher we are merging two complementary existing search protocols that have been developed within digital library projects in the United States, namely SDLIP and STARTS, into a combined protocol, which we refer to as SDARTS [14].

SDLIP (Simple Digital Library Interoperability Protocol) [27], jointly developed by Stanford University, the University of California at Berkeley and at Santa Barbara, the San Diego Supercomputer Center, the California Digital Library, and others, defines a layered, uniform interface to search over each collection. SDLIP is a flexible, extensible protocol, and can "host" different query languages and metadata specifications for the sources to export. In particular, the requirements for the metadata interface are minimal, but extensions are allowed and encouraged. As a result, a perfect complement for SDLIP is STARTS (Stanford Protocol Proposal for Internet Retrieval and Search) [13]. STARTS is a high-level protocol that defines, among other things, the specific metadata that sources should export to facilitate metasearching.

SDLIP and STARTS complement each other naturally. At Columbia University, we have combined them into a protocol named SDARTS, which extends SDLIP by instantiating its query language and by defining a rich metadata interface according to what STARTS dictates. The result is a simple, expressive protocol that facilitates the construction of metasearchers [14]. In addition, we have developed a software toolkit to simplify the indexing of "local" document collections so that they are SDARTS compliant, as well as to simplify the construction of "wrappers" around external collections over which we do not have any control. SDARTS and its associated software toolkit provide the necessary infrastructure to incorporate collections into our project with minimal effort.

Metasearchers let users query over distributed collections. An alternative mode of interaction is for users to *browse* Yahoo!-like directories to locate collections of interest and then submit queries to these databases. Recently, commercial web sites have started to *manually* organize web-accessible text collections into hierarchical classification schemes (e.g., see InvisibleWeb at `http://www.invisibleweb.com`). Automating this classification is challenging, since many times the contents of searchable collections on the web are not available other than by querying. For example, consider the PubMed medical database from the National Library of Medicine, which stores medical bibliographic information and links to full-text journals accessible through the web. This database is accessible through a query interface at `http://www.ncbi.nlm.nih.gov/PubMed/`. If we query PubMed for documents with keyword `angina`, PubMed returns 36,150 matches, corresponding to high-quality citations to medical articles. The abstracts and citations are stored locally at the PubMed site and are not distributed over the web. Unfortunately, the high-quality contents of

PubMed are not "crawlable" by traditional search engines. A query on AltaVista for all the pages in the PubMed site with keyword "angina" (i.e., `angina host:www.ncbi.nlm.nih.gov`) returns only three matches. This example illustrates that often we cannot classify a valuable text collection by extracting and analyzing all the documents that it contains.

To automate the classification of searchable collections like PubMed, we have developed a novel technique that learns a small number of *query probes* to issue off-line to the collections [19]. We start with a comprehensive, predefined topic hierarchy with an associated training set of preclassified documents. We then characterize these documents by selecting the best features (i.e., words) using an information theoretic feature selection algorithm that eliminates the words that have the least impact on the class distribution of documents [21]. This step eliminates the features that either do not have enough discriminating power (i.e., words that are not strongly associated with one specific category) or features that are redundant given the presence of another feature. After this feature selection step, we train a rule-based document classifier [8] to produce rules like the following:

IF ibm AND computer THEN Computers
IF diabetes THEN Health
IF cancer AND lung THEN Health

For example, a document having the word "diabetes" will be classified into category "Health" according to this classifier. The next step is to transform each of these rules into query probes, and to adaptively issue the queries to the collections that we want to classify, extracting only the number of matches for each query. The number of documents that match a specific query at a database (e.g., "cancer AND lung") represents the number of documents that would match the corresponding classifier rule if we could run it over every document in the collection. Finally, our method classifies the collections using simply the number of query matches, without retrieving any documents from the collections. As a result, our strategy efficiently produces an accurate collection classification using a small number of query probes (typically fewer than 200 queries of a few words each are needed to classify a collection). Users can then browse the hierarchy of categories to identify the collections that match their information need.

4.2 Reranking Search Results

The query and search modules of our system allow the user to specify and adapt questions according to the clinical context and retrieve a wide variety of relevant documents from multiple sources. However, these modules are primarily *query-oriented*; while some patient information is used to direct the search, the primary focus of these modules is to include documents in the results that match the entered query.

Yet many of the documents that are generally relevant to a query about "unstable angina" may not be of high priority to a specific patient. For example, an article describing complications in patients who have both angina and diabetes should be ranked lower when the patient is not diabetic. On the other hand, given the sample patient record and article sentences shown in Figure 2, we can assume that the given article is very likely to be relevant to the patient.

Patient Record:
This is a 44 year old female past medical history of coronary artery disease, status **post myocardial infarction** in 1983, status post **CABG** in 1989 [...] The patient was admitted to New York Presbyterian Hospital on 12/3/99 [sic] a worsening CHF and **unstable angina** for evaluation for heart transplant.

Medical Article:
This was a multicenter prospective study of consecutive patients admitted to coronary care units with **unstable angina**. Baseline characteristics were age 60.18 ± 16 years, history of **prior myocardial infarction** in 336 patients (32%) [...] In-hospital treatment consisted of [...], angioplasty, or **coronary artery bypass grafting (CABG)** in 25.1% ...

Figure 2: Term Matches between Patient Record and Medical Article

PERSIVAL takes advantage of the patient record information to filter out documents that match the query well but the patient record poorly by reranking the results of the search according to how well they match with key elements of the patient record. To determine the degree of this match, more computationally expensive, natural language processing techniques are applied to the small portion of the entire set of the documents that match the query in the first place.

We base our comparison between patient records and medical journal articles on a common representation of both as lists of important technical terms, with associated values when they occur. Terms, that is words and phrases that capture technical content and have a fixed meaning within a specific domain, contain a large part of the information present in an article or patient record; demographics, diseases, treatment procedures, and drugs are all likely to be represented in the text via terms. In some cases, the term is associated with a value (e.g., "base heart rate over 90", where *base heart rate* is the technical term and *over 90* is the value). Representing both the patient record and the documents as vectors of term-value pairs provides a basis for converting document information into a form that can be used for quantitative comparisons. In particular, we represent terms as UMLS unique identifiers which capture the *semantic concepts* conveyed by the terms.

Patient records are analyzed by MedLEE as described in section 3; scientific articles, which employ more general language and less structure than patient records, are analyzed by the procedure described here. To find terms within scientific articles, we use a variety of surface indicators for each candidate term:

- Its relative frequency in medical and general text; terms are expected to be far more frequent in medical texts.
- Its distributional characteristics across different documents; terms usually bunch together more than ordinary words [1].
- Measures of cohesion between adjacent words help identify multiword terms; the component words of multiword terms and collocations occur together much more frequently than would be expected from their individual marginal frequencies [2].

- Syntax places constraints on terms; usually, terms consist of nouns, possibly premodified by adjectives and postmodified by a single prepositional phrase.

To collect this information from the text, we process it with tokenization and part of speech tools, as well as with a finite state grammar that enforces syntactic constraints on terms, expands invisible term connections from conjunctions (e.g., "unstable and stable angina pectoris" is a variant of "unstable angina pectoris") and associates terms with values by capturing attributive and predicative modifications between terms and numeric or adjective phrases (e.g., "*acute* myocardial infarction, *severe* unstable angina, systolic blood pressure *of 113.6*"). The numeric information from our statistical criteria (the first three indicators of terms above) is combined in a log-linear model [25], a supervised learning technique. By training on a list of established terms (the large scale vocabulary test (LSVT, [24]) we obtain the weights for the variables, producing a measure of how likely a word or phrase is to be a term.

After terms are identified, another algorithm performs the actual matching, measuring the overall importance of a term within both patient record and article. Two factors come into play: the relative specificity of a term and its semantic category, since more specific terms and terms that refer to diseases, treatments, and drugs are more likely to influence the matching. We use the semantic hierarchy in the UMLS to retrieve the semantic category, and, along with term frequency, to measure term specificity [28].

Once terms, values, and semantic categories have been obtained from both the patient record and the document, we calculate their degree of matching as the cosine product of these two vectors, with each term weighed according to the importance value assigned to it. In the example of Figure 2, "unstable angina" and "myocardial infarction" provide a large part of the matching score, since as a matched symptom and a matched disease, respectively, they contribute more than a matched theurapeutic procedure like "CABG". When values are present, a further matching step is executed, which alters the sign and magnitude of the match at that term according to how well the values match. Currently, we detect incompatibilities between values and partial matches between quantitative ranges of values.

The term recognition and matching modules are also used to provide anchor points for the two text summarization modules described below, and "topics" of summaries for the video module, in order for that module to assign labels to and access textbook examples and actual patient ECG videos.

4.3 Search and Organization of Echocardiogram Video

PERSIVAL also provides efficient tools for automatic indexing of echocardiogram videos and searching over large echochardiogram video collections. Echocardiography is an important imaging technique, which assists the cardiologist in the diagnosis of heart abnormalities. Because this method is non-invasive and cheap it is usually available in almost every major healthcare center. For example, at New York Presbyterian hospital there are thousands of echo videos taken and archived each year. However, very few tools and computer facilities are available for indexing and accessing such large video collections. Most echo videos are still stored on analog tapes with limited annotations.

In PERSIVAL, we envision that users will be able to access, search, and interact with digital echo videos efficiently and effectively. Video data will be integrated with other modalities of information and presented to the right users in the right context. For example, doctors, clinicians or medical students may retrieve echo videos of related cases with certain abnormalities from the library in order to compare with prior findings. Such facilities will be very useful in diagnosis, surgical planning, or teaching processes.

Echo video presents unique research challenges and opportunities for video indexing and summarization. Unlike other types of video addressed by existing work, echo video does not include speech, audio, or transcript information that can be used to index the video content. Information is predominantly contained in visual form. On the other hand, there are usually predictable structures in the production of echo videos. Sonographers usually follow a recommended sequence of transducer positions for capturing the two-dimensional echocardiograms. In addition, there is associated information from other modalities, such as ECG and diagnosis reports, which can be used in analyzing the video content or providing useful annotations.

To achieve these goals, our current research involves the following objectives and approaches:

Index video at the syntactic and semantic levels. Working with the domain specialists, we identified the syntactic structures and semantics of echo video. In particular, we developed a view transition model to represent rules used in the echo video scanning procedure. Characterized by the unique position and angle of the transducer, each view captures information about specific anatomical structures of the heart from a specific orientation. One of our objectives is to develop automated algorithms and tools for segmenting and recognizing constituent views in the video. To do this, we analyze unique spatio-temporal visual patterns of anatomic parts and apply a domain-specific view transition model. Results of the automatic tools that we have developed will allow users to randomly access views of interest, interactively browse constituent views at an intuitive level, and selectively transmit important views over networks.

Annotate and organize large collections of video. The video segments and summaries described above can be annotated by labels from view recognition and information contained in the diagnosis reports associated with each video. For example, descriptions of abnormalities related to certain parts (e.g., valves and muscles) can be linked to views in which such abnormalities are most visible. In addition, we are developing a taxonomy for classifying and organizing large collections of representative cases that can be used for research and teaching purposes.

Develop intuitive content-based video search tools. A query to the echo video library may be based on concepts in textual form or multimedia form. For example, terms describing specific abnormalities can be first used to retrieve specific segments of videos and their associated diagnostic findings. After seeing the returned videos (entirely or in a summary form), users may use graphic tools to select regions in the video and ask the system to find other videos showing similar visual patterns that are related to important clinical concepts (e.g., speed and volume of blood flows shown in the video). By combining such search tools using visual features with clinically meaningful categorization, users will be able to find more efficiently specific videos in large collections.

Currently, we have achieved promising results in view segmentation, recognition, and key frame extraction by using automatic image analysis algorithms and view transition models [9]. We are in the process of constructing a video library containing several hundreds of cases with important abnormalities, and evaluating our current tools.

5. PRESENTATION

PERSIVAL must formulate a concise and effective presentation that enables the user to understand the main points of the retrieved documents without having to examine them directly. It must also maintain links to the documents from the summary, allowing users to easily select the documents they judge most relevant. Our approach involves summarization of both text and video, including the ability to provide definitions for unknown terms. A layout component will integrate and cross-link search results, ultimately presenting summaries along with original documents for easy viewing and manipulation. Since PERSIVAL will be made available to end users on a variety of platforms, including low-end PCs, it is important that processing is efficient. We use a thin client server for this purpose.

5.1 Textual Summarization

A textual summary is generated to describe important information across the set of textual documents returned in a search. The method used to generate this multi-document summary depends in part on the document genre. Clinicians are more likely to be interested in seeing medical journal articles or textbooks that are relevant to the patient's case while patients will be more interested in consumer health information. These genres are quite different in how they are structured and thus, we use different techniques to produce a summary in each case. Furthermore, patients and clinicians are likely to be interested in different kinds of information, also requiring different processing. For both types of summaries, our approach involves a unique integration of statistical processing to select relevant phrases with symbolic processing to edit and weave phrases together to form the summary.

Summaries for Clinicians. Medical journal articles, of interest to clinicians, are written in a relatively rigid form, with sections (e.g., results) coming in a more or less predetermined order. Information extraction techniques [32] can take advantage of this structure to locate particular pieces of information. Experimental research articles typically present the outcome of a clinical study for several groups of patients (at least one test group and one control group). In a user study on summarization [26], we found that physicians can determine relevance of an article by quickly skimming the results and recommendations pertaining to the patient under their care. Thus, this is the information that should be included in a summary.

Our summarization module for clinicials [10] extracts this patient-specific information from a medical article. Structure is exploited to find the "results" section, then text categorization techniques are used to separate out sentences within the section that actually describe results. PERSIVAL has been trained to find words and phrases that are good indicators of results (e.g., "predictor"). This stage selects on average one third of the Results section, which is in turn only a portion of the full article. In the final stage of this component, we use pattern matching techniques to find particular types of results (e.g., "Multivariate analysis showed ... ") and select only those portions of the sentence that match the patient. The phrases which match information that is currently extracted for the query "What are the risk factors for unstable angina?" are shown in Figure 3; for example, the first sentence in that summary matches a patient who has atrial fibrillation. We are currently working on generating these target phrases from the extracted information. Following each sentence is a pointer to the article in which it was found. Ultimately, this information will be used by the layout component to directly link pieces of the summary to specific documents, enabling selective access to the related documents.

In a univariate analysis, NYHA class, pulmonary artery systolic, and atrial fib were associated with a decreased event free survival ([ajc]). But only NYHA class was considered as associated in a multivariate analysis ([ajc]). Prior angina was considered in both univariate and multivariate analysis a predictor of in-hospital morbidity ([1]). Atrial fib was not significant in multivariate analysis ([5]). The occurrence of angina after admission showed a strong univariate relation with the incidence of in-hospital acute MI or death ([5]).

Figure 3: Extracted summary phrases from results reported in journal articles

Summaries for patients. For patients, we summarize the set of consumer health documents that are determined to be relevant. In this case, we cannot assume that all patients will be interested in a specific type of information such as results. Instead, we provide information that is commonly repeated across documents and thus provides a synopsis of the set of documents. We follow this with "indicative" descriptions which characterize the kind of information contained within the documents, indicating which documents provide more detail on what topics and which documents are different from others in either content or form.

For the synopsis, we use a similarity tool that we developed for summarization of news [15]. Using statistical measures of pairwise similarity between sentences followed by clustering, it identifies sets of sentences across articles where each set describes similar information. From each set of similar sentences, it extracts one representative sentence to form part of the summary.

For the indicative part of the summary, the system uses a hierarchical representation of common topics found across all documents retrieved in the search. From this tree structure, it can determine the portion of the tree common to all documents, the different formats used, when a document presents more detail than all other documents, and when a document provides information that is not related to information presented in other documents. The summary that PERSIVAL generates in response to the search query "What is unstable angina?" is shown in Figure 4.

5.2 Explanations for Technical Terms

Currently, our summarization module uses terms and phrases that are found in the documents being summarized. However, some of these terms may not be familiar to patients. Ultimately, we want to be able to provide definitions for unfamiliar terms in the summary. To do this, we have

Treatment is designed to prevent or reduce ischemia and minimize symptoms. Angina that cannot be controlled by drugs and lifestyle changes may require surgery. Angina attacks usually last for only a few minutes and most can be relieved by rest. Most often, the discomfort occurs after strenuous physical activity or an emotional upset. A doctor diagnoses angina largely by a person's description of the symptoms. The underlying cause of angina requires careful medical treatment to prevent a heart attack. Not everyone with ischemia experiences angina. If you experience angina, try to stop the activity that precipitated the attack.

Highlighted differences between the documents include:

- This file (5 minute emergency medicine consult) is close in content to the summary
- The Merck manual of medical information contains extensive information on the topic.

Figure 4: Generated summary of documents retrieved for the query "What is unstable angina?"

developed a component of PERSIVAL that can identify and extract medical terminology, along with their definitions and modifiers, from reliable online resources such as the Heart Information Network (HIN, www.heartinfo.org/reviews).

In our study, we automatically analyze these resources in order to explore structural and linguistic methods for the identification and extraction of definitions and the terms they define, complementing our work on term extraction (see Section 4.2). This component of PERSIVAL, called DEFINDER (Definition Finder), uses rule-based techniques on text along with the Universal Medical Language System (UMLS) knowledge base. For the definition extraction, DEFINDER uses both shallow text processing (based on cue phrases, structural, and linguistic indicators) and a rich, dependency-oriented lexicalist grammar, the English Slot Grammar [23], for analyzing more complex linguistic phenomena. For example, our analyzer identifies that verapamil and diltiazem should both be included in the category of calcium channel blockers, from text such as: "Nifedipine is one of the three widely prescribed calcium channel blockers. The others are verapamil and diltiazem." In this example, the referring anaphoric phrase "the others" indicates that these three items belong in one category.

Our results show that medical texts for the popular audience, when of high quality, provide a valuable source of medical terminology and definitions. We performed two evaluations: 1) for the definition extraction method and 2) for the quality of defined terms. In the first case our system obtained 84% precision and 83% recall. For the second evaluation we choose a set of 93 terms and their definitions from our corpus and compare them with 3 other online dictionaries (including UMLS). The results presented in [20] show that the dictionaries appear to be incomplete (e.g only 60% of our term set are present and defined in UMLS; 24% of the terms are present but undefined; and 16% were absent altogether). A recent comparison between definitions in the UMLS and those automatically produced by DEFINDER shows that the latter are more useful and readable to lay people. Examples of our results include: (1) angina—the chest pain that occurs when the heart is deprived of oxygen due to diminished blood flow; (2) atrial fibrillation—improper contraction of the upper left chamber of the four-chambered heart; and (3) hypertension—high blood pressure. The output of our system can be used in the creation and enhancement of online terminological resources and in summarization.

5.3 Presentation and Summarization of Echo Video

For the video data, PERSIVAL provides efficient tools to present and summarize the retrieved videos. For example, from a patient's record, we know the patient was diagnosed to possibly have mitral valve regurgitation. Using this term as a query input, we retrieve video segments related to this abnormality from the patient's echo video or other videos in the library. After seeing the displayed videos, users may also use the content-based search tools described earlier to find more specific videos showing visual characteristics revealing important clinical information.

Depending on the context and needs, users may want to view returned videos at different lengths. Based on their inherent structures and semantics, echo videos can be summarized at different levels with different lengths. The first level includes presentation of key frames and associated data showing the most informative view of the heart in each segment. At the second level, each segment is represented by one complete heart cycle (clinical summary) that also shows the dynamics of the heart. At the third level, a highlight version of video can be produced by concatenating several short video clips each of which represents one single heart cycle from selected views. Such video highlights will be very useful for accessing the video library through bandwidth limited networks, such as wireless networks. Figure 5 shows a window including a key frame summary of selected video segments related to mitral valve.

5.4 Automated Layout

To help create a high quality user interface, we are developing methods for laying out automatically the material being presented, so that coherent, understandable transitions are employed as the presentation changes (e.g., by adding or deleting a display, or changing a display's contents). We are building on our previous research on automated generation and layout, which uses hierarchical decomposition planning techniques [33]. Unlike that earlier work, which generates all components on a single display, one challenge in PERSIVAL is to manage a set of displays, including some that are externally generated (e.g., the existing WebCIS and PatCIS systems). To determine a suitable approach, we surveyed previous work on automated user interface layout [22], and have begun to apply some of the best existing ideas to PERSIVAL. New directions include employing evaluation techniques to resolve inconsistencies in automatically generated constraints, and adding zoomable user interface components [29] to the set of rendering possibilities available to the automated layout system.

6. COGNITIVE STUDIES

At the same time as we develop the system, we are also carrying out formative evaluation that can help us determine how best to implement personalization. Evaluative work to date has focused on identifying how physicians assess relevancy of information in the context of particular patients, for tasks involving both searching (i.e. finding articles relevant to a patient) and summarization (i.e. extracting from an article information relevant to a patient). The objective

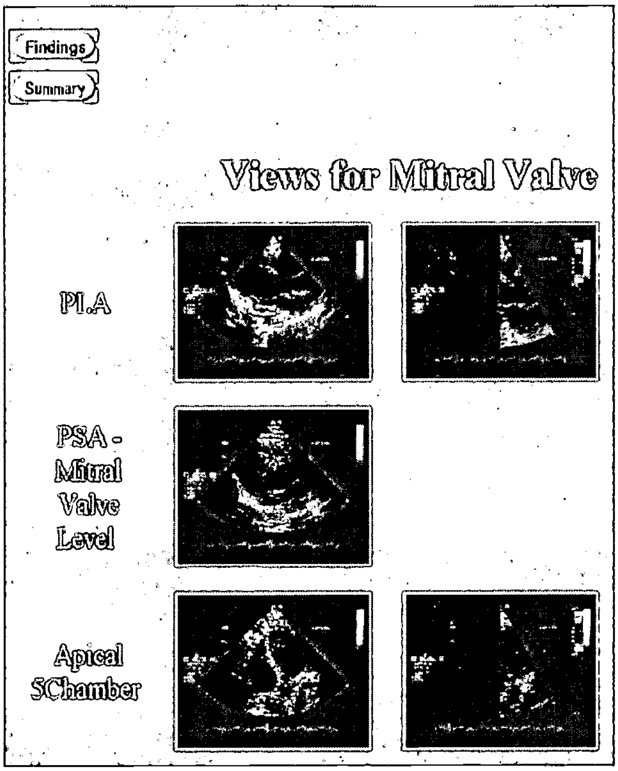


Figure 5: Echocardiogram Visual Summary

of this work has been to provide input to the design of system components, based on the empirical data, and to lay the groundwork for subsequent formative and summative evaluation of the corresponding PERSIVAL components.

In our first studies a test set of articles in the area of cardiology were collected. Other study materials included medical records from three cardiac patients, including electrocardiogram and echogram reports, as well as a written description of each patent. The study task involved having ten physicians review the information about the patients (one at a time) and then indicate whether each of the articles were relevant to care of the patient. Subjects were asked to "'think aloud" while doing this task and the audio recorded sessions were analyzed for strategies used in assessing relevance. Physicians were also asked to indicate which statements in the articles should be included in a summary for that patient. Results to date indicate that a number of different strategies were applied by subjects in scanning the articles and in deciding on their relevance. Differences in ratings were found to be related to a number of factors, including individual interests and level of physician expertise.

We are currently extending our study of relevance rating to a design where physician subjects are asked to rate relevancy of articles in the context of specific medical situations. The approach will later be extended to evaluation of PERSIVAL components in order to compare processing of information by physicians with that of automated system components. Plans are also being made for assessing the usability of other PERSIVAL components as prototype implementations become available, including visual summarization and presentation. Related ongoing evaluations involve audio recording and analysis of actual questions asked during intensive care rounds and assessment of information needs in naturalistic health care settings.

7. CONCLUSIONS AND CURRENT DIRECTIONS

We have shown how information from the patient record can be used to personalize the processes of search and summarization across multimedia information. To date, our work has focused on developing the components of PERSIVAL; we have developed a user query component that can use clinical context to help the user formulate meaningful queries and extract important information from the record, a distributed online multimedia search component that uses machine learning to find relevant sources and patient information to rerank articles, and a multimedia presentation component that uses patient information to determine relevance, automatically finds explanations of terms, uses segmentation and domain knowledge to summarize echocardiogram video, and ultimately will integrate this information in automated layout. Our next steps will be to develop interfaces between the currently separated components and automatically feed information from one stage to the next. We will also be focusing on further formative evaluation that can be used to improve personalization in PERSIVAL.

ACKNOWLEDGEMENTS

This paper is based upon work supported by the National Science Foundation under Digital Library Initiative Phase II Grant No. IIS-98-17434. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The authors would like to acknowledge the support of students involved in the project who have had many of the ideas and done most of the programming, including Kathy Dunn, Noemie Elhadad, Shahram Ebadollahi, Yun Ho, Panagiotis Ipeirotis, Simon Lok, Min-Yen Kan, Eneida Mendonça, Smaranda Muresan, and Sergey Sigelman.

REFERENCES

[1] K. W. Church. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, volume 1, pages 180–186, Saarbrücken, Germany, August 2000.

[2] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**:22–29, 1990.

[3] J. Cimino, J. Li, E. Mendonça, S. Sengupta, V. Patel, and A. Kushniruk. An evaluation of patient access to their electronic medical records via the world wide web. In *Proceedings of the Annual Fall Symposium of the American Medical Informatics Association*, pages 151–5, Los Angeles, CA, November 2000.

[4] J. J. Cimino. From data to knowledge through concept-oriented terminologies: Experience with the Medical Entities Dictionary. *Journal of the American Medical Informatics Association*, **7**(3):288–297, 2000.

[5] J. J. Cimino, A. Aguirre, S. B. Johnson, and P. Peng. Generic queries for meeting clinical needs. *Bulletin of the Medical Library Association*, **81**(2):195–206, 1993.

[6] J. J. Cimino, S. B. Johnson, A. Aguirre, N. Roderer, and P. D. Clayton. The MedLine button. In *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care*, pages 81–85, Baltimore, Maryland, November 1992.

[7] P. D. Clayton, R. V. Sideli, and S. Sengupta. Open architecture and integrated information at Columbia-Presbyterian Medical Center. *M.D. Computing*, **9**(5):297–303, 1992.

[8] W. W. Cohen. Learning trees and rules with set-valued features. In *Proceedings of AAAI'96*, 1996.

[9] S. Ebadollahi, S.-F. Chang, H. Wu, and S. Takoma. Indexing and summarization of echocardiogram videos. In *American College of Cardiology*, March 2001.

[10] N. Elhadad and K. R. McKeown. Towards generating patient specific summaries of medical articles. In *Proceedings of the NAACL Workshop on Automatic Summarization*, June 2001.

[11] J. W. Ely, J. A. Osheroff, M. H. Ebell, G. R. Bergus, B. T. Levy, M. L. Chambliss, and E. R. Evans. Analysis of questions asked by family doctors regarding patient care. *BMJ*, **319**(7206):358–361, 1999.

[12] C. Friedman, G. Hripcsak, W. DuMouchel, S. B. Johnson, and P. D. Clayton. Natural language processing in an operational clinical information system. *Natural Language Engineering*, **1**(1):83–108, 1995.

[13] L. Gravano, C.-C. K. Chang, H. Garcia-Molina, and A. Paepcke. STARTS: Stanford proposal for Internet meta-searching. In *Proceedings of the 1997 ACM International Conference on Management of Data (SIGMOD-97)*, May 1997.

[14] N. Green, P. G. Ipeirotis, and L. Gravano. SDLIP + STARTS = SDARTS: A protocol and toolkit for metasearching. In *Proceedings of the First ACM and IEEE Joint Conference on Digital Libraries (JCDL 2001)*, June 2001.

[15] V. Hatzivassiloglou, J. L. Klavans, and E. Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212, College Park, Maryland, June 1999.

[16] G. Hripcsak, J. Cimino, and S. Sengupta. WebCIS: Large scale deployment of a Web-based clinical information system. *Journal of the American Medical Informatics Association*, **6**:804–8, 1999.

[17] G. Hripcsak, C. Friedman, P. I. Alderson, W. DuMouchel, S. B. Johnson, and P. D. Clayton. Unlocking clinical data from narrative reports: A study of natural language processing. *Annals of Internal Medicine*, **122**(9):681–688, May 1995.

[18] B. Humphreys, D. Lindberg, H. Schoolman, and G. Barnett. The Unified Medical Language System: An informatics research collaboration. *Journal of the American Medical Informatics Association*, **5**, 1998.

[19] P. G. Ipeirotis, L. Gravano, and M. Sahami. Probe, count, and classify: Categorizing hidden Web databases. In *Proceedings of the 2001 ACM International Conference on Management of Data (SIGMOD 2001)*, May 2001.

[20] J. Klavans and S. Muresan. DEFINDER: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. In *Proceedings of the Annual Fall Symposium of the American Medical Informatics Association*, 2000.

[21] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML-97)*, pages 170–178, 1997.

[22] S. Lok and S. Feiner. A survey of automated layout techniques for information presentations. In *Proceedings of Smart Graphics 2001 (Int. Symp. on Smart Graphics)*, Hawthorne, NY, March 2001.

[23] M. McCord. English slot grammar. Technical report, IBM, 1990.

[24] A. T. McCray, M. L. Cheh, A. K. Bangalore, K. Rafei, A. M. Razi, G. Divita, and P. Z. Stavri. Conducting the NLM/AHCPR large scale vocabulary test: A distributed Internet-based experiment. In *Proceedings of the Annual AMIA Fall Symposium*, 1997.

[25] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

[26] K. R. McKeown, D. A. Jordan, and V. Hatzivassiloglou. Generating patient-specific summaries of online literature. In *Proceedings of the 1998 AAAI Spring Symposium on Intelligent Text Summarization*, pages 34–43, Stanford, California, March 1998.

[27] A. Paepcke, R. Brandriff, G. Janee, R. Larson, B. Ludaescher, S. Melnik, and S. Raghavan. Search middleware and the Simple Digital Library Interoperability Protocol. *D-Lib Magazine*, **6**(3), 2000.

[28] R. J. Passonneau, K. K. Kukich, J. Robin, V. Hatzivassiloglou, L. Lefkowitz, and H. Jing. Generating summaries of work flow diagrams. In *Proceedings of the International Conference on Natural Language Processing and Industrial Applications*, pages 204–210, New Brunswick, Canada, June 1996.

[29] K. Perlin and D. Fox. PAD: An alternative approach to the computer interface. In *Proc. SIGGRAPH '93*, pages 57–64, Anaheim, California, August 1993.

[30] T. Richardson, Q. Stafford-Fraser, K. Wood, and A. Hopper. Virtual network computing. *IEEE Internet Computing*, **2**:33–38, 1998.

[31] D. L. Sackett, R. B. Haynes, G. H. Guyatt, and P. Tugwell. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Little, Brown and Company, Boston and Toronto, 2nd edition, 1991.

[32] B. M. Sundheim, editor. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, San Mateo, California, 1996.

[33] M. Zhou and S. Feiner. Top-down hierarchical planning of coherent visual discourse. In *Proc. IUI '97 (1997 Int. Conf. on Intelligent User Interfaces)*, pages 129–136, Orlando, Florida, January 1997.

An Approach to Search for the Digital Library

Elaine G. Toms
Faculty of Information Studies
University of Toronto
416-978-7802
toms@fis.utoronto.ca

Joan C. Bartlett
Faculty of Information Studies
University of Toronto
416-978-4715
bartlett@fis.utoronto.ca

ABSTRACT

The chief form of accessing the content of a digital library (DL) is its search interface. While a DL needs an interface that integrates a range of options from search to browse to serendipity, in this work we focus on analytical search. We propose using Bates' search tactics as a basis for the re-design of search interfaces. We believe this approach will help to identify the types of tools that need to be supported by a DL interface.

Categories and Subject Descriptors

H..3.3 Information Search and Retrieval

General Terms

Design; Human Factors.

Keywords

Searching; Digital Libraries; Search Interface; Search Tactics.

1. INTRODUCTION

The public face of a digital library (DL) is its interface – the part that users see and manipulate, the part that handles the communication process between a user and the DL components and ultimately the part that contributes to the success or failure of a user's interaction with a DL. The user is both rewarded by the wealth of options and limited to what the interface allows. Although a DL may provide a range of services, from analytical search to browsing and serendipitous processes, we focus specifically on the search service in this aspect of our research.

At present searching is the preferred and most popular means of finding specific information on the WWW and sometimes the only means of accessing DL content. But, despite their popularity, search engines seem to be no more proficient in aiding the user to find what they need than online catalogues or text databases. The problem is partially concerned with query formulation, and with users' understanding of the task and the system's 'understanding' of how the user performs that task. The problem is best expressed in a playful definition: "search engines can be quite tiresome and not very fruitful if you don't know how to use them correctly."[8] We prefer however to approach the problem from the other perspective: *systems do not know what to provide for users.*

In DLs, searching is served by a WWW page that contains a pane(s) for entering a search expression(s), options for selecting search constraints on that expression, and sometimes user directions. However, the design of these pages follows conventional electronic form design practice, rather than being empirically based on how people search for information. This paper will propose an approach to the design of search interfaces that is based on prior research into search tasks and processes, and grounded in Bates [1]'s information search tactics.

2. SEARCH TASK

In general, a quest in a DL starts with an information problem that is conceptualized as a question. The user transforms that question into a query; the system returns a set of references from which the user assesses relevance and may revise the query and re-submit.

In this research we are concerned with the process from expression of need through to query generation and system response. The difficulty is in expressing in words what is in one's head. The process requires the user to translate need into words and format that are acceptable to the system. This process requires not only the selection of appropriate words, but also knowledge of system specific attributes, e.g., truncation, phrase specification.

Over the past four decades of information retrieval, users have interacted with systems using a variety of styles, most of which were dictated by the level of systems development of the day. From command line interfaces, to simple menu selection, to the graphical user interface, control has shifted from the system to the user, and the requisite skill level from expert-only to include the novice. The search interface in vogue today relies on an interface design technique – 'form filling' and DL search interfaces provide little beyond a single box and a search command button to support this process.

3. MODELS OF THE SEARCH PROCESS

To date multiple approaches to the search processes have been suggested [5, 7, 10, 11]. The search process tends to be discussed at a macro level, e.g., Kulthau [6] who provides only guidance and orientation on how that task can be implemented at the interface level. But, at the 'keystroke' level, the search task is procedural; commands are entered and responses received. We need an understanding of component steps used to perform the search task at that micro level [9].

Four procedural models offer insights into the search process from different disciplinary perspectives. Three of these are: 1) how students locate information in textbooks [5]; 2) a task analysis of information-seeking activities [7]; and, 3) an information systems framework [11]. These have comparable stages but at differing levels of granularity and have some elements of both problem-solving and analogical reasoning; that, is they include both the

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

encoding of a goal and stimulus, to compare features and to infer similarities.

The fourth is Norman [10]'s seven stages of human-computer interaction (HCI), considered one of the most influential models in HCI. It too is grounded in problem-solving theory, but describes a generic process that is applicable to each selection at the interface. Thus, for any single stage in the first three, there will be multiple renditions of the fourth, one for each human input that takes place within that stage. Norman's model is useful for assessing micro stages in the search models, e.g., what intention is taking place? what action sequence must be specified?

4. SEARCH AT THE MICRO LEVEL

While information search models provide generic views of the search process, none are sufficiently detailed to prescribe the interface components needed to support aspects of the search task. Another approach to the search process is that of Bates [1, 2, 3] who proposed 29 information search tactics which she defined as "move[s] made to further a search" – a heuristic meant to assist the meeting of short term goals and the procedures taken during the search process. She grouped these tactics into four categories:

a) Monitoring: "tactics to keep the search on track and efficient." These range from evaluating search formulation to spell checking.

b) File structure: "techniques for threading one's way through the file structure ... to the desired file, source, or information within source." These include tactics such as breaking a complex search down into smaller components, or using an indirect route to access information.

c) Formulation: "tactics to aid in the process of designing and re-designing the search formulation." These involve broadening and narrowing the search with search terms, and Boolean operators.

d) Term: "tactics to aid in the selection and revision of specific terms within the search formulation." These tactics relate to the selection of search terms, by exploiting the relationships between terms, such as broad and narrow concepts, or spelling variations.

Bates' tactics were derived from her analysis of search strategies for bibliographic systems and at that time were meant to aid the search process in both manual and online systems. Each tactic might include one or more moves. A move is more fine-grained and can be equated with a cycle of Norman [9]'s seven stage model. Despite the attention that these tactics have received, they remain a theoretical perspective, having never been tested. However, they can prescribe the sorts of tools that one needs at the interface level to support steps in the search process.

In addition to search tactics, Bates also defined idea tactics [2], which are meant to aid the thinking/creative process. These are most useful at the early stages of the search process during problem understanding. She also devised stratagems [3], which are broader, providing guidance for a specific class of search task, e.g., chasing footnotes. These tend to handle particular contexts.

Approaching the search process from the perspective of Bates' tactics replaces the emphasis on search functionality with one that focuses on the user's task: tools are needed to do the job? We are not alone in taking this approach. Fuhr [4] too suggests that Bates' four-level plan could be used to implement "strategic support of the information seeking process." We are merging Bates' tactics with a procedural model of the search process that is emerging from HCI and assessing the appropriateness of each tactic for each stage of that process. For each tactic we are identifying the type of interface tool that is needed to support that tactic. At the same time we are updating the tactics. For example, no tactic suggests the need for an overview of the digital library content, which has long been recognized as essential information for the user in resource selection [7].

5. CONCLUSIONS

We believe that Bates' tactics provide a useful approach with which to conceptualize and model the search interface. The tactics need updating and amplification to conform to systems development of today and they need testing. In addition, the tools that are derived from those tactics must be integrated into the DL interface so that they support not only the analytical search function but also all information exploration processes.

6. ACKNOWLEDGMENTS

This research is supported by a Connaught Fund Grant and NSERC grant to the first author.

7. REFERENCES

[1] Bates, M.J. Information search tactics. *Journal of the American Society for Information Science* 30 (4 1979), 205-214.

[2] Bates, M.J. Idea tactics. *Journal of the American Society for Information Science* 30(5 1979), 280-289.

[3] Bates, M.J. Where should the person stop and the information search interface start? *Information Processing and Management* 26 (5 1990), 575-591.

[4] Fuhr, N. Information Retrieval in Digital Libraries: Dealing with Structure, presented at the first DELOS Workshop on Information Seeking, Searching and Querying in Digital Libraries, (http://www_dbs.inf.ethz.ch/DELOS/invited.html

[5] Guthrie, J. T. Locating information in documents: examination of a cognitive model. *Reading Research Quarterly*, *23*(2 1988, 178-199.

[6] Kuhlthau, C. Inside the search process: information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5 1991), 361-71.

[7] Marchionini, G. *Information seeking in electronic environments*. New York: Cambridge University Press, 1995.

[8] Netlingo (URL: http://www.netlingo.com)

[9] Norman, K. L. Models of the mind and machine: information flow and control between humans and computers. *Advances in computers*, *32* (1991), 201-254.

[10] Norman, D. A. Cognitive engineering. In D. A. Norman & S. W. Draper (Eds.), *User centered system design: new perspectives on human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1986, 31-61.

[11] Shneiderman, B., Byrd, D & Croft, W.B. Sorting our searching: a user interface framework for text searches. *Communications of the ACM* 41 (4 1998), 95-98.

TilePic: A File Format for Tiled Hierarchical Data

Jeff Anderson-Lee
Electronics Research Laboratory
UC Berkeley
Berkeley, CA 94720-1776
1-510-643-6447
jonah@eecs.berkeley.edu

Robert Wilensky
Division of Computer Science
UC Berkeley
Berkeley, CA 94720-1776
1-510-642-7034
wilensky@cs.berkeley.edu

ABSTRACT

TilePic is a method for storing tiled data of arbitrary type in a hierarchical, indexed format for fast retrieval. It is useful for storing moderately large, static, spatial datasets in a manner that is suitable for panning and zooming over the data, especially in distributed applications. Because different data types may be stored in the same object, **TilePic** can support semantic zooming as well. It has proven suitable for a wide variety of applications involving the networked access and presentation of images, geographic data, and text. The **TilePic** format and its supporting tools are unencumbered, and available to all.

1. INTRODUCTION

The **TilePic** file format was designed in response to a specific, yet common, data management need. Many applications, especially networked applications, require access to and display of data objects that are too large to be handled as a single entity within the client. Because only a small portion of each dataset is actually visible at any one time at full resolution, it is possible to break the data into smaller "tiles", each of which can be loaded on demand. Also, since less detail is needed when zooming out, it is reasonable to create pre-zoomed subsets at a variety of scales.

TilePic is a file format designed to accommodate this need. It offers a way to encapsulate a large amount of related, static data into a single file. The data should represent a single, one or two-dimensional dataset, possibly at multiple scales of resolution or abstraction. The data should be divisible into equal sized chunks (tiles) based on x,y coordinates so that relevant, localized portions can be accessed quickly without having to look at all of the data. **TilePic** also supports the storage of data at multiple levels of resolution or abstraction. It does so by supporting multiple layers of tiles, where each layer is related to the next by the same (integral) scale factor, allowing applications to zoom in from a more abstracted and less detailed representation to a less abstracted and more detailed one.

This contrasts with trying to compress the data so that it will fit into the client, or using on-the-fly, server-side data abstraction services. These approaches are also valid for some applications.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

2. PREVIOUS WORK

There are several pre-existing alternatives to **TilePic**. Unfortunately, there were several characteristics that we desired, and none of the existing formats satisfied all of these.

First, we wanted to avoid tying the format to one particular type of data, e.g., *tiff* [12] or *jpeg* images. We foresaw the need to tile multiple kinds of data and wanted to support one form of tiling rather than several. This constraint eliminated from consideration most pre-existing image formats that support tiles and or multiple images, including FlashPix [4], GridPix [1], MrSID [6], *tiff*, and *gif*.

We also wanted quick, indexed access to individual tiles. This constraint ruled out stream-based archiving formats such as *tar* and *zip*, in which files must be searched for in an archive.

3. ADVANTAGES

TilePic is simpler because it just focuses on tiling. It stores the tile data plus fixed and user-definable metadata attributes in the file. The data for each tile is contiguous within the file for faster access. An index further assists fast retrieval of tile data.

While it is possible to realize some of **TilePic**'s benefits by on-the-fly computations, there are several advantages to creating and storing tiles in advance. In pre-computing the tiles, the computation can be done once and used multiple times. This makes the use of better quality abstractions such as pixel color-averaging feasible in place of the faster pixel sub-sampling strategy often used for on-the-fly image scaling. Moreover, storing several layers of tiles, each one-half the scale of the previous usually takes up only about one-third more space than the original data, which is an effective space-time trade-off.

One benefit of **TilePic** is that the abstraction scheme is not hardwired. This gives application designers more flexibility. When scaling a large image to create a smaller one, details like lines, symbols, and textual annotations can eventually become smaller and indistinct. Alternatively, if the line-width or typeface is kept constant while zooming out, the presentation soon becomes cluttered. Therefore, simple scaling is often less useful than using a new representation designed for the new resolution. One option is use semantic zooming [5], i.e., leave out less important objects at higher resolution, such as minor roads or landmarks on a map. Thus, rather than simply scaling maps, it is better to substitute different map products at different resolutions, whenever possible. **TilePic** readily accommodates doing so. This flexibility is very useful for applications such as GIS where the user may wish to zoom over many levels.

Using **TilePic** with vector data, many solutions are also possible on a case-by-case basis. It is possible to use semantic zooming as with raster maps, leaving out minor features. Another option is to use algorithms to reduce the detail by reducing the number of points, simplifying the lines and polygons as resolution is decreased [3]. Furthermore, with **TilePic** one can substitute another data type entirely, such as using a small image to cover a large region, and yet still have the benefit of precisely positioned data points at the higher resolutions, i.e., just when they are needed. When the tiles are pre-computed, it is easy to substitute alternative abstract representations for some layers. Furthermore, thanks to the tiling, only the areas of interest need to be sent to the client at the highest level of detail.

Packaging the data into conceptually meaningful units makes it easier to manage. For example, we have a large amount of tiled data, primarily images. Packaging the tiles for one image into a single file makes tasks such as retrieving them from an archival storage system much easier, as all the tiles corresponding to a single object can be transferred en-masse from the archive, rather than as thousands of separate files.

4. LIMITATIONS

TilePic requires that data can be tiled into equal-sized chunks, in one or two-dimensions. Another consideration is that it makes sense to abstract the data with integral scale changes between layers. However, single layer **TilePic** files can still be useful if the data may be tiled but does not need to be scaled.

TilePic was designed with read-only access to static data in mind. As such it provides no easy way to update the contents of a single tile. Instead, one must unpack the **TilePic** file, modify the data, and then generate a new **TilePic** file. For data that changes frequently, this approach would be costly.

Another limitation of the **TilePic** representation is one of maximum size. The format is limited to 4GB files by virtue of using 32-bit unsigned offsets in the metadata. Since many file systems impose a 1GB or 2GB limit on file sizes, we viewed this limitation as a reasonable design compromise.

5. TILEPIC TOOLS

We have made available [11] several tools for working with **TilePic** files. The basic **TilePic** utilities are fairly simple, are written in C, and require no additional libraries. However, most users may wish to use these in the context of specific typed tile data such as *png* and *jpeg* images and/or cgi-bin scripts. Hence, we have written a number of supporting programs and *perl* scripts to help with these applications. For the most part, these rely on additional software utilities and/or libraries that are freely available elsewhere.

The code was written for a Unix/C environment. We have tried to indicate all of the additional packages that may be required, along with one source where each may be found. In addition, we provide detailed documentation of the **TilePic** format, **TilePic** tools, a simple API in C that can be used by other programs, and sample applications. Interested users are advised to examine the GIS Viewers [10] developed by our project; these can be used either in conjunction with other geo-located data [9] or simply to display images [8], and fully exploit the **TilePic** format.

6. CONCLUSIONS

We have found **TilePic** to be an extremely useful, if simple idea. We have used it successfully in a number of collaborative applications such as with the Museum of Vertebrate Zoology [7] for displaying photographs and notebook pages of scientific and historical value and with CalFlora [2] for displaying base maps onto which botanical observation data is overlaid. The list of applications for it continues to grow, and to surprise us. We look forward to other developers finding additional applications and development ideas.

7. ACKNOWLEDGEMENTS

Many members of the UC Berkeley Digital Library Project contributed to the context from which **TilePic** emerged. Loretta Willis and Howard Foster played particularly substantial roles in the creation of related applications. Thanks also go to Satoshi Asami for his work on GridPix which helped to stimulate the work on **TilePic**. We would also like to thank Sun Microsystems, Microsoft Research, IBM, ADIC and Intel for their contributions of hardware and software to help make this research possible. This research was supported by the Digital Libraries Initiative, under grant NSF CA98-17353.

8. REFERENCES

[1] Asami, S. and Patterson, D.A., GridPix: Presenting Large Image Files Over the Internet, Report No. UCB/CSD-00-1099, May 2000, Computer Science Division, EECS, University of California at Berkeley, CA 94720-1776

[2] CalFlora, http://www.calflora.org/ .

[3] Douglas, D.H. and Peucker, T.K., 1973. "Algorithms for the Reduction of the Number of Points Required to Represent a Line or Its Character," The American Cartographer 10(2):112-123.

[4] Eastman Kodak, FlashPix Executive Summary, http://www.kodak.com/US/en/digital/flashPix/.

[5] Fox, David. Tabula Rasa: A Multi-scale User Interface System. PhD Thesis, Department of Computer Science, New York University, May 1998.

[6] LizardTech, LizardTech, Inc. - Imaging Software, http://www.lizardtech.com/index.html.

[7] Museum of Vertebrate Zoology, Museum of Vertebrate Zoology Data Access, http://dlp.cs.berkeley.edu/mvz/ .

[8] Museum of Vertebrate Zoology, Museum of Vertebrate Zoology Image Catalog: Great Gray Owl, http://elib.cs.berkeley.edu/gislite/examples/owl.html.

[9] The UC Berkeley Digital Library Project, CalFlora Occurrence Database example, http://elib.cs.berkeley.edu/gis/examples/test11.html.

[10] The UC Berkeley Digital Library Project, GIS Viewer 3.0, http://elib.cs.berkeley.edu/gis/.

[11] The UC Berkeley Digital Library Project, TilePic home page, http://elib.cs.berkeley.edu/tilepic/.

[12] Warmerdam, F. and Welles, M., TIFF Software, http://www.libtiff.org/.

Long Term Preservation of Digital Information

Raymond A. Lorie
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
001-408-9271720
lorie@almaden.ibm.com

ABSTRACT

The preservation of digital data for the long term presents a variety of challenges from technical to social and organizational. The technical challenge is to ensure that the information, generated today, can survive long term changes in storage media, devices and data formats. This paper presents a novel approach to the problem. It distinguishes between archiving of data files and archiving of programs (so that their behavior may be reenacted in the future).

For the archiving of a data file, the proposal consists of specifying the processing that needs to be performed on the data (as physically stored) in order to return the information to a future client (according to a logical view of the data). The process specification and the logical view definition are archived with the data.

For the archiving of a program behavior, the proposal consists of saving the original executable object code together with the specification of the processing that needs to be performed for each machine instruction of the original computer (emulation).

In both cases, the processing specification is based on a Universal Virtual Computer that is general, yet basic enough as to remain relevant in the future.

Categories and Subject Descriptors

E.7 [**Electronic Publishing**]: Long-term preservation of digital information.

General Terms

Standardization, Languages.

Keywords

Digital library, preservation, archival, digital information, digital documents, emulation.

1. INTRODUCTION

The problem that libraries are facing today is well known [1]. For centuries, paper has been used as the medium of choice for storing text and images. Today more than ever, some of the archived objects (books, newspapers, scientific papers, government and corporate documents, etc.) are in danger of becoming unreadable. The preferred solution is to digitize the documents and store the binary files in a digital library (DL). As a result, an object can be copied repeatedly without degradation and its content can be sent remotely and accessed at will. Also, the physical space needed to store the object becomes smaller and smaller as storage density increases.

Beside digitization, a high percentage of the data to be preserved is, today, generated directly in digital form. Spreadsheets, word processor documents, e-mail messages, as well as audio CD's or DVD's are obvious examples.

Whatever the origin of the digital information is, we are left with the same challenge: to ensure that the information can survive long term changes in storage media, devices, and data formats. An excellent introduction to this problem is given in [2].

2. THE PROBLEM

Suppose we use a computer (identified as M2000) to create and manipulate digital information today. For the purpose of archiving the data for preservation, the digital information is catalogued in a DL index; the content of the document may be stored in the DL itself, say in a file F2000. Suppose that, in 2100, a client wants to access the archived data. Assuming the catalog entry is still accessible and still refers to the document, three conditions must be met in order to recover its content:

1. F2000 must be physically intact (bit stream preservation)
2. A device must be available to read the bit stream.
3. The bit stream must be correctly interpreted.

Condition 1: some researchers predict very long lifetimes for certain types of media, but others are much less optimistic. Anyway, if a medium is good for N years, what do we do for N+1 years? Whatever N is, the problem does not go away. The only possible solution consists of rejuvenating the information periodically by copying it from the old medium onto a newer one.

Condition 2: machines that are technologically obsolete are hard to keep in working order for a long time. Actually, this condition is more stringent than the previous one. Here also, rejuvenation is needed: it moves the information onto a new medium that can be read by the latest generation of devices. Thus, conditions 1 and 2 go hand in hand. Note that rejuvenation is not an overhead incurred only for preservation; it is also a means of taking advantage of the latest storage technology.

Condition 3: the two conditions above insure that a bit stream saved today will be readable, as a bit stream, in the future. But one must be able to decode the bit stream to recover the information in all its meaning. This is quite a challenging problem, and this paper sketches a solution.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

3. THE TYPES OF DOCUMENTS

We distinguish three cases; by increasing order of complexity:

Case 1: The document is represented as a simple data structure. The rules to decode and understand the document can be explained in natural language, and saved in the DL. In 2100, a program can be written to decode the data[1].

Case 2: When the data structure and the encoding reach a certain level of complexity, it becomes impractical or even impossible to explain them in natural language. The alternative is to save a *program* that decodes the data; this may be the only way to be sure that the decoding is specified completely. The program is written in some language L. In 2100, the M2100 system must be able to interpret it.

Case 3: At the end of the spectrum, we may be interested in archiving a computer program or system for its own sake. In this case, it is the ability to run that program which must be preserved by the archiving mechanism. Not only the bit stream that constitutes the program must be archived, but we must also make sure that the code can be executed at restore time. Therefore, enough information on how to execute an M2000 code on an M2100 machine must also be archived.

The overall case analysis can be simplified by noting that case 1 can be handled by any method that handles case 2. In both cases 2 and 3, a program is being saved. The only – but very significant - difference is that, for case 3, the language in which the program is written must be the native M2000 machine language, while the language L, used for handling case 2, can be arbitrary. We refer to case 2 as *data archiving* and to case 3 as *program (or behavior) archiving.*

4. PREVIOUSLY PROPOSED SCHEMES

Two schemes have been proposed earlier; both have serious drawbacks.

4.1 Conversion

Conversion [1] has been used for decades to preserve operational data in data processing installations. When a new system is installed, it coexists with the old one for some time, and all files are copied from one system to the other. If some file formats are not supported by the new system, the files are converted to new formats and applications are changed accordingly.

Conversion is quite reasonable when the information that is being converted will, most likely, be used in the near future (a bank account record, for example). However, in the type of archiving that interests us here, the information may not be accessed for a long time, potentially much longer than the period between two system changes. Then, the conversion becomes a burden and is not really necessary. Another disadvantage of conversion is that the file is actually changed repeatedly – and it is hard to predict the cumulative effect that such successive conversions may have on the document.

4.2 Emulation

In [2], Jeff. Rothenberg sketched out an overall method based exclusively on *emulation.*

In summary, it consists of saving, together with the data,

- the original program P, also as a bit stream, that was used to create and manipulate the data (this program runs on M2000), including the operating system and other components when necessary,
- the detailed description of the M2000 architecture,
- a mostly textual description of how to use the program P and what its execution returns.

Then, in 2100, an emulator of the M2000 machine can be built, based on the architecture description. Once this is done, the program P can be run and will produce the same results that P used to produce in 2000. Let us point out right away that building an emulator from the description of the M2000 architecture is not a simple endeavor. It can be done only if the description of the M2000 architecture is perfect and complete (a notoriously difficult task in itself). But even then, how do we know the emulator works correctly since no machine M2000 exists for comparison? In [3], the same author suggests the use of an *emulator specification*. Although no particular specification means is proposed, we can imagine that these specifications could be prepared in 2000 from the M2000 architecture, facilitating the actual generation of an emulator in 2100 (by a human or a machine).

Note that the method hinges on the fact that the program P is the original executable bit stream of the application program that created or displayed the document (including the operating system). This is justifiable for behavior archiving but is overkill for data archiving. In order to archive a collection of pictures, it is certainly not necessary to save the full system that enabled the original user to create, modify and retouch pictures. If Lotus Notes is used to send an e-mail message in the year 2000, it is superfluous to save the whole Lotus Notes environment and have to reactivate it in 2100 in order to restore the note content. But there is an even worse drawback: in many cases, the application program may display the data in a certain way (for example, graphically) without giving explicit access to the data itself. In such a case, it is impossible to move the actual data from the old system to the new one.

Other authors have voiced a similar reservation. For example, D. Bearman in [4] notes that "Rothenberg is fundamentally trying to preserve the wrong thing by preserving the information system functionality rather than the record".

5. OUR PROPOSAL

One important characteristic of the method introduced in [5] and summarized below, is that it differentiates between data archiving which does not require full emulation, and program archiving which does.

For data archival, we propose to save a program P that can extract the data from the bit stream and return it to the caller in an *understandable way*, so that it may be transferred to a new system. The proposal includes a way to specify such a program, based on a Universal Virtual Computer (UVC). To be

[1] A piece of text encoded in a well known alphabet such as ASCII is a particularly easy instance of a case 2 document; the only requirement is that the DL catalog keep the alphabet definition.

understandable, the data is returned with additional information, according to the metadata (which is also archived with the data).

For the archival of a program behavior, emulation cannot be avoided. But here also, the Universal Virtual computer can be used to write the M2000 emulator in year 2000, without any knowledge of what M2100 will be.

5.1 Data Archival

Consider fig.1. The data contained in the bit stream is stored in 2000, with an arbitrary internal representation, Ri. In 2100, The data is seen by a client as a set of data elements that obey a certain *schema* Sd, in a certain *data model.* A decoding algorithm (*method*) extracts the various data elements from Ri and returns them to the client, tagged according to Sd. A language L is used in 2000 to specify the details of the needed decoding. In addition, a mechanism allows the client to read Sd as if it were data. It relies on a schema Ss, a schema to read schemas. The schema Ss is simple and intuitive; it should endure for a long time to come, and be published in many places so that it remains known. In the following sections, we describe each component in more detail.

5.1.1 The logical data model

The choice of an appropriate data model and a means to describe Sd is based on the following premises:

1) it must be simple in order to minimize the amount of description that must accompany the data and decrease the difficulty of understanding the structure of the data;

2) it is only used to restore the data and not to work with it (actually, once it is restored, the data will generally be stored in the repository of the system used at restoration time, maybe under a different model).

Flat files, or tables similar to those used in the relational model, certainly satisfy the requirement. So do hierarchies. Simple and powerful, inherently easy to linearize, they have been used in many areas, often under a different name: *repeating groups* or *non-first-normal-forms* [6] in databases, *Backus-Naur Form* [7] in syntax specification, and - more recently - *XML* [8]. We choose an XML-like approach.

5.1.2 An example

Consider a file containing a collection of pictures of historical buildings in Mycity, including both formatted data and gray scale pictures. The file is a list of records. Each record consists of a sequence of fields. Each field can itself be a list of records made of fields that can be records, etc. A table showing the populated hierarchical structure is shown, much abridged, in fig. 2.

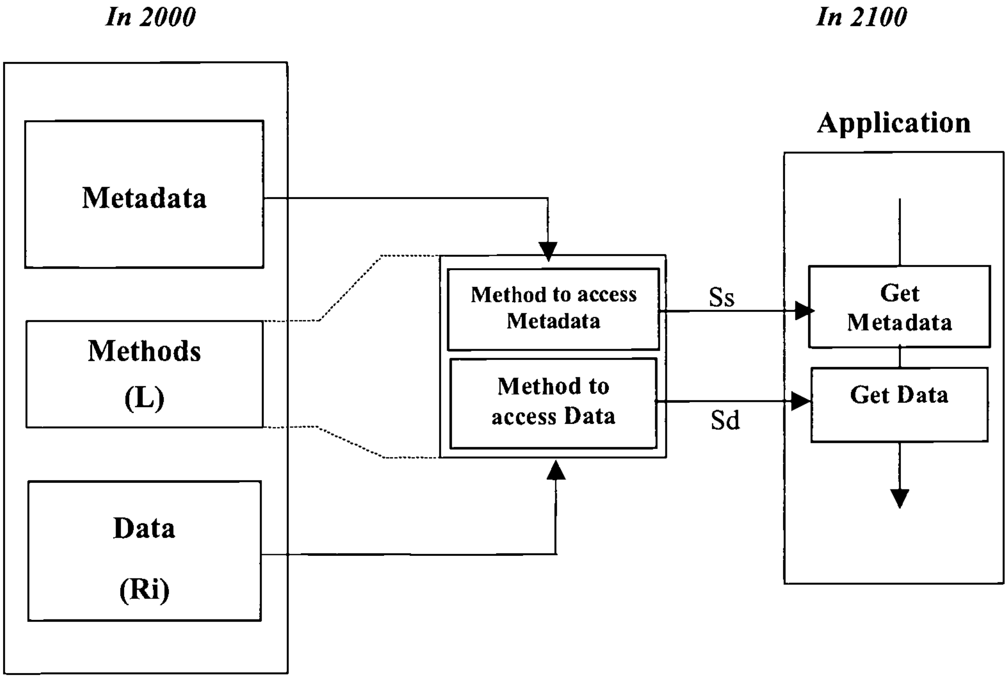


Figure 1: Overall Mechanism for Data Archival

building						
name	address	picture				
		year	nbr_lines	dots-per-line	line	
					No	gray_value
ABC Building	12 Main street	1903	1200	2100	1	102 104 116 ...
					2	211 234 ...
		1924	900	1300	1	125 ...
XYZ Building	9 North street	1917	2180	2700	1	202

Figure 2: Populated table for a Collection of buildings

A reader is able to understand what the various data elements mean because the header, displayed at the top of the table, describes the schema Sd. In addition, the indentation of the data allows the reader to "parse" the data according to the schema. For the digital equivalent of the data in fig. 2, a representation of Sd is also needed. The proposal consists of storing, together with the data, a representation Ri of Sd. That representation is nothing else than the linearized form of a construct similar to a Data Type Definition (DTD) in XML; it defines the application-dependent tags. The DTD for the application is shown in fig. 3.

In plain English, it would read as:

> a collection is a list of buildings; a building is associated with an address, a name, and a list of pictures; a picture is associated with a date (year), the number of dot lines in the picture, the number of dots per line, and a list of lines; a line has a number and a list of gray values.

The * token stands for 0 or more; + means at least 1; ? means optional.

The DTD is the metadata to understand the data (it is clearly application-dependent). At this point, let us assume that the client knows the DTD.

5.1.3 *Invocation and functionality of the methods*

The method to access the data supports the retrieval of all values in the tree according to a depth-first traversal. At a logical level, the restore application in 2100 simply executes the following (pseudo-code) statements:

```
open
while (more) {
        get_field (tag, x)
}
```

```
DOCTYPE  Building_collection [
 ! This is a collection of gray scale pictures of historical buildings
   in Mycity.
 ! A building has an address, and an (optional) name; it can have
   several pictures (for different years).
 ! The gray value is between 0 (white) and 255 (black).

   ELEMENT Collection (building+)
   ELEMENT building (name?, address, picture+)
   ELEMENT picture (year, nbr_lines, dots-per-line, line+)
   ELEMENT line (nbr, gray_value+)

   ELEMENT name (CHARr)
   ELEMENT address (CHAR)

   ELEMENT year (NUM)
   ELEMENT nbr_lines (NUM)
   ELEMENT dots_per_line (NUM)

   ELEMENT nbr  (NUM)
   ELEMENT gray_value (NUM)
```

Figure 3: The DTD for our application

For each field, the value is returned in variable x, together with a <tag>. The tags, although slightly different from those used in XML, unambiguously indicate the type of each element. In the example, the first repetitive calls to get_field would return the following:

```
<Collection>
    <building>
        <name>                      ABC_Building
        <address >                  12  Main street
        <picture>
            <year>                  1903
            <nbr_lines>             1200
            <dots_per_line>         2100
            <line>                  ...
```

5.1.4 *A schema to read schemas*

Contrary to what we assumed earlier, the client may not know the information contained in the DTD. Therefore, we need to provide a way to retrieve that information as well. A simple solution consists of adopting for the schema a method similar to the one proposed for the data: the schema information is stored in an internal representation Ri, and accompanied by a method to decode it. Logically, the Ss looks like this:

```
DOCTYPE Metadata [
    ELEMENT fields (root_name, comment, field+)
    ELEMENT field  (level, name, description, type?, attribute?)

    ELEMENT root_name (CHAR)
    ELEMENT comment (CHAR)
    ELEMENT level (NUM)
    ELEMENT name (CHAR)
    ELEMENT description (CHAR)
    ELEMENT type (CHAR)
    ELEMENT attribute (CHAR)
]
```

The *level* specifies the depth of a record in the hierarchy. The same code "open... get_field..." can be used to retrieve the metadata. Fig. 4 shows the initial section of the results.

The mechanism presented above accomplishes the following: it defines a simple interface for accessing the archived data. That interface is simple because the decoding rules are all contained in the methods; it will therefore easily survive for a very long time (its definition may have to be stored in more than one place but it certainly does not need to be stored with each archived object). The same is true for the mechanism to read schemas.

It should be noted that the examples used above are oversimplified, but they are sufficient for illustrating the proposed mechanism. It is clear that the DTD's will need identifiers and references such as those available in XML.

5.1.5 *Specification of methods*

The responsibility of extracting the logical data elements from the data stream lies with the methods, supposedly written in L. But what should L be? Let's try some possibilities:

1. A natural language. The difficulties are well known; and computer scientists have invented all kinds of codes and pseudo-codes to avoid them, leading to the next item:

2. A high level language. High-level languages are designed to facilitate the writing of programs by large communities of programmers. Language developers, then, always try to incorporate the latest features that may facilitate program development. Every five or ten years, something new seems to come along and the current language gets obsolete.

3. The machine language of the computer on which the algorithm runs in 2000. This is the option that requires a full emulation of the M2000 to be written at restore time; we have discussed its difficulties earlier in this paper.

```
<fields>
        <name>                      Collection
        <comment>                   this is a set of pictures of historical buildings
        <field>
                <level>             0
                <name>              building
                <description>       list of building(s) which have pictures
                <attribute>         +
        </field>
        <field>
                <level>             1
                <name>              name
                <description>       name of the building
                <type>              CHAR
                <attribute>         ?
        </field>
        <field>
                <level>             1
                <address>           postal address of building
                .........
        .........
        .........
</fields>
```

Figure 4: Retrieving a schema definition (partial results)

Instead, we propose to describe the methods as programs written in the machine language of a ***Universal Virtual Computer (UVC).*** The UVC is a Computer in its functionality; it is Virtual because it will never have to be built physically; it is Universal because its definition is so basic that it will endure forever.

The UVC program is completely independent of the architecture of the computer on which it runs. It is simply interpreted by a UVC Interpreter. A UVC Interpreter can be written for any target machine.

This approach does not have the drawbacks of the method 3 above. If a UVC program is written in M2000, it can be tested on a UVC interpreter written in 2000 for an M2000 machine. If x years later, in 2000+x, a new machine architecture comes up, a new UVC interpreter can be written. For quality control, a set of UVC programs can be run through both the 2000 and 2000+x interpreter, and should return exactly the same sequence of tagged data elements. Also, a flaw in the interpreter will never damage any archived document; a programmer may simply have to refer to the UVC specifications to fix the problem. Actually, it is safe to assume that the source code used to implement the year 2000 interpreter can be used as the base for developing the 2000+x version. Also, the same source can be compiled for various target computers, still decreasing the size of the task.

In addition, the UVC can be very simple - and at the same time very general, so that writing an interpreter at any time remains a simple task, far from the complexity of writing a full machine emulator.

5.1.6 The UVC Architecture

The details of the UVC specification are not important at this point. Clearly, its architecture may be influenced by the characteristics of existing real computers or virtual machines developed for different purposes, such as Java. What is important is that it does not need to be implemented physically. Therefore there is no actual physical cost. For example, the UVC can have a large number of registers; each register has a variable number of bits plus a sign bit. The UVC has an unlimited sequential bit-oriented memory. Addresses are bit-oriented (so that a 9-bit "byte" computer can be emulated as easily as an 8-bit one). Also, speed is not a real concern since M2100 will be much faster and these programs are run mostly to restore the data and store them in an M2100 system, and a small set of instructions is sufficient. This reduces the amount of work involved in developing an interpreter of the UVC instructions onto a real M2100 machine.

The fact that the instruction set is kept to a minimum may complicate the writing of a program at the machine instruction level. But, as in any RISC machine, it is anticipated that the methods will be coded in some high level language and automatically compiled onto UVC instructions.

5.1.7 The UVC Interface

In 2100, a machine M2100 will come with a restore program that will read the bit stream in a virtual memory and then issue requests to the UVC Interpreter. The interface must be independent of the conventions used in 2000 or 2100. It uses software registers, filled with single values (of elementary types), according to the following list:

- Reg 0: an integer (k) indicating which function is being invoked.
- Reg 1: the completion code returned by the function.
- Reg 2: a pointer *p-data*, pointing to the data bit stream.
- Reg 3: a pointer *p_out* to some memory set aside to receive a data element.
- Reg 4: a pointer *p_tag* to some memory set aside to receive the tag of the data element.
- Reg 5: a pointer *pw* to a working area.

There is a single entry point to the beginning of the UVC code for the methods. That code will branch to the appropriate method depending upon the value k in Reg 0.

5.1.8 Highlights of the approach for data archiving

The use of the UVC gets rid of the need for a full machine emulator. It also eliminates the need for agreeing on standardized formats. Anybody who wants to preserve a file can use any format but must make sure that the UVC method is supplied to interpret the format. The only standards needed are now the UVC and the data model for data and metadata; they are simple enough to endure. Only the UVC interpreter will have to be written (or re-compiled) when a new machine architecture emerges. There is no impact on the archived information.

5.2 Program Archival

When the behavior of a program needs to be archived, the M2000 code must be archived and later emulated. If the program is only a series of native instructions of the M2000, it may not require the saving of any other package or operating system. However, if the object is a full-fledged system with Input/Output interactions, then the operating system must be archived as well.

We have mentioned earlier the difficulties implied in writing emulators in the future. The UVC approach can be naturally extended to support the archiving of programs, providing for a way to essentially write the emulator in the present. Instead of archiving the UVC method to decode the data, the actual M2000 program will be archived, together with UVC code that emulates the instruction set of M2000. This time, in 2100, the UVC interpreter will interpret that UVC code; that interpretation will then yield the same results as the original program on an M2000. This suffices if the program does not have any interaction with the external world (Input/Output operations or interrupts).

Things get more complicate when Input/Output operations are involved. Suppose the program prints a black/white document on an all-point-addressable printer. The program somewhere issues a Start I/O operation with some data. Clearly the execution of that instruction is not part of the M2000. The M2000 only sends the data to an output device processor P which computes an output-oriented data structure S (such as a bit map), and sends it to the last process, the one that actually prints the page. Our proposal for extending the method to support such operations is as follows.

In addition to archiving the UVC program that interprets the M2000 code, another UVC program that mimics the functioning of P must also be archived. It will produce the structure S. It is impossible to anticipate in 2000 the output technology that will exist in 2100. But, if S is simple and well documented, it will be relatively easy to write in 2100 a mapping from S to the actual device. For an all-point-addressable B/W printer, S is simply a bit map. The bit map becomes the interface to an abstract printer, independently of what the new technology will be. This technique, again, ensures that the difficult part (which depends heavily on the details of the device) is written in 2000 when the device exists. It can be fully tested in 2000 by mapping the abstract device into a 2000 device.

Abstract devices must be similarly defined for sequential tapes (with operations such as R, W, Rewind, Skip), for random access storage units (R, W at a particular record address), for sequential character output or input (screen, keyboard), for x/y positioning (mouse, touch-screen, cursor), etc.

6. SUMMARY AND CONCLUSIONS

In this paper, we analyzed the challenges of archiving digital information for the very long term.

We made a distinction between the archiving of data and the archiving of a program behavior.

The same technique is used to solve both problems: both rely on a Universal Virtual Computer. For archiving data, the UVC is used to archive methods which interpret the stored data stream. For archiving a program behavior, the UVC is used to specify the functioning of the original computer.

In summary, for data archiving:

1) In 2000, whoever creates a new data format needs to produce a UVC program to decode the data. For at least one platform, a UVC interpreter must be developed. It can be used to test the correctness of the UVC program.

2) In 2100, every machine manufacturer needs to produce a UVC interpreter.

For program emulation,

1) In 2000, for each platform, the manufacturer needs to provide an emulator of M2000 written as UVC code. Manufacturers of devices in 2000 need to provide the UVC code that emulates the device control unit.

2) In 2100, every machine manufacturer needs to produce a UVC interpreter, and every manufacturer of an I/O device needs to produce an implementation of the abstract device on the real 2100 device.

What the proposed method accomplishes is to provide a reliable framework where preparatory work can be done in 2000 - when the information is well known - rather than in 2100 when the difficulty would be much greater. It also avoids the cumbersome need for defining standards under which the data should be stored. These standards would have to be defined for all types of applications, and would have to remain valid for centuries; this is just unpractical. Instead, the proposed solution replaces the need for a multitude of standards (one for each format) by a single standard on the UVC method. That standard should cover: the UVC functional specifications, the interface to call the methods, the model for the schema and for the schema to read schemas. Each of these components can be kept general and simple enough to remain relevant in the future.

In this paper, we couch the preservation issue in the framework of a digital library. The proposed solution calls for the archiving of the data bit stream, some UVC code, and some metadata describing the data schema and the interface to invoke the methods. A DL system generally contains two databases: the one that contains the metadata and the one that contains the data itself (the archive store). Although it is a matter of choice, we suspect all of the items above will be kept in the archive store. But the meta-database may be used to store some information necessary to bootstrap the restore process (for example, we need to know the alphabet before we start reading any text). The longevity of the meta-database can be ensured by migration and that information will therefore remain available.

Nevertheless, another environment may be worth considering, in which the document is not part of a digital library. Then, the whole information is stored on a removable storage object such as CD-ROM or tape and needs to be restored in a distant future by using only information that *it* contains. The technology presented here remains applicable, with relatively minor additions to solve the bootstrap problem.

It would be naive to think that solving the archiving problem is simply a technical challenge. For example, the success of any effort would hinge on a minimal agreement of all parties involved in generating new technologies or creating new types of data. But this cannot happen before a certain level of technical know how is reached. Thus, it is important for the computer science community to start developing the technology, and the purpose of this paper is to document some initial ideas. Our research project is currently investigating design issues and developing an early prototype to prove the validity of the concepts and evaluate our design decisions. The "real life" aspects of our current work are provided by a joint study with the Koninklijke Bibliotheek, the national library of the Netherlands, The Hague.

7. ACKNOWLEDGMENTS

Robert Morris, then Manager of Data Systems Technology at the IBM Almaden Research Center, brought the importance of the problem to the attention of the author. His encouragements and those of Robin Williams are gratefully acknowledged. Many thanks go to Henry Gladney for many lively discussions on the subject.

8. REFERENCES

1. Waters, D, and Garret, J.: Preserving Digital Information. Report of the Task Force on Archiving of Digital Information, Commission on Preservation and Access and the Research Libraries Group, Inc., May 1996.
2. Rothenberg, J. Ensuring the Longevity of Digital Documents. Scientific American, 272(1), January 1995.
3. Rothenberg, J.: Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. A report to the Council on Library and Information Resources, January 1999.
4. Bearman, B. Reality and Chimeras in the Preservation of Electronic Records, D-Lib Magazine, Vol. 5, No 4, 1999.
5. Lorie, R.: Long Term Archiving of Digital information, IBM Research Report RJ 10185, July 2000.
6. Dadam, P. & al.: A DBMS Prototype to support Extended NF2 Relations, ACM SIGMOD, May 1986.
7. Aho & al.: Compilers: Principles, Techniques, and Tools. Addison-Wesley, 1986
8. Harold, E.R. XML, Extensible Markup Language. IDG Books Worldwide, 1998.

Creating Trading Networks of Digital Archives

Brian Cooper and Hector Garcia-Molina
Department of Computer Science
Stanford University
{cooperb,hector}@db.stanford.edu

ABSTRACT

Digital archives can best survive failures if they have made several copies of their collections at remote sites. In this paper, we discuss how autonomous sites can cooperate to provide preservation by trading data. We examine the decisions that an archive must make when forming trading networks, such as the amount of storage space to provide and the best number of partner sites. We also deal with the fact that some sites may be more reliable than others. Experimental results from a data trading simulator illustrate which policies are most reliable. Our techniques focus on preserving the "bits" of digital collections; other services that focus on other archiving concerns (such as preserving meaningful metadata) can be built on top of the system we describe here.

KEYWORDS

preservation, digital archiving, replication, fault tolerance, data trading

1. INTRODUCTION

Digital materials are vulnerable to a number of different kinds of failures, including decay of the digital media, loss due to hackers and viruses, accidental deletions, natural disasters, and bankruptcy of the institution holding the collection. Archives can protect digital materials by making several copies, and then recover from losses using the surviving copies. Copies of materials should be made at different, autonomous archives to protect data from organization-wide failures such as bankruptcy. Moreover, cooperating archives can spread the cost of preservation over several institutions, while ensuring that all archives achieve high reliability. Several projects [4, 12, 24, 10] have proposed making mutiple copies of data collections, and then repeatedly checking those copies for errors, replacing corrupted materials with pristine versions.

A key question for a digital archive participating in a replication scheme is how to select remote sites to hold copies of collections. The archivist must balance the desire for high reliability with factors such as the cost of storage resources and political alliances between institutions. To meet these goals, we propose that archives conduct peer-to-peer (P2P) *data trading*: archives replicate their collections by contacting other sites and proposing trades. For example, if archive A has a collection of images it wishes to preserve, it can request that archive B store a copy of the collection. In return, archive A will agree to store digital materials owned by archive B, such as a set of digital journals. Because archive A may want to make several copies of its image collection, it should form a *trading network* of several remote sites, all of which will cooperate to provide preservation.

In previous work [5], we have studied the basic steps involved in trading and the alternatives for executing these steps. For example, in one step a local site selects a trading partner from among all of the archive sites. This requires the local site to choose some strategy for picking the best partner. In another step, the local site asks the partner to advertise the amount of free space it is willing to trade. Then, the local site can determine if the partner will trade enough space to store the local site's collections. We summarize our conclusions from this previous study for these and other issues in Section 2.2 below.

In this paper, we discuss how a digital archive can use and extend these basic trading building blocks to provide preservation services. Archives must take into consideration real-world issues that impact the decisions they make while trading. For example, an archive may have budgetary constraints that limit the amount of storage it can provide. Storage resources cost more than just the expense of buying disk space. In particular, an archive must also provide working servers, administrators to maintain those machines, network access to the servers, and so on. Here, we study how the amount of storage a site provides impacts its ability to trade and the number of copies it is able to make.

Another issue that archives must confront is that they may choose trading partners for a number of reasons beyond simply achieving the highest reliability. For example, the libraries of a particular state university system may be directed to cooperate by the university's board of regents. We call such a grouping of sites a trading *cluster*. The cluster may be large enough to serve the needs of its member sites, or sites may need to seek binary *inter-cluster* links

*This material is based upon work supported by the National Science Foundation under Award 9811992.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

with other archives to expand their trading networks. We examine the ideal cluster size as well as the number of inter-cluster links that must be formed to compensate for a too-small trading cluster.

A site also may have to deal with trading partners that are more or less reliable than itself. For example, a very reliable site must decide whether to trade with all archives or only with those that also have high reliability. We examine these issues to determine how sites can make the best decisions in the face of varying site reliabilities.

Other researchers have examined using redundancy to protect against failures in systems such as RAID [21], replicated file systems [8], and so on. Our work is similar to these systems in that we use replication, we balance resource allocation and high reliability, and we attempt to ensure high data availability.

Unlike these previous systems, our data trading scheme is focused on respecting the differences between individual digital archives, even as these archives cooperate to achieve reliability. Thus, a primary concern of ours is site autonomy. Archivists should be able to decide who they trade with, what types of collections they store and how much storage they provide. Such local decisions are not as important in a system such as RAID, in which a central controller makes all of the decisions. Archives also may have differing reliability goals, such that one archive is willing to expend more resources and expects correspondingly higher reliability in return. It may therefore be important to consider different policies for high and low reliability sites, such that both kinds of sites can protect their data. Similarly, different archives may experience different rates of failure, and an archive may wish to take these failure rates into account when replicating collections. An array of similar components (such as RAID) does not face this issue. Finally, an archivist has unique concerns that are not addressed in traditional systems. It is often important to establish the provenance of collections, and this task is difficult if the collections are moved from site to site frequently or without the archivist's control. An archivist may also wish to keep collections contiguous, so that they can be served to users as a complete unit. Our trading mechanism is flexible enough to address all of these concerns, from autonomy to contiguous collections, while still providing a great deal of protection from failures.

In this paper, we examine how a digital archive can preserve its collections by forming and participating in P2P trading networks. In particular, we make several contributions:

- We present a trading mechanism that can be used by an archive to reliably replicate data. This mechanism is tuned to provide the maximum reliability for the archive's collections, and can be extended if necessary in consideration of individual archivists' needs and goals.

- We identify how to configure an archive for trading by examining the amount of storage that the site should provide and the number of copies of collections a site should try to make.

- We examine the impact of trading with remote partners chosen for political reasons, as opposed to trading with all archive sites. We also discuss the optimal trading network size, and examine when an archivist may wish to seek out additional trading partners.

- We discuss how an archive might trade with sites that have different *site reliabilities*, or rates of failure, by adjusting its trading policies to take these reliabilities into account. We also discuss the importance of accurately estimating the reliabilities of other sites.

In order to evaluate each of these issues, we have used a simulator that conducts simulated trading sessions and reports the resulting reliability. Our concern is primarily in selecting remote sites for storing copies of archived collections. Once trades have been made and collections are distributed, archivists can use other existing systems to detect and recover from failures, enforce security, manage metadata, and so on. Other projects have examined these issues in more detail [4, 22, 17, 23, 19]. It is also possible to enhance our basic techniques to deal with digital objects which change over time, or trades with sites that provide a specialized service (such as storage for a fee). In ongoing work, we are extending our model to provide negotiation for access services (such as search) in addition to storage services. We are also extending our model to deal with trades of other commodites, such as money or processing power, in addition to storage space.

This paper is organized as follows. In Section 2 we discuss the basic trading mechanism, as well as extensions to the basic mechanism for trading networks of digital archives. Section 3 presents evaluations of alternative trading policies using simulation results. Section 4 discusses related work, and in Section 5 we present our conclusions.

2. DATA TRADING

Data trading is a mechanism for replicating data to protect it from failures. In this section, we summarize the techniques used in data trading. We also discuss the extensions and enhancements to data trading that are needed to use the mechanism for digital archives. A full discussion of the basic data trading algorithm, as well as analysis of the tradeoffs involved in tuning the algorithm, is presented elsewhere [5].

2.1 Archival services

Our model of a digital archiving service contains the following concepts:

Archive site: an autonomous provider of an archival storage service. A site will cooperate with other autonomous sites that are under the control of different organizations to achieve data replication. The focus of this paper is the decisions made by a particular archive site; we refer to this site as the *local site*.

Digital collection: a set of related digital material that is managed by an archive site. Examples include issues of a digital journal, geographic information service data, or a collection of technical reports. Although collections may consist of components such as individual documents, we consider the collection to be a single unit for the purposes of replication. Here, we assume that all collections are equally important and require the same effort to preserve.

Archival storage: storage systems used to store digital collections. Some of the storage, called the *public* storage, is dedicated to remote sites that have concluded trades with the local site, and is used to store collections owned by the remote sites. An archive site must decide how much public

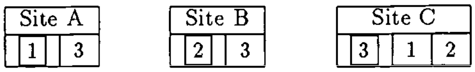


Figure 1: Reliability example.

storage P_{total} to provide. Here, we assume that a site uses a *storage factor* F, such that if the site has N bytes of archived data, it purchases $F \times N$ total storage space. The site uses N bytes of this space to store its own collections, and has $P_{total} = F \times N - N$ extra space to trade away.

Archiving clients: users that deposit collections into the archive, and retrieve archived data. When a client deposits a collection at an archive site, that site is said to "own" the collection, and takes primary responsibility for protecting it.

Trading network: a local site must connect to *remote sites* and propose trades. In the general case, any site can connect to any other site. In a digital archiving domain, it may be more desirable to select a set of "trusted" sites to trade with. This issue is discussed in more detail below.

Automation: The archive should operate as automatically as possible, while allowing librarians or archivists to oversee its operation and adjust its configuration. Thus, an archiving site may automatically replicate copies of a digital collection, but would do so according to the desired goals and constraints important to the administrators.

These concepts are used to study the properties of a trading network primarily concerned with protecting the data itself from loss. While we do not consider other archival concerns (such as preserving access or establishing chain of custody) for simplicity, our model can be extended to deal with such concerns.

Each archive site can fail (lose data), and we model the possibility of failures as a *site reliability*: a number which indicates the probability that the site does not experience data loss. Although site reliabilities may change over time, here we assume for simplicity that reliabilities are constant. Given site reliabilities and a placement of copies of collections at these sites, we can calculate two values:

- *global data reliability*: the probability that no collection owned by any site is lost.
- *local data reliability*: the probability that no collection owned by a particular site is lost.

Thus, global data reliability measures the success of the trading mechanism as a whole, while local data reliability measures the success of decisions made by an individual site participating in data trading. For example, consider Figure 1. This figure shows three sites, each of which owns one collection (shown boxed), while storing copies of collections owned by other sites. Let us assume that the site reliability of each site is 0.9, that is, each site has a ten percent chance of experiencing data loss. In one possible scenario, sites B and C fail but site A does not, while in another scenario, none of the sites fail. We can calculate *global data reliability* by examining all possible scenarios of sites failing or surviving; in this case, there are eight possibilities. For each scenario, we assign the score "0" if at least one collection is lost, or "1" if no data is lost. Thus, in the scenario where sites B and C fail, collection 2 is lost and we assign the score 0. We then weight each score by the probability of the scenario; the situation where B and C fail but A does not will occur with probability $0.1 \times 0.1 \times 0.9 = 0.009$. Finally, we sum the weighted scores to find the expected probability of data loss. The distribution of collections shown in Figure 1 has a global reliability of 0.981, indicating that there is less than a two percent chance of data loss.

We can calculate *local data reliability* in much the same way, except that we only consider the collections owned by a particular site when assigning scores. For example, if we wish to calculate the local reliability of site A, we examine all possible scenarios, and assign a score of "0" if collection 1 is lost, or "1" if collection 1 is preserved. In this way, we can calculate the local data reliability of site A and site B to be 0.99, while the local data reliability of site C is 0.999. Site C enjoys higher data reliability because it has made more copies of its collection.

We can interpret local and global data reliabilities as the probability that data will not be lost within a particular interval, say, one year. Then, we can calculate the expected number of years before data is lost, known as the *mean time to failure* (MTTF). An increase in reliability from 0.99 to 0.999 actually represents an increase in the MTTF from 100 years to 1000 years. Because MTTF better illustrates the results of a particular policy by giving an indication of how long data will be protected, we report our simulation results in Section 3 in terms of MTTF.

In this paper we are primarily concerned about evaluating the choices made by individual sites, and preserving the autonomy of those sites. Therefore, we will examine data trading from the perspective of local data MTTF. In previous work [5] we have assumed that all sites have the same probability of failure, but here we consider the possibility that different sites have different reliabilities.

2.1.1 The trading network

There are two reasons why a local site may choose a particular remote site as a P2P trading partner. First, the remote site may have a reputation for high reliability. Second, there may be political or social factors that bring several autonomous archives together. An archive must make trades that take both reliability and politics into account.

We refer to the set of potential trading partners for a local site as that site's *trading network*. In our previous work we have assumed that a site's trading network includes all other archive sites. However, a local site may participate in one or more *clusters*, or sites that have agreed to form partnerships for political, social or economic reasons. For example, all of the University of California libraries may join together in one cluster. A local site may also have individual *intercluster* links for political or reliability reasons. If an archive at say MIT is well known for high reliability, one of the University of California libraries may form a partnership with MIT in addition to the California cluster. Once a site has found trading partners, it can continue to consider politics and reliability when proposing trades. In Section 2.2 we discuss how a site can use site reliabilities to select sites for individual trades.

There are two challenges that face a site when it is constructing a trading network. The first challenge is deciding how many sites should be in the network, and what intercluster partnerships to form. The second challenge in constructing a trading network is estimating the site reliabilities of other sites. One possible method is to examine the past behavior of the site. Sites with many failures are likely to have more failures in the future, and are assigned a lower site reliability than sites that rarely fail. Another method is to examine components of the archive's storage mechanism [6]. Sites that use disks that are known to be reliable or security measures that have successfully protected against hackers should be given a higher site reliability. A third possibility is to use the reputation of the site or institution hosting the site. Thus, even the perceived reliability of a site can be influenced by political or social factors.

We evaluate the ideal size for trading clusters, and give guidelines for how many inter-cluster partnerships should be formed in Section 3. We also examine the impact of site reliability estimates in that section.

2.2 Conducting trades

When a client deposits a collection at an archive site, the site should automatically replicate this collection to other sites in the trading network. This is done by contacting these sites and proposing a trade. For example, if site A is given a collection of digital journals, site A will then contact other sites and propose to give away some of its local archival storage to a site willing to store a copy of the journals.

We have developed a series of steps for conducting trades in previous work [5]. These steps are summarized in the DEED_TRADING algorithm shown in Figure 2. This is a distributed algorithm, run by each site individually without requiring central coordination. A deed represents the right of a local site to use space at a remote site. Deeds can be used to store collections, kept for future use, transferred to other sites that need them, or split into smaller deeds. When a local site wants to replicate a collection, it requests from a remote site a deed large enough to store the collection. If the remote site accepts, the local site compensates the remote site with a deed to the local site's space. In the simplest case, the deed that the local site gives to the remote site is equal to the deed that the remote site gives to the local site. There are other possibilities; see below.

Several details of the DEED_TRADING algorithm can be tuned to provide the highest reliability:

<S>: The *trading strategy* <S> dictates the order in which other sites in the trading network will be contacted and offered trades. The best strategy is for a site to trade again with the same archives it has traded with before. This is called the *clustering* strategy, because a site tries to cluster its collections in the fewest number of remote sites. If there are several sites that have been traded with before, the local site selects the remote site holding the largest number of the local site's collections. If there is still a tie, or if there are no previous partners, the local site chooses a remote site randomly. For the special case where sites have small storage factors (e.g. $F = 2$), the *best fit* strategy is best. Under best fit, the remote site with the smallest advertised free space is chosen. In [5] we examine several other strategies, such as *worst fit*, where the site with the most advertised free space is preferred. If different sites have different reliabilities, as we assume in this paper, it is possible to adjust the strategy to reflect those reliabilities; see below.

<A>: A site must decide how much of its storage space to offer for trades. The best advertising policy <A> is the *data-proportional* policy, where a site advertises some multiple y of the total amount of data N owned by the site. If the amount of remotely owned data stored so far is P_{used}, and the amount of free public space is P_{free}, then the advertised amount is:

$$MIN(N \times y - P_{used}, P_{free})$$

Thus, the amount of advertised space is the total amount of "available" public space minus the amount of public space used so far, except that a site cannot advertise more public space than it has free. Our experiments show that the best setting for y is $y = F - 1$, where F is the site's archival storage factor (see Section 2.1).

<U>: If a local site has a deed for a remote site, it can use that deed to make a copy of any collections that fit in the deed but do not already exist at the remote site. A site must decide when to use a deed that it holds to make more copies of collections. The *aggressive* deed use policy, which provides the highest reliability, dictates that a site will use its deeds to replicate as many collections as possible, in order of rareness. Thus, a site holding a deed will use it to replicate its "rarest" collection (the collection with the fewest copies) first. If some of the deed is left over, the site will make a copy of the next rarest collection, and so on. These collections are replicated even if they have already met the replication goal <G>.

<R>: If a site is unable to make <G> copies of a collection C_L, it can try to trade again in the future to replicate the collection. The *active retries* policy says that a site will not wait to be contacted by other sites to make copies of C_L, but instead will run DEED_TRADING again after some interval to replicate C_L. A site must choose an appropriate event to trigger the retry; for example, the site may wait one week before trying again.

DEED_TRADING also uses the following policies, which are investigated in this paper:

<G>: A site tries to make <G> copies of a collection. Once this target is met, the site does not have to make any more trades. Appropriate values of <G> are discussed in Section 3.

<D>: The deed that L gives to R may or may not be the same size as the deed that R gives to L. In our previous work, we have assumed that the two deeds were of equal size. Here, we investigate the possibility that the deed size is influenced by the site's reliability. This issue is discussed below.

2.2.1 *Adapting trading policies for differing site reliabilities*

We can extend the basic trading framework presented in [5] (summarized above) to allow a local site to use the estimated reliabilities of its partners in order to make good trading decisions. There are two aspects of DEED_TRADING that could be modified based on site reliabilities: the trading strategy <S>, and the deed size policy <D>.

One way to change the trading strategy <S> is to look only at site reliabilities when making trades. In the *highest reliability* strategy, a site seeks to trade with partners that have the best reliability. The idea is to make trades that will best protect the local site's collections. In contrast, the *lowest reliability* strategy seeks out sites with the worst reliability. Although each trade may be less beneficial, the low reliability sites may be more desperate to trade than high reliability sites, meaning that the local site can make more copies of its collections. Finally, the *closest reliability* strategy seeks to find the sites with reliability closest to the local site's; the local site must then estimate its own reliability.

```
I. The local site $L$ repeats the following until it has made <G> copies of collection $C_L$, or until all sites in the trading network have been contacted and offered trades:
    1. Select a proposed deed size $D_L = size(C_L)$.
    2. Select a remote site $R$ in the trading network according to the trading strategy <S>.
    3. If $L$ has a deed for $R$ then:
        (a) If the deed is large enough to store $C_L$, then use the deed to make a copy of $C_L$ at $R$. Return to step I.
        (b) Otherwise, set $D_L = D_L - size$(existing deed).
    4. Contact $R$ and ask it to advertise its free space $<A>_R$.
    5. If $<A>_R \ < D_L$ then:
        (a) Contact sites holding deeds for $R$. Give those sites deeds for local space (at $L$) in return for the deeds for $R$. Add these deeds to the existing deed $L$ holds for $R$. Adjust $D_L$ downward by the total amount of the newly acquired deeds.
        (b) If $L$ cannot obtain enough deeds this way, then it cannot trade with $R$, and returns to step I.
    6. $R$ selects a deed size $D_R$ according to the deed size policy <D>.
    7. If $L$'s advertised free space $<A>_L \ < D_R$, the trade cannot be completed. Return to step I.
    8. The trade is executed, with $L$ acquiring a deed of size $D_L$ for $R$'s space, and $R$ acquiring a deed of size $D_R$ for $L$'s space.
    9. $L$ uses its deeds for size $R$ to store a copy of $C_L$.
II. If the goal of <G> copies for $C_L$ is not met, $L$ can try this process again at some point in the future, according to the retry policy <R>.
III. At any time a site may use a deed that it posesses to replicate its collections, according to its deed use policy <U>.
```

Figure 2: The DEED_TRADING algorithm.

Another way to change the trading strategy is to use site reliabilities in combination with other factors. In the clustering strategy, the local site chooses the remote site holding the most copies of collections owned by the local site. In the *weighted clustering* strategy, the local site weights the number of collections by the reliability of the site. For example, site A (reliability 0.5) might hold three collections while site B (reliability 0.9) might hold two collections. We consider the partnership value of site A to be $0.5 \times 3 = 1.5$, while the partnership value of site B is $0.9 \times 2 = 1.8$; thus, site B is chosen. Other strategies could be weighted in a similar manner. In the case of best fit and worst fit, we can multiply the advertised space by the site's reliability, and use the weighted value in the best fit or worst fit calculations. In this way, we are calculating the "expected" amount of space at the remote site based on the probability that the space will actually be available.

The deed size policy <D> can use reliabilities to encourage a "fair" trade between sites. Under the (previously studied) *same size* policy, the local site and remote site exchange deeds that are the same size. However, if the reliabilities of the two sites differ, then a deed for the more reliable site may be considered "more valuable," and the less reliable site will have to give a larger deed to compensate. We can denote the site reliability of site i as P_i, and the size of the deed that the site gives in trade as D_i. Then, we can calculate the *reliability-weighted* value of the deed as $P_i \times D_i$. The *weighted size* policy dictates that the reliability-weighted values of the exchanged deeds must be equal, e.g. if the local site L trades with the remote site R then $P_L \times D_L = P_R \times D_R$. The local site chooses a deed size based on the collection it wants to replicate, so the size of the deed that the remote site must give in return is $D_R = (P_L \times D_L)/P_R$.

A local site must be able to estimate the site reliability of its trading partners (and possibly itself) in order to make decisions which take reliability into account. We can denote site i's estimate of site j's reliability as $P_{i,j}$. In an ideal situation, each site could calculate reliabilities exactly, such that $P_{i,j} = P_j$. However, it is difficult to predict which sites will fail, and thus reliability estimates may be innaccurate. A local site can use information about a remote site's reputation, any previous failures, and the reliability of the storage components to estimate the reliability. Thus, it is likely that sites which are in fact highly reliable are known to be reliable, while more failure prone sites are known to be less reliable. In other words, $P_{i,j} \approx P_j$.

In Section 3.3 we examine the reliability resulting from trading strategies that account for reliability and the impact of the same size and weighted size policies. We also examine the effects of innaccurately estimating site reliabilities.

3. RESULTS

3.1 The data trading simulator

In order to evaluate the decisions that a local site must make when trading, we have developed a simulation system. This system conducts a series of simulated trades, and the resulting local data reliabilities are then calculated. Table 1 lists the key variables in the simulation and the initial base values we used; these variables are described below.

The simulator generates a *trading scenario*, which contains a set of sites, each of which has a quantity of archival storage space as well as a number of collections "owned" by

Variable	*Description*	*Base values*
S	Number of sites	2 to 15
F	Site storage factor	2 to 7
P_{MIN}, P_{MAX}	Min/max site reliability	$P_{MIN} = 0.5$ or 0.8 $P_{MAX} = 0.99$
P_{est}	P_i estimate interval	0 to 0.4
$CperS_{MIN}$, $CperS_{MAX}$	Min/max collections per site	$CperS_{MIN} = 4$, $CperS_{MAX} = 25$
$Csize_{MIN}$, $Csize_{MAX}$	Min/max collection size	$Csize_{MIN} = 50$ Gb, $Csize_{MAX} = 1000$ Gb
$Ctot$	Total data at a site	$Ctot_{MIN}$ to $Ctot_{MAX}$
$Ctot_{MIN}$, $Ctot_{MAX}$	Min/max value of $Ctot$	$Ctot_{MIN} = 200$ Gb, $Ctot_{MAX} = 10,000$ Gb
<G>	Replication goal	2-15 copies
<S>	Trading strategy	9 strategies tried
<D>	Deed size policy	*same size* and *weighted size*

Table 1: Simulation variables.

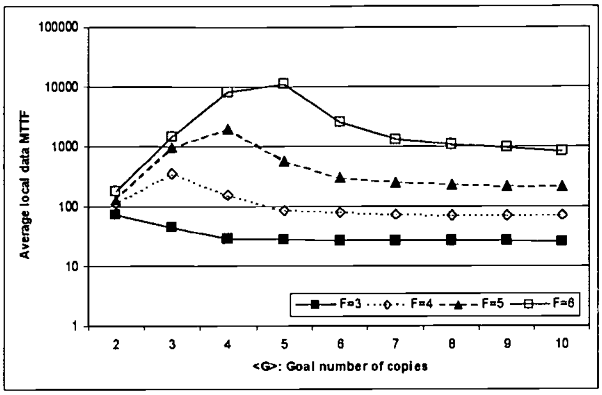


Figure 3: The trading goal and storage capacity.

the site. The number of sites S is specified as an input to the simulation. The number of collections assigned to a site is randomly chosen between $CperS_{MIN}$ and $CperS_{MAX}$, and the collections assigned to a site all have different, randomly chosen sizes between $Csize_{MIN}$ and $Csize_{MAX}$. The sum of the sizes of all of the collections assigned to a site is the *total data size Ctot* of that site, and ranges from $Ctot_{MIN}$ to $Ctot_{MAX}$. The values we chose for these variables represent a highly diverse trading network with small and large collections and sites with small or large amounts of data. Thus, it is not the absolute values but instead the range of values that are important.

The archival storage space assigned to the site is the storage factor F of the site multiplied by the $Ctot$ at the site. In our experiments, the values of F at different sites are highly correlated (even though the total amount of space differs from site to site). By making all sites have the same F, we can clearly identify trends that depend on the ratio of storage space to data. Therefore, we might test the reliability that results from a particular policy when all sites use $F = 2$. In this case, one site might have 400 Gb of data and 800 Gb of space, while another site might have 900 Gb of data and 1800 Gb of space. The scenario also contains a random order in which collections are created and archived. The simulation considers each collection in this order, and the "owning" site replicates the collection. A site is considered "born" when the first of its collections is archived. A site does not have advance knowledge about the creation of other sites or collections. Our results represent 200 different scenarios for each experiment.

We model site failures by specifying a value P_i: the probability that site i will not fail. This value reflects not only the reliability of the hardware that stores data, but also other factors such as bankruptcy, viruses, hackers, users who accidentally delete data, and so on. In our experiments, we consider the situation where all sites are relatively reliable (e.g. $0.8 \leq P_i \leq 0.99$) as well as the case where some sites are quite unreliable (e.g. $0.5 \leq P_i \leq 0.99$). To consider site reliability estimates, we assume that site i's estimate $P_{i,j}$ of site j's reliability is randomly chosen in the range $P_j \pm P_{est}$.

3.2 Local configuration issues

An archive site should have enough space to store the collections deposited by local clients. In order to participate in data trading, a site also needs extra *public* storage space that it trades away. We call the ratio of total space to locally owned collections the *storage factor* F. In this section we examine the best value of F, which indicates the appropriate amount of extra storage a site must provide.

A related issue is the number of copies of collections that a site will attempt to make. If more copies are made, higher reliability results. However, remote sites must have enough storage to hold all of the copies, and the local site must have enough public storage space to trade away to make these copies. In other words, the *goal* <G> number of copies is related to the storage factor F.

To examine the relationship between <G> and F, we tested a situation where 15 archive sites replicate their collections; each site had a reliability of 0.9. We varied F in the range $2 \leq F \leq 6$ and tested goals from 2 to 15 copies. The results are shown in Figure 3. Note that the vertical axis in this figure has a logarithmic scale, and that there are separate data series for $F = 3, 4, 5$ and 6. As expected, providing more storage increases the local reliability. The best reliability (11,000 years MTTF) is obtained when $F = 6$ and sites try to make five copies. (We are mainly concerned with finding the policy that has the highest reliability, regardless of the actual magnitude of the MTTF value.) Trying to make more copies results in decreased reliability because there is not enough space to make more than five copies of every site's collections. If one site tries to make too many copies, this site uses up much of the available space in the trading network, resulting in decreased reliability for other sites.

Sites may wish to purchase less space than six times the amount of data for economic reasons. Our results show that with $F = 5$ and <G>= 4, sites can achieve 2,000 years MTTF, and with $F = 4$ sites can achieve 360 years MTTF if the goal is three copies. Therefore, while buying a lot of space can provide very high reliability, archives can still protect their data for hundreds of years with a more modest investment.

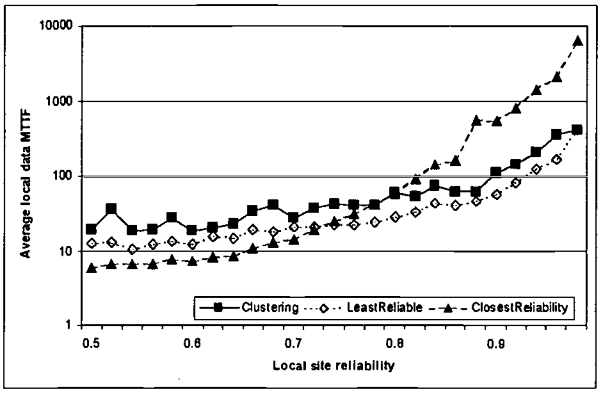


Figure 4: Trading strategies.

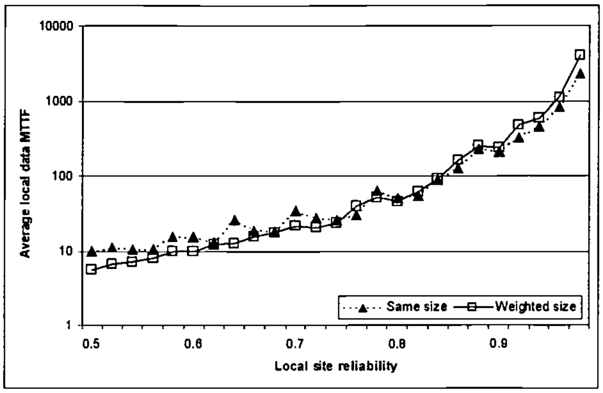


Figure 5: The deed size policy.

3.3 Trading policies that consider reliability

Archive sites can use reliability information about other sites to make trading decisions (Section 2.2). First, we examined trading strategies by running simulations where each site had different reliabilities; site reliabilities were randomly chosen in the range $0.5 \leq P_i \leq 0.99$. In this experiment, there were 15 sites, each with a storage factor of $F = 4$ and a target <G> of three copies. We also assumed (for the moment) that each site was able to predict site reliabilities accurately, so that $P_{i,j} = P_j$. The results are shown in Figure 4. (For clarity, not all strategies are shown; the omitted strategies are bounded by those in the figure.) Recall that the clustering strategy is to trade again with previous trading partners, the closest reliable strategy is to trade with sites of reliability close to that of the local site, and the least reliable strategy is to prefer the least reliable site. The results indicate that the clustering strategy is best for sites with relatively low reliability, but that sites with $P_i \geq 0.8$ are better off using the closest reliability strategy. For example, a site with $P_i = 0.9$ achieves a local data MTTF of 540 years using closest reliability, versus 110 years MTTF resulting from clustering. These results assume that all sites are using the same strategy. We ran another experiment where the high reliability sites ($P_i \geq 0.8$) used one strategy, but the lower reliability sites used another. These results (not shown) confirm that it is always best for the high reliability sites to use the closest reliable strategy, and for the low reliability sites to use clustering. We ran similar experiments with $0.8 \leq P_i \leq 0.99$, and reached the same conclusions, although the range of high reliability sites that should use closest reliability was $P_i \geq 0.9$.

High reliability sites clearly benefit by trading among themselves, so that every trade they initiate places a copy of a collection at a very reliable site. If low reliability sites were to try to trade only among themselves, they would lose reliability by excluding the benefits of trading with high reliability sites. If low reliability sites were to try to trade preferentially with the high reliability sites (as in the highest reliability strategy), they would quickly find the high reliability sites overloaded. Therefore, the best strategy is to make as many trades as possible in a way that is neutral to the remote sites' reliability, and this is what the clustering strategy does. The high reliability sites will not seek out low reliability sites to make trades, but will *accept* trade offers made by those sites.

In order to use strategies that depend on site reliabilities, a site must be able to estimate the reliabilities of itself and its trading partners. We examined the importance of accuracy in these estimates by allowing the probability estimate interval P_{est} to vary. The failure probability P_i of each site is selected at random from the range $0.5 \leq P_i \leq 0.99$, and sites with $P_i \geq 0.8$ used closest reliability while other sites used clustering. Each local site i's estimate of the remote site j's reliability was randomly chosen in the range $P_j \pm P_{est}$. The results (not shown) indicate that the best reliability results in the ideal case: when the estimates are completely accurate. As long as sites are able to make estimates that are within seven percent of the true value, their local data reliability is quite close to the ideal case. However, as the error increases beyond seven percent, the local data reliability drops. For example, when estimates are innaccurate by 30 percent, archives using closest reliability can only achieve a local MTTF of 200 years, versus 500 in the ideal case. If sites can estimate a site reliability close to the true value, they can usually separate high reliability archives from low reliability archives, and select the high reliability sites for trading. If estimates are very innaccurate (e.g. by 25 percent or more) very high reliability sites (e.g. $P_i \geq 0.94$) achieve better reliability using the clustering strategy. However, moderately reliable sites ($0.8 \leq P_i \leq 0.94$) still achieve better MTTF with the closest reliability strategy.

Another policy that can take site reliabilities into account is the deed size policy <D>. We have compared the weighted size policy with the same size policy in an experiment with 15 sites, where $0.5 \leq P_i \leq 0.99$, the storage factor $F = 4$, and the target <G>= 3. The results are shown in Figure 5. (In this experiment, the high reliability sites, $P_i \geq 0.8$, used the closest reliability strategy, and other sites used clustering.) The figure indicates that the weighted size policy, which considers deeds from reliable sites to be more valuable, is good for high reliability sites ($F \geq 0.8$). For example, a site with $P_i = 0.9$ can achieve 240 years MTTF using the weighted size policy, a 14 percent increase over

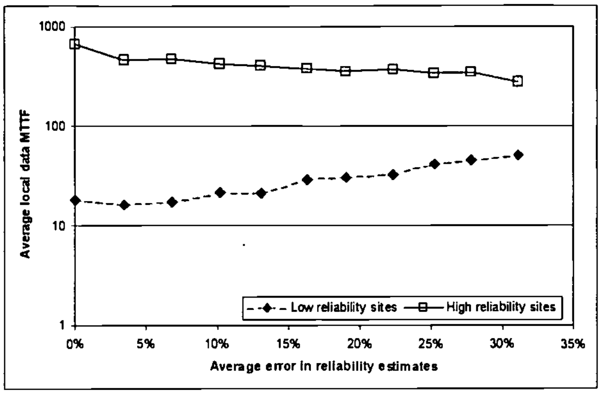


Figure 6: The impact of estimating site reliabilities.

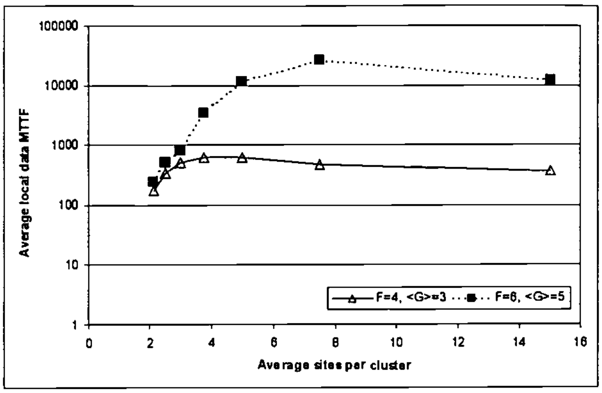


Figure 7: The impact of cluster size.

the same size policy MTTF of 210 years. In contrast, low reliability sites are hurt by the weighted size policy, with as much as a 50 percent decrease in MTTF (from 25 years to 12 years) when $P_i = 0.64$. High reliability sites are the beneficiary of the weighted size policy because they receive more space in trades, and the most reliable sites can demand the most space from other sites. These results indicate that it may be better for low reliability sites to avoid paying the high penalties of the weighted size policy by trading only with other low reliability sites. However, the results (not shown) of another experiment we conducted indicate that it is still better for low reliability sites to try to trade with high reliability archives, even when the weighted size policy is used. If the low reliability sites ignore the high reliability sites by using closest reliability instead of clustering, they experience an average decrease in local data MTTF of 15 percent (from 16 years to 14 years).

Once again, we have examined the effect of estimating reliabilities. Figure 6 shows the impact on local data MTTF versus the accuracy of the estimates. In this experiment, $0.5 \leq P_i \leq 0.99$ and sites estimated reliabilities randomly in the range $P_j \pm P_{est}$ such that a larger P_{est} resulted in a larger average error (shown on the horizontal axis in Figure 6). These results show that high reliability sites suffer when estimates are innacurate, while low reliability sites benefit. This is because a low reliability site can be mistaken for a high reliability site, and thus can get larger deeds from its trading partners. Similarly, high reliability sites can be mistakenly judged to have less reliability, and must accept correspondingly smaller deeds. Nonetheless, most high reliability sites ($0.8 \leq P_i \leq 0.98$) still achieve higher MTTF under the weighted size policy than under the same size policy, even when estimates are as much as 30 percent wrong on average.

In summary, if some archives are more reliable than others:

- Highly reliable sites should trade among themselves. However, if site reliability estimates are off by 25 percent or more, then the clustering strategy is better.

- Less reliable sites should continue to use clustering.

- Highly reliable sites can use the weighted size policy to extract larger deeds from low reliability sites.

- Less reliable sites should try to trade using the same size policy, but should continue to trade with highly reliable sites even if the weighted size policy is used.

3.4 The trading network

In this section, we investigate the ideal trading network size. Specifically, we examine the effects of *clusters*, or groupings of sites that cooperate for political or social reasons. If the cluster is not large enough to serve a site's trading needs, the site will have to seek *inter-cluster* partnerships to expand the trading network. Note that in previous sections, we assumed a local site could potentially trade with any remote site. Even with the clustering strategy, any site was eligible to become a trading partner. In this section we consider the case where clusters are pre-ordained.

In order to determine the ideal cluster size, we ran a simulation in which 15 archive sites were divided into N clusters, where $N = 1, 2...7$. In this experiment, each cluster is *fully isolated*: there are no inter-cluster links. Thus, when $N = 1$ all sites trade with each other, but when $N = 3$ there are three clusters of five sites, and sites trade only within a cluster. We examined the case where $F = 4$ and <G>=3, as well as $F = 6$ and <G>=5. The results are shown in Figure 7. When space is tight ($F = 4$), a cluster of about 5 sites provides the best reliability (with a MTTF of 630 years). In contrast, when there is more space ($F = 6$), then about seven sites is the best cluster size, with a MTTF of 26,000 years. In both cases, larger clusters are actually detrimental, decreasing the local data reliability of the member sites. Large clusters mean that a member site must trade with many other archives, and this can cause some sites to become overloaded; thus their public storage becomes filled up. When this happens, the overloaded sites are less able to make trades, and their reliability suffers. *Therefore it is not necessary or even desirable to form very large clusters in order to achieve reliability.*

If sites participate in trading clusters that are smaller than the ideal size, they can seek inter-cluster partnerships to enhance reliability. We have simulated a situation where 12

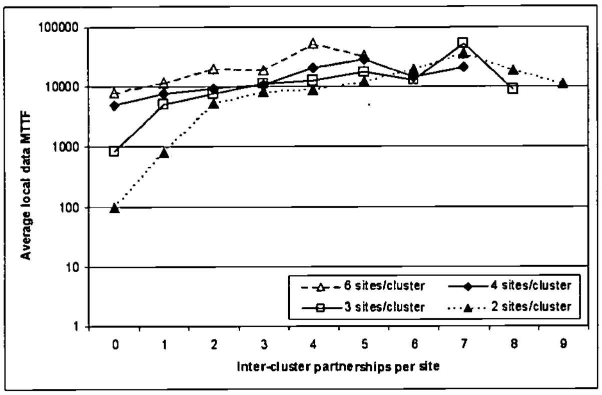


Figure 8: Inter-cluster partnerships, $F = 6$.

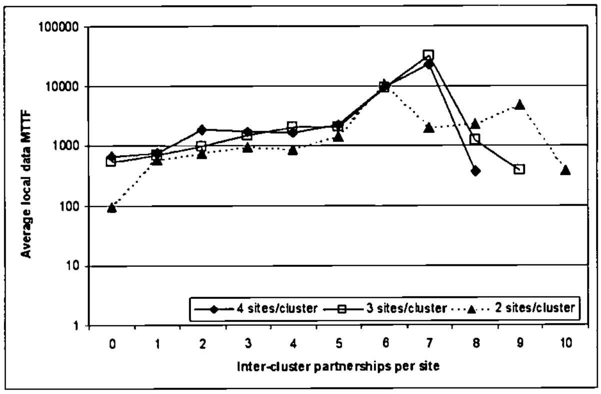


Figure 9: Inter-cluster partnerships, $F = 4$.

sites were divided into small clusters, and each site randomly chose partners outside of its own cluster. Figure 8 shows the results for $F = 6$, where average local data reliability is plotted against the number of inter-cluster partnerships per site. The results show that smaller clusters must seek out many inter-cluster partnerships to achieve the highest reliability. Thus, sites in clusters of three or fewer archives must find roughly seven partners in other clusters, while clusters with four sites should find roughly five additional partners. Even sites in relatively large clusters (e.g. with six sites) can benefit by seeking four inter-cluster partnerships. Seeking too many inter-cluster partners can hurt reliability. A local site may try to find partners outside the cluster, but unless the partners are fully integrated into the cluster, then the local site must field all of the partner's trading requests, and quickly becomes overloaded. Similarly, when $F = 4$, inter-cluster partnerships are beneficial. Our results, shown in Figure 9, indicate that for clusters of less than five sites, six or seven inter-cluster partnerships are needed to achieve the best reliability.

In summary:

- Sites in clusters of about five archives (for $F = 4$) or seven archives (for $F = 6$) achieve the highest reliability.
- Sites in smaller clusters can seek inter-cluster partnerships to improve their reliability.
- If a cluster is too large or if a site has too many inter-cluster partners, reliability can suffer.

4. RELATED WORK

The problems inherent in archiving data are well known in the digital library community [11]. Researchers have confronted issues such as maintaining collection metadata [23, 17], dealing with format obsolescence [25, 19, 14], or enforcing security policies [22]. These efforts complement attempts to simply "preserve the bits" as exemplified by projects like SAV [4], Intermemory [12], LOCKSS [24], or OceanStore [10]. The work we present here can be used to replicate collections in order to best preserve the bits, and can be augmented if necessary (e.g. with a metadata management scheme.)

Many existing data management systems use replication to provide fault tolerance. However, these systems tend to focus on access performance and load balancing [7, 26, 27], whereas we are primarily concerned about reliability. Sites using our clustering strategy attempt to emulate *mirrored disks* [2]. In contrast, database systems tend to prefer a strategy called *chained declustering* [15], which trades some reliability for better load balancing after a failure [18]. Digital archives, which are primarily concerned with preservation, prefer the more reliable mirrored disks; hence, they use the clustering strategy. Moreover, we are concerned with placing archived data that is not likely to change, and therefore are not as concerned as previous researchers with the ability to correctly update distributed replicates [1, 13]. Thus, while a distributed transaction protocol could be added if necessary, efficient or correct updates are less important than preserving the data.

Other systems (such as Coda [16] or Andrew [9]) use replication in the form of *caching*: data is moved to the users to improve availability. Then, if the network partitions, the data is still readable. Our goal is to place data so that it is most reliably stored, perhaps sacrificing short term availability (during network partitions) for long term preservation. Specifically, Andrew and Coda eject data from caches when it is no longer needed. Our scheme assumes that data is never ejected.

The problem of optimally allocating data objects given space constraints is well known in computer science. Distributed bin packing problems [20] and the File Allocation Problem [3] are known to be NP-hard. Trading provides a flexible and efficient way of achieving high reliability, without the difficulties of finding an optimal configuration.

5. CONCLUSIONS

In this paper, we have examined how archives can use and extend peer-to-peer data trading algorithms to serve their data preservation needs. This provides a reliable storage layer that can be enhanced with other services (such as format migration or authenticity verification) to create a complete archiving solution. In order to trade effectively, a site must make certain policy decisions. We have provided guidelines for selecting the amount of storage a local site must provide. We have presented and evaluated trading policies that exploit site reliability estimates, significantly improving reliability. In particular, we have shown that high reliability sites should trade amongst themselves, while low reliability sites should try to trade their collections using the clustering strategy. Finally, we have examined the impact of trading clusters shaped by political and social concerns, and how many extra trading partners a member of such a cluster must find to achieve the highest reliability.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments.

6. REFERENCES

[1] F. B. Bastani and I-Ling Yen. A fault tolerant replicated storage system. In *Proc. ICDE*, May 1987.

[2] Andrea Borr. Transaction monitoring in Encompass [TM]: Reliable distributed transaction processing. In *Proc. 7th VLDB*, September 1981.

[3] W. W. Chu. Multiple file allocation in a multiple computer system. *IEEE Transactions on Computing*, C-18(10):885–889, October 1969.

[4] Brian Cooper, Arturo Crespo, and Hector Garcia-Molina. Implementing a reliable digital object archive. In *Proc. ECDL*, pages 128–143, September 2000. In LNCS vol. 1923.

[5] Brian Cooper and Hector Garcia-Molina. Peer to peer data trading to preserve information. http://dbpubs.stanford.edu/pub/2001-7, 2001. Technical Report.

[6] Arturo Crespo and Hector Garcia-Molina. Modeling archival repositories for digital libraries. In *Proc. ECDL*, pages 190–205, September 2000. In LNCS vol. 1923.

[7] Xiaolin Du and Fred Maryanski. Data allocation in a dynamically reconfigurable environment. In *Proc. ICDE*, February 1988.

[8] Barbara Liskov et al. Replication in the Harp file system. In *Proc. 13th ACM SOSP*, October 1991.

[9] J. H. Morris et al. Andrew: A distributed personal computing environment. *CACM*, 29(3):184–201, March 1986.

[10] John Kubiatowicz et al. OceanStore: An architecture for global-scale persistent storage. In *Proc. ACM ASPLOS*, November 2000.

[11] John Garrett and Donald Waters. Preserving digital information: Report of the Task Force on Archiving of Digital Information, May 1996. Accessible at http://www.rlg.org/ArchTF/.

[12] Andrew Goldberg and Peter Yianilos. Towards an archival intermemory. In *Proc. ADL*, 1998.

[13] Jim Gray, Pat Helland, Patrick O'Neal, and Dennis Shasha. The dangers of replication and a solution. In *Proc. ACM SIGMOD*, June 1996.

[14] Alan Heminger and Steven Robertson. Digital Rosetta Stone: A conceptual model for maintaining long-term access to digital documents. In *Proc. 6th DELOS Workshop on Preservation of Digital Information*, June 1998.

[15] Hui-I Hsiao and Devid DeWitt. Chained declustering: A new availability strategy for multiprocessor database machines. In *Proc. ICDE*, February 1990.

[16] J. J. Kistler and M. Satyanarayanan. Disconnected operation in the coda file system. *ACM TOCS*, 10(1):3–25, February 1992.

[17] Carl Lagoze, Jane Hunter, and Dan Brickley. An event-aware model for metadata interoperability. In *Proc. ECDL*, September 2000. In LNCS vol. 1923.

[18] Edward Lee and Chadramohan Thekkath. Petal: Distributed virtual disks. In *Proc. 7th ACM ASPLOS*, October 1996.

[19] Nuno Maria, Pedro Gaspar, Antonio Ferreira, and Mario Silva. Information preservation in ARIADNE. In *Proc. 6th DELOS Workshop on Preservation of Digital Information*, June 1998.

[20] Silvano Martello and Paolo Toth. *Knapsack Problems: Algorithms and Computer Implementations*. J. Wiley and Sons, Chichester, New York, 1990.

[21] David Patterson, Garth Gibson, and Randy H. Katz. A case for redundant arrays of inexpensive disks (RAID). *SIGMOD Record*, 17(3):109–116, September 1988.

[22] Sandra Payette and Carl Lagoze. Policy-carrying, policy-enforcing digital objects. In *Proc. ECDL*, September 2000. In LNCS vol. 1923.

[23] Arcot Rajasekar, Richard Marciano, and Reagan Moore. Collection-based persistent archives. In *Proceedings of the 16th IEEE Symposium on Mass Storage Systems*, March 1999.

[24] David S. H. Rosenthal and Vicky Reich. Permanent web publishing. In *Proc. USENIX Annual Technical Conference*, June 2000.

[25] Jeff Rothenberg. Ensuring the longevity of digital documents. *Scientific American*, 272(1):24–29, January 1995.

[26] Harjinder Sandhu and Songnian Zhou. Cluster-based file replication in large-scale distributed systems. In *Proc. ACM SIGMETRICS*, June 1992.

[27] Ouri Wolfson, Sushil Jajodia, and Yixiu Huang. An adaptive data replication algorithm. *ACM TODS*, 2(2):255–314, June 1997.

Cost-Driven Design for Archival Repositories *

Arturo Crespo and Hector Garcia-Molina
Computer Science Department
Stanford University
Stanford, CA 94305-2140, USA
{crespo,hector}@db.stanford.edu

ABSTRACT

Designing an archival repository is a complex task because there are many alternative configurations, each with different reliability levels and costs. In this paper we study the costs involved in an Archival Repository and we introduce a design framework for evaluating alternatives and choosing the best configuration in terms of reliability and cost. We also present a new version of our simulation tool, ArchSim/C that aids in the decision process. The design framework and the usage of ArchSim/C are illustrated with a case study of a hypothetical (yet realistic) archival repository shared between two universities.

1. INTRODUCTION

Digital information can be lost for a variety of reasons: magnetic decay, format and device obsolescence, human or system error, among many others. A solution is to build an *archival repository* (AR), a system capable of storing and preserving digital objects (e.g., movies, technical reports) as technologies and organizations evolve [1].

Designing an AR is difficult, as there are many configuration options and uncertainties about the future. For example, one must decide how many sites to use, what types of disks or tape units to use, what and how many formats to use to store documents, how frequently to check existing documents for errors, what strategy to use for error recovery, how often to migrate documents to a more modern format, and so on. On top to this, the designer needs to predict future events such as the reliability of sites and disks, survivability of formats, how many resources will be consumed by the recovery algorithms, how frequently the recovery algorithms will be invoked, how many user accesses will be made to the documents, and many other uncertainties.

Two important factors must be considered in AR design: the *level of assurance* (e.g., on average a document will not be lost for 1000 years) and the *cost* (e.g., an initial investment of 1 million dollars and yearly expenses of 100 thousand dollars). There has been some research on predicting the level of assurance of a given AR [4], but there has been little or no work on predicting the cost of an AR.

Predicting the cost of an AR is difficult task. First, we need to estimate the cost for each "event" such as the AR creation, the failure and repair of a disk, etc. For many of these events, we may also have to predict when they will happen. For example, since we do not know when a disk will fail, we cannot deterministically predict when and how often we will pay for its repair. Second, we may not know for certain future costs, so we may have to represent them with probability distributions (e.g., the price of a disk may be between $100 and $150). As we will see in this paper, deriving cost estimates and likelihoods for a given AR requires a lot of "guess work." However, the alternative of ignoring costs altogether can easily lead to systems that are overdesigned and overpriced, or that do not meet user expectations.

In this paper we show how AR costs (and failures) can be modeled, albeit in a rough way, so that rational decisions can be made. In particular, we present a complete design framework for making cost-driven decisions about ARs, and a powerful simulation tool, *ArchSim/C* that aids in the process. Our design framework is based on Decision Analysis (DA) theory [9] and we believe that it is a good way of structuring the design of archival repositories. ArchSim/C can model important configuration options, such as multiple formats, preventive maintenance, and failure distribution functions. By using specialized techniques, ArchSim/C is able to provide cost and reliability information for a configuration in a time frame that allows the exploration and testing of different policies. To illustrate the framework, we use as a running example a case study based on a hypothetical AR of MIT/Stanford technical reports.

The contributions of this paper are:

- An in-depth study of the costs involved in an AR.
- A comprehensive design framework for making AR cost decisions.
- A new version of our simulation tool that can predict the reliability *and* the cost of an AR.
- A demonstration of the framework in a case study.

*This work was partially supported by NSF.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

2. ARCHIVAL REPOSITORIES

We define an *Archival Repository* (AR) as a repository that guarantees long-term data survivability. In this section we study the elements of a typical archival repository (AR), so we can later evaluate their reliability and cost. A typical AR is formed by a *data store* that can fail and an *archival system* (AS) that ensures long-term survivability of its documents. The AS provides fault tolerance by managing multiple *materializations* for each document. A materialization is the set of all the *components* necessary to provide some sort of human access to a document. For example, a materialization may include the bits, disks, and format interpreters necessary to display a technical report. Figure 1 shows the AS main functions (in solid-line boxes), the non-fault-tolerant store (in a dashed-line box), and the archival documents. The arrows represent the runtime interactions between the elements. This representation can model many existing archival systems including the Computing Research Repository [8], the Archival Intermemory Project [7, 2], and the Stanford Archival Vault [3].

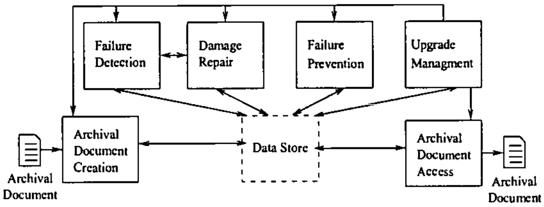


Figure 1: Archival Repository Model

The data store encompasses the set of components, such as sites, disks, or format interpreters that make materializations accessible. Because the store is not fault tolerant, materializations may be lost. A materialization is considered lost when *any* of its components has failed. If *all* of the materializations of a document are lost, then the document is considered lost.

The AS monitors materializations, and when a failure is detected, attempts to repair it. Further, the preventive maintenance module take actions to avoid failures. For example, the AS may copy components that are stored on a disk that is close to the end of its expected life, onto a newer disk. The AS may also initiate system upgrades to take advantage of newer technologies. In summary, the main functions of the AS are: document creation, document retrieval, failure detection, failure repair, preventive maintenance, and upgrade management.

3. AR COSTS

We can see the life of an AR as a sequence of events such as the failure of a disk, user access to a document, and making a copy of a document. A *cost event* is an event that has an economic impact. The definition of what has an economic impact will vary from organization to organization. For example, an organization may define economic impact as anything that has an impact on the accounting books (e.g., expenses and depreciation of capital equipment). Another organization may extend the definition to include expenses incurred by the users (e.g., expenses because of unavailability of the system). Cost events may or may not be triggered by a physical event. For example, an organization may buy a maintenance contract for disks under which an annual fee is paid in exchange for free repairs of all disks that may fail during the year. In this case, the failure of a disk (a physical event) will not have any economic impact (and thus it is not a cost event), while the annual payment for the maintenance contract (which is not an AR physical event) will be a cost event.

How can we compare the total cost of two ARs? Ideally, we would like to assign a monetary value (e.g., dollars) to each event in the sequence, and then, aggregate those costs into a single value. Having done that, we can simply choose the system with the lowest cost. If we know the sequence of costs events and each future cost can be deterministically computed (e.g., disk prices will decreased by 5% annually from current prices), this is a feasible task. In this case, we compute the monetary value of each event and we aggregate them by computing the average annual system cost, or ASC (e.g., the AR will cost $100,000 annually).

However, as we explained in the introduction, we may not know the exact sequence of cost events. In addition, we may not know deterministically future costs and we may have to represent costs as probability distributions. In this case, the system is characterized by a probability distribution of ASCs (e.g., with probability 0.3, the annual cost will be $100,000; with probability 0.7 it will be $150,000). Although in simple cases the ASC distribution can found analytically, in general, we have to rely on simulations to obtain an approximation of the distribution.

With probabilistic ASCs, choosing the best AR is not straightforward. The general problem of choosing between two probability distributions of costs has been studied in [11]. In the extended version of the paper [5], we describe several ways of choosing between distributions, but in this paper, we will use the simplest way of selecting the best of two cost distributions: namely, we will choose the one with the lowest mean (average). Given this, throughout the paper, we will frequently talk about the mean annual system cost (MASC) as representative indicator for the distribution of the average annual system cost.

3.1 A Taxonomy of Cost Sources

In this section, we classify the cost sources in an Archival Repository. Our goal is to understand those sources, so we can use them as building blocks for cost events. The problem of classifying cost sources for computer systems has been studied in [6], but we are not aware of any studies for the specific case of Archival Repositories. The most common cost sources in an Archival Repository can be broken down into the following categories:

Hardware and Software: This category includes all the expenses (including lease fees) for servers, clients, disks, software, the network, and peripherals. Although, this is the most obvious source of cost for a computer system, it only represents about 20% of the total cost for the system [6]. Usually, it is easy to estimate the cost of the initial hardware and software, as we can just use market prices. However, for replacement hardware and software, this is a more complicated process as we need to predict future prices. Moreover, this prediction is often obtained in the form of a probability distribution, based on current trends, as well as possible future technological developments. For example, when predicting disk costs ten years from now, we may conclude that with 60% probability a terabyte will cost $10 or less, with

80% probability it will cost $15 or less, and with 99% probability it will cost $50 or less.

Non-labor Operational costs: This category includes all the costs (different than labor) necessary to maintain the AR operational. For example, these cost will include the electricity consumed by the system, air conditioning, and physical space. As with the Hardware-and-Software category, it is easy to estimate the initial cost of non-labor operational costs. For future costs, the major challenge in estimating theses costs is trying to predict the future needs of the components of the AR. For example, technological improvements may reduce the need for physical space, but they may increase the need for air conditioning.

Labor costs: This category includes all the human-related costs necessary for the AR. In particular, this will include management (e.g., system administrator), support (e.g., help desk), and development (e.g., application developers).

Information acquisition: Information is sometimes free (e.g., technical reports, thesis), but in general, libraries need to pay for information (e.g., journals). This payment may be a one-time fixed cost, periodic payments (subscriptions), or, more infrequently, pay per use. In some context, we may choose to ignore this cost and considered it "the cost of doing business" (i.e., the library will have to provide access to information even if they do not have an AR). However, we should consider this cost if the creation of the AR will change the way the library pays for information (for example, moving from a paper-based library to a digital library with publisher charging different amounts for paper journals than for digital journals).

Insurance: We define insurance as any agreement where an outside party takes the risk of a specific failure in an AR component in exchange of a fixed payment. An example of insurance is a maintenance contract where the library pays a fixed amount to a company that replaces failed disks. Insurance is important not only because of its direct cost, but also because it can reduce the *variance* of the AR cost. If we are able to "insure" all uncertain events, then we will have a deterministic ASC.

Unavailability: If the system is not available, there may be an economic impact for the organization. Unavailability may be caused by a system failure, but it can also occur when system resources are diverted to maintenance or repair tasks. For example, a user may be blocked because the system is checking the storage device that holds the requested document for errors. Similarly, the system may only be able to handle a fraction of the normal users when it is migrating documents to a new format.

Measuring the cost of unavailability is a difficult task. If users pay for access, we may be able to assign a direct cost corresponding to the lost income. If users to do not pay directly, we still want to penalize the system for unavailability, lest we end up with a design that disregards user needs. One way is to assume that users will access an alternate system (even if the content is not available elsewhere). We could, for instance, assume that the alternate system is equivalent to our AR. Thus, if it costs $500 per day to operate the AR, the cost of unavailability will be $500. We could also consider a commercially available alternate system. For instance, an average search on Dialog (SciSearch database) costs $6, so if we cannot satisfy say 1500 requests while doing preventive maintenance, then the additional cost will be $9000.

Cost of losing a document: Even though our objective is to preserve all documents in the repository, in some circumstances one can put a price on document loss. For example, an organization may choose between archiving certain documents or recreating them. In this case, the cost of losing the document would be the cost of recreating it. Of course, there are cases when we cannot put a dollar value on losing a document (e.g., the diary of a famous person), so we can use an arbitrarily large cost.

3.2 A Taxonomy of Cost Events

In the previous section we studied cost sources. In this section, we use those sources as building blocks for the most common AR cost events. Cost events can be broken down into the following categories.

AR creation: Starting an archival repository involves a large number of expenses. Hardware and software need to be bought, infrastructure needs to be put in place, new personnel needs to be hired, and so on. For instance, the creation of an AR with 100 disks would involve a server (about $5000), the disks ($500 per disk for a total of $50000), installation costs (one consultant at $1000), renting and furnishing an office space ($800 for the realtor that finds the place and $2000 for furniture and other necessary improvements for the rented space), and loading of the documents (five days of work supervised by a system administrator, about $1200) for a total of about $60000. The AR creation cost can be amortized over time. Amortization can done by either charging a fraction of the startup cost over fixed periods of time (in which case it would be an operational cost) or over each usage of the system (in which case it would be a document access cost).

Document Access: When accessing a document, the AR may incur acquisition costs or labor costs (e.g., the cost of the operator who retrieves and mounts a tape).

AR operation: The total operational cost of the AR would include the office space taken by the repository, the necessary utilities (electricity, network, etc.), and the cost of the people in charge of keeping the system running. For example, in San Francisco, the average cost of office space is $380 per square meter per year, so if we assume the repository occupies a small office of $8m^2$, the annual space cost will be $3040 per year. Reasonable estimates for utilities are $4000 for electricity and $3000 for network connectivity. Finally a quarter-time system administrator and a 1/8th librarian would cost about $20000 per year. This results in a total operational cost of about $30000 per year.

Failure Detection: To enhance reliability, an AR needs to periodically check for failed components (e.g., corrupted tape). When performing failure detection, we should not only take into account the cost of the detection itself (e.g., moving a tape from storage, mounting the tape on the reader, checking the tape, and returning the tape to storage), but also the cost of unavailability that it may generate. In Section 5.4 we will see an specific example of how to compute these costs.

Repairs: When the AR fails and needs to be repaired, cost events may be generated (if we do not have a maintenance contract). For example, when a hard drive fails, we may need to buy a new hard drive (about $500), remove and install the new one ($100 for the time of the technician), and restore the content of the failed drive into the new disk ($200 for the network cost and the unavailability caused by the transfer).

Preventive Maintenance: Before a component fails, we may want to transfer the information to a new component. For example, if we know that tapes can survive 20 years, we may decide to copy old tapes into new ones after 10 or 15 years. The cost associated with a preventive maintenance event, includes the cost of the new media, the transfer of the information, and the possible unavailability that this task may create in the AR.

Upgrades: Upgrades are similar to preventive maintenance, i.e., we transfer information from old components to new ones. However, the motivation and the cost implications of an upgrade are different. We perform upgrades to obtain some advantage from modern technology. These advantages may go beyond improved reliability (which is the reason for preventive maintenance) and may include reduced cost. For example, when upgrading to modern hard drives, we may gain reduced operational costs (e.g., if they require less administrator time, less power, or less physical space). Therefore, after an upgrade, we need to reconsider all other costs in the system and change the cost events appropriately.

4. AR RELIABILITY

The reliability of an AR gives the likelihood that the system will work for a given period of time. Formally, the reliability is the conditional probability that no "failures" have happened in the time interval $[0, t]$ given that the system was operational at $t = 0$. There are many ways we can define an AR failure. It could be the loss of a document, the loss of a certain fraction of the collection, or even the loss of some specific set of documents. In this paper we will take the most stringent criteria: loss of any single document.

As with costs, we summarize the probability distributions for time to failure by its mean. So, we use the mean time to failure (MTTF) as representative indicator for the distribution of the time to failure. For instance, an AR with a MTTF of 100 years is expected to survive 100 years, i.e., if we build say 10 identical ARs, and average the time when each fails, we get about 100 years.

It is important to note that MTTFs can be used to compare ARs, even if we expect their configuration to change relatively soon. For instance, say we compare two ARs, A and B, using a current hardware configuration, and find that A's MTTF is 50 years, while B's is 200 years. One may be tempted to think that because the current hardware will be replaced in say 15 years, then the longer MTTFs are meaningless. However, this reasoning is incorrect. System B, with its longer MTTF, is significantly less likely to fail during the first 15 years than A and is hence preferable. In 15 years, when we change the configuration of the AR, we can re-evaluate its MTTF, and again decide what are the best options based on the predicted MTTFs at that time. In summary, MTTFs can be used to compare systems even over short periods of time.

To estimate system reliability, we need to identify the undesired events, such as the failure of a disk or an operator error, that may lead to a failure. A document is lost if the bits that represent it are lost, and also if the necessary components that give meaning to those bits are lost. An undesired event does not necessarily cause information loss. In fact, we have seen that if the AR keeps two copies of a document, and the disk holding one of the copies fails, then the document is not lost. It would take a second undesired event affecting the second copy to cause information loss.

5. DESIGNING AN AR

Our goal is not simply to evaluate a given AR, but instead to design an AR that meets our cost and reliability targets. For instance, we need to decide how many document copies to keep, what formats to store them in, how frequently to check for errors, and so on, in order to attain some desired reliability and maximum cost. To aid the design, we use a framework based on Decision Analysis [10]. We show our design framework in Figure 2. The framework is a cycle where we first formulate our objectives. Then, we identify the uncertainties (e.g., when a disk will fail). A large number of uncertainties can make the system difficult to analyze, so we next identify and eliminate the uncertainties that do not have a critical influence on the overall performance of the system. Then, we assess the probability distribution of the uncertainties and predict the performance of the AR design. Finally, we perform a sensitivity analysis to appraise our design. If we find a problem with the recommended design (for example, we discover that one of our initial assumptions is incorrect), then we iterate over the cycle.

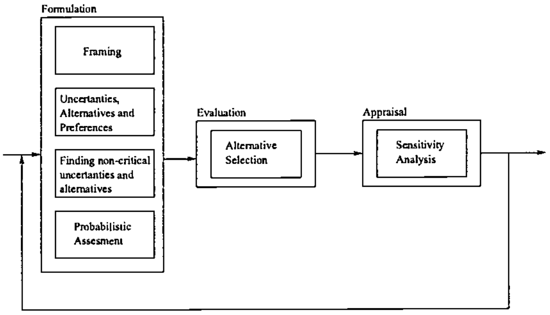


Figure 2: AR Design Cycle

To illustrate the design process, we will use a case study of a Stanford-MIT technical reports AR. At each step, we will discuss the assumptions or decisions made in this sample design. We will also present simulation results for this case study to show the types of conclusions that can be reached.

5.1 Framing the problem

The first is to clearly define the success criteria for the design. Typically, the criteria will include the archival guarantees (MTTF) and the cost of the AR (MASC). Possible goals can be: (i) Maximize the MTTF of an AR such that the MASC is less than a given amount. (ii) Minimize the MASC of an AR for a minimum MTTF. (iii) Maximize some combination of the MTTF and MASC. In other words, we want to transform the MTTF and MASC to a common metric (let us say dollars) and maximize its combination. For example, if documents can be recreated, the organization may be able to assign a dollar value to losing a document as we discussed in Section 3.1.

In our case study, the goal will be to have a repository with a MTTF at least equivalent to that of standard paper (100 years) with the minimum MASC possible. A failure is defined as the loss of one or more documents.

When designing an AR, some decisions are taken before starting the design process (policies), others are delayed until the implementation of the system (tactics), while the rest are the focus of the design (strategies). For example, in our case study we assume that the AR will cover Stanford and

MIT technical reports (a policy) and that the decision on the specific brand of the hard drives that the AR will use can be defered until implementation time (a tactic). It is important for the design team to agree on which decisions are policies or tactics, as no time should be spent studying them during the design process.

5.2 Identifying Uncertainties, Alternatives, and Preferences

Uncertainties are probabilistic factors that affect the AR (e.g., the time when the disk will fail). Despite their name, we might actually have some control over an uncertainty. For instance, even though we do not know when a disk will fail, we may be able to choose between disk with different MTTFs. When we control the value of an uncertainty completely, we will call it a *variable*. For example, if we assume that disk prices will decrease exactly 5% per year (and we know the current price), then the cost of a replacement disk becomes a variable.

Alternatives are the different designs that we have available. For example, in our case study, we may consider:

- ARs with disks with MTTF of either 3, 5, 10, or 20 years.
- ARs with failure detection intervals of 30, 60, 120, or 720 days.

The combination of these different values results in 16 possible configurations that we need to evaluate.

5.3 Modeling an AR

An important decision is the level of granularity in the model. If we have too little granularity, then we will have complex uncertainties that are difficult to analyze. If we have too much granularity, the number of variables will be high, making the analysis of the model difficult and even impossible. For example, in our case study we decide to use disks, sites, and formats as the lowest level of detail (in contrast to choosing documents, files, or even bits). Thus, we only need to quantify how much money disks, sites, and formats will cost and how will they affect MTTF. This is much simpler than trying to find the MASC and MTTF of the AR as a whole (not enough granularity), or the MASC and MTTF of every single file (too much granularity).

To model and evaluate a particular AR configuration, we propose an extension of the model presented in [4]. In particular, our extension adds cost events and their associated cost distributions. Recall that our model of an AR has two major elements: a non-fault-tolerance data store and an archival system (AS) that ensures long-term survivability of the information. To model the store, we need to define:

- How many component instances and types are present in the system: that is, how many disks, formats, etc., are available.
- Time distributions for component failures. Many components have two different failure distributions, one during archival and another during access. For example, a tape is more likely to fail when it is being manipulated and mounted on a reader than when it is stored. Therefore, each component may have two failure distributions: during archival and during access. For some components, such as disks or sites, the access and archival distributions will be the same.
- Time for performing a component check. This distribution describes how long it takes to discover a failure (or to determine that a component is good), from the time the check process starts. For example, consider checking a tape. This may involve getting the tape from the shelf, mounting the tape, and scanning the tape for errors.
- Time for repairing a component failure. This distribution describes how long it takes to repair a component. This distribution may be deterministic (if the component can be repaired in a fixed amount of time). Repair time may be "infinite" if the component cannot be fixed.

In addition, there is an important interdependency between components. Specifically, the failure of one component may cause the failure of another component. For example, if a site fails (e.g., because it was destroyed by a fire), then all the disks at the site will also fail. This failure dependency is captured by a directed graph. For example,

- **AR Description**
  - Initial collection: 200,000 documents. No documents created after startup. Each document, d, will have materializations:
    - $\langle d, MIT, disk_i \rangle$,
    - $\langle d, MIT, disk_k \rangle$,
    - $\langle d, Stanford, disk_x \rangle$,
    - $\langle d, Stanford, disk_y \rangle$.

    Where MIT and $Stanford$ are the two sites; and $disk_i$, $disk_k$, $disk_x$, and $disk_y$ are different storage devices.
  - Number of components and types: 100 storage devices in each site, 2 sites.
  - Failure dependency graph: $site \rightarrow disk$, when the disk is in the given site.
- **Policies**
  - Document Creation policy: for each document, two materializations are created, one in each site.
  - Document to Materialization: read from any materialization.
  - Failure detection algorithm: complete scan of all disk. Site failure detection is instantaneous.
  - Damage Repair algorithm: discard bad component and replace with new component instantaneously.
  - Failure prevention algorithm: none
  - Upgrade policy: none
- **Distributions (unknown for now)**
  - Disk Failure dist. during access (time)
  - Disk Failure dist. during archival (time)
  - Disk Failure Detection success dist. (probability)
  - Disk Repair success dist. (probability)
  - Disk Failure Detection interval dist(time)
  - AR Creation Cost. (dollars)
  - AR Operational cost dist. (dollars)
  - Disk Failure Detection cost dist. (dollars)
  - Disk Repair cost dist. (dollars)

Figure 3: Archival Repository Model Parameters

Variable	low	base	high
Disk MTTF during access (years)	20×0.9	20	20×1.2
Disk MTTF during archival (years)	20×0.9	20	20×1.2
Success of a Failure Detection (probability)	1	1	1
Success of a Failure Repair (probability)	1	1	1
Failure Detection Interval (days)	120×1.5	120	120×0.9
AR Creation Cost (dollars)	55000	60000	70000
AR Operational Cost (dollars/year)	200	300	400
Failure Detection Cost (dollars/run)	$1000 + 1200 * 6$	$1200 + 1500 * 6$	$1400 + 1800 * 6$
Repair Cost (per replaced disk)	$450 + 100 + 164$	$500 + 100 + 204$	$600 + 100 + 244$

Figure 4: Base values

an arrow between "Site A" and "Disk 1" in the interdependency graph means that if "Site A" fails, then "Disk 1" will also fail.

To model the AS we need to define:

- Document Creation algorithms and their associated cost distributions: When a new document is added to the AR, the AS uses the document creation algorithm to create enough materializations to ensure survivability of the document. This action may create one or more cost events, each with a different cost distribution.
- Document Access algorithms and their associated cost distributions: how a document request is transformed into requests for the appropriate components, and the associated costs of that operation.
- Failure Detection algorithm and their associated cost distributions: As explained earlier, the AS scans the store looking for damaged or lost materializations. When a damaged materialization is found, a damage repair algorithm is started (as described below).
- Damage Repair algorithms and their associated cost distributions: After a failure has been detected, the AS attempts to repair damaged components. There are many strategies to repair a damaged document that are discussed in [4].
- Failure Prevention policies and their associated cost distributions: The AS scans the store and takes preventive measures so materializations are less likely to be damaged. For example, the AS may copy components that are stored on a disk that is close to the end of its expected life, into a newer disk.
- Upgrade algorithms and their associated cost distributions: A technology upgrade may change the algorithms used by the AS as well as the cost distributions.

Figure 3 summarizes the AR model for our case study. The failure and cost distributions for the model and described in the next subsections. Note that for simplicity the model assumes no format or site failures. (Our methodology can of course handle a more general model.)

5.4 Transforming Non-Critical Uncertainties into Variables

We can simplify the AR analysis by considering as variables the uncertainties that have little impact on MTTF and MASC. For example, if the distribution for disk prices introduces little variation on the total cost, we might as well replace it with its mean. Eliminating uncertainties can save substantial analysis and simulation effort. We call uncertainties that have a large impact on MTTF or MASC the *critical uncertainties*. The remaining ones are called *non-critical uncertainties* or, given that we are fixing them, just *variables*. In this subsection we will see how can we identify critical and non-critical uncertainties.

To determine the impact of an uncertainty, we need to find its distribution. Obtaining an exact probability distribution for each uncertainty may take a significant effort with a limited payoff, so instead we approximate the distributions by using just three values: low, base, and high which correspond to the distribution 10, 50, and 90 percentile. Finding the appropriate low, base, and high values for an uncertainty is more an art than a science. Only experience and a good understanding of the AR components allow one to make these predictions.

After approximating the distributions of the uncertainties, we assess their impact by using a Tornado Diagram. A Tornado Diagram shows the system performance (MTTF or MASC) for the low/base/high value of each uncertainty (while keeping all other uncertainties at their base values). An example of a tornado diagram can be found on Figure 5. We will explain this diagram in detail later in this section, but for now, we can see some uncertainties (such as the Disk Failure) impact MTTF significantly while others (such as Failure Detection Cost) have little or no impact.

Returning to our case study, let us consider the case where disks have a MTTF of 20 years and we scan the repository every 120 days. (In practice we would do a similar evaluation for each of the other 15 alternatives discussed in Section 5.2). First, we obtain a *rough* range for the values of the variables. These ranges are shown in Figure 4. The choice of these values is highly subjective, but, nevertheless, we will attempt to describe the rationale that an expert may have followed to reach these values.

Disk failure during access and archival: For these two uncertainties, we choose to have the same distributions, since disks do not fail significantly more when accessed. We use as base value the MTTF advertised by the manufacturer (20 years). We assume that there is little variation in MTTF, so we will assign a low value of 90% of the advertised MTTF and a high value of 120% of the advertised MTTF.

Success of failure detection and a repair: For these two uncertainties, we assume that the probability of success is 1. In other words, we are assuming that there are no hidden failures (i.e., if a disk is defective, we can always tell) and

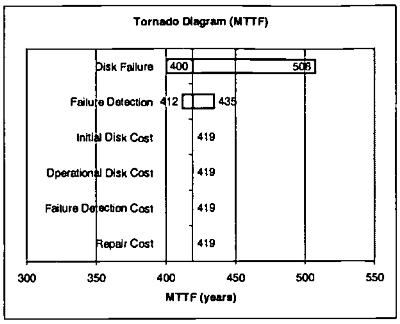


Figure 5: Tornado Diagram (MTTF)

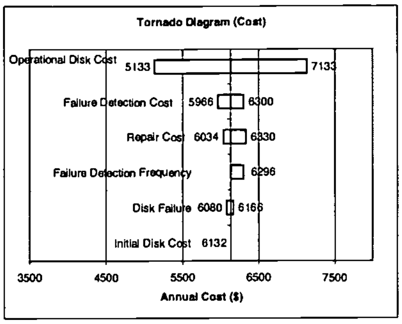


Figure 6: Tornado Diagram (Cost)

that all repairs are successful (i.e., we can always replace a defective disk with a working one). Note that the later does not mean that we can always repair a document. It just means that we are always able to install a new disk and to copy the content of the failed disk from alternative sources *if it is available.*

Failure Detection Interval: This uncertainty shows that the assignment of low, base, and high valued need not be symmetrical. For instance, we assigned a low value of 1.5 times the targeted mean time to detection (120 days), while we assigned a high value of 0.9 times the detection time. In other words, it is more likely that detection will be slower rather than faster.

AR Creation Cost: Using the rationale presented in Section 3.2, we will estimate the initial AR cost to be $60000. To allow for error, we will choose a low value of $55000 and a high value of $70000. We will assume that this will be a one-time cost (i.e., no amortization will be done over time).

Operational Cost of a Disk: As illustrated in Section 3.2, we use a total operational cost of about $30000 per year with a low value of $20000 and a high value of $40000 per year to allow for errors. This total operational cost divided by the number of disks (100 per site) results in a operational cost per disk of $200 to $400.

Cost of the Disk Failure Detection Algorithm: This is probably the hardest uncertainty to estimate. We divide the cost of the detection algorithm in two components. The first component represents the direct cost of running the algorithm, while the second component reflect the cost of service unavailability. The direct cost of the failure detection algorithm includes the time required by the System Administrator to start the scan and correct any problems with the scan (assuming these tasks are not included in the administrator's salary already). If we assume that failure detection involves 5 days of part-time work, the cost will be about $1200 per run (see Section 3.2).

To estimate the unavailability cost, we will assume that users will use an alternate commercial service. Using the costs of Section 3.2 for 1500 missed user requests, we price unavailability at $9000. Therefore, the total cost of running the failure detection algorithm is about $10200. To allow for error, we will choose a low value of $8200 and a high value of $12200 per run.

Cost of the Disk Repair: The repair cost is equal to the cost of adding a new disk ($450 to $600) plus the cost of removing the disk ($100) and a fixed amount for the resources involved in copying the data from the alternate sources onto the new disk (equal to twice the cost of running the detection algorithm on a single disk, this is, for the base cost, $10200/100 * 2 or $204).

We are now ready to generate the Tornado Diagrams. We use ArchSim/C to simulate the performance of the system. At this stage, we do not want to run the full fledged simulations (which may take a significant amount of time). Instead, we run *fast* simulations with broad confidence intervals (requiring fewer repetitions) and considering all uncertainties, except disk failures, to be deterministically fixed at their base values. Fixing the value of the variables speeds up the simulation as we do not need to compute a random value for each event and allows us to group events. For instance, instead of generating a random value for each repair cost, we just count the number of repairs that were performed during the simulation and multiply by the fixed cost of making a repair. We treat disk failures differently because deterministic failure times would cause all disks to fail at the same time (and all data would be lost).

To generate the Tornado Diagrams, we evaluate the AR reliability and cost for the proposed design with all the variables at their base values. Then, we modify each variable independently (while keeping all others at their base value) to its high and low value and evaluate performance again. Each tornado diagram summarizes $1 + 2 \times$ variables simulations (one simulation for the base case and two for each variable). In our case, this results in a total of 13 simulation per tornado diagram. We show the result of our simulations in Figure 5. We can see that most of the MTTF variation (95.7%) comes from the disk MTTF variation. Therefore, with respect to this metric, we can safely assume that the other uncertainties are noncritical and can be fixed at their base values.

Figure 6 shows the equivalent diagram for costs. In this case, 94% of the cost variation is produced by the operational cost of the disks. Therefore, with respect to this metric, we can safely assume that the other uncertainties are non critical and can be fixed at their base values.

In conclusion, we only need to consider disk failures and disk operational costs as critical uncertainties, for the case of a design with disks having a MTTF of 20 years and failure detection interval of 120 days. To complete the analysis, we need to repeat the process with the other 15 configurations. Although we do not show the results for the other cases, the conclusion is the same: only disk MTTF and cost are critical. (In general, the conclusions could vary from scenario to scenario, but this does not occur in our case study.)

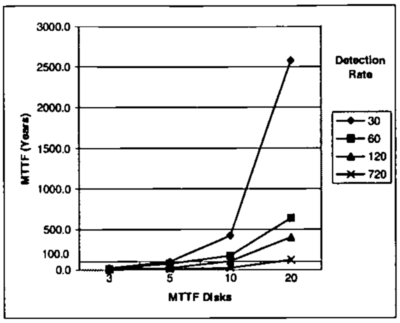


Figure 7: MTTF for base values

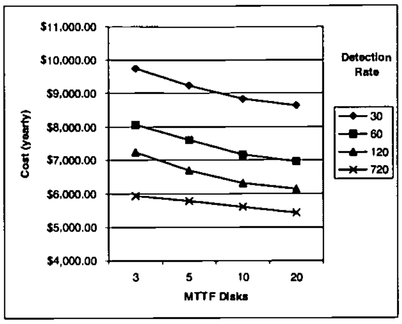


Figure 8: MASC for Base Values

5.5 Eliminating Futile Alternatives

Let us turn our attention to the available alternatives. For that purpose, we used ArchSim/C again to run fast simulations. The results are in the graphs of Figure 7 and 8. From the graphs, we can see that a detection interval of 720 days never achieves our required minimum of a MTTF of 100 years (it barely achieves it for a disks with MTTF of 20 years, but the 90% confidence interval includes values below 100 years). Similarly, disks with MTTF of 3 years also never achieve our required minimum MTTF. Note that these results are based on fast simulation where all uncertainties, except the MTTF of disks, are fixed. If we are very aggressive and eliminate too many alternatives, we might eliminate the alternative that may happen to be the best when running the full simulation. On the other hand, by eliminating some alternatives, the time to run the full-fledged simulations later is reduced. For this case study, we not consider further disks with a MTTF of 3 years or detection time of 720 days. If we were more aggressive, we could have also eliminated disks with MTTF of 5 years (except when the detection interval is 30 days).

Regarding MASCs, the preliminary analysis shows a surprising result. The MASC of an AR with costly, but more reliable, disks ends up lower than that of an AR with the cheap, less reliable, disks. This is because of the cost of buying a new disk (when the cheap disk fails) and transferring the information to it. Therefore, we drop disks with a MTTF of 3 years, and detection intervals of 720 days, and reduce our alternatives from the original 16 to just 9.

5.6 Probabilistic Assessment of Uncertainties

For our final analysis we may need to assess the probability distributions of uncertainties more precisely. In practice, we will rely on experts to produce these distributions. Techniques for probability distribution elicitation are described in [12].

To illustrate, in our case study, we model disk failures with an "infant mortality" distribution. This kind of distribution, typical for electronic devices, has two phases. First, when most manufacturing defects will cause a failure, the probability of failure is high, but drops sharply over time. In the second phase, the probability of failure is constant. To model this distribution we use three parameters: time span for the first phase, percentage of devices failing in the first phase, and the probability of failure in the second phase. For our disks, we use a distribution where 10% of the disks fail within the first 30 days (i.e., an exponential distribution with mean 285 days) and, after that disks fail following an exponential distribution with mean 20 years.

To model the operational cost of disks, we assume that the library will sign one-year maintenance contracts. Although the price is fixed for one year, from year to year, the price specified in the contract may change due to market conditions. We will use an uniform distribution between $200 and $400 per disk to capture those market fluctuations. Using these more complex distributions makes our predictions more accurate, but also makes evaluation much harder. Fortunately, ArchSim/C can handle such general distributions.

5.7 Evaluating an AR Design

ArchSim/C receives as input an AR model (including costs), a stop condition (e.g., stop when the first document), a simulation time unit (minutes, hours, days, etc.), and the number of repetitions. ArchSim/C outputs the mean time to failure (mean time to stop condition), a cost metric, and a confidence interval for both the MTTF and the MASC.

ArchSim/C follows the structure of a traditional simulation tool. Each component of the AR model registers future events in a timeline. For example, when a disk is created, the simulation uses the disk failure distribution to compute when the disk will fail; then, it registers the future failure event in the timeline. The simulation engine advances time by calling the module that registered the first event. This module may change the state of the repository and register more events in the timeline. Additionally, the module may contact the Cost Manager and record some cost involved with its operation. After the module returns, the simulation engine checks for the stop condition and, if it has not happened, it advances to the next event, in chronological order. If the stop condition has occurred, the simulation stops and records the point on the timeline when this happened and the total cost incurred up to that time. The engine keeps re-running the simulation until the number of repetitions requested by the user is reached. At that moment, ArchSim/C computes the MTTF and the MASC by averaging the recorded time to failure and costs at the end of each repetition. ArchSim/C also compute the confidence interval for those values. Further details on ArchSim/C, including techniques and features that speed up significantly the simulation can be found in [4].

In the previous sections we concluded that the most promising alternatives were the ones with disks with a MTTF of 5 to 20 years and a detection/repair interval of 30 to 120 days. We also concluded that we would consider the MTTF of the

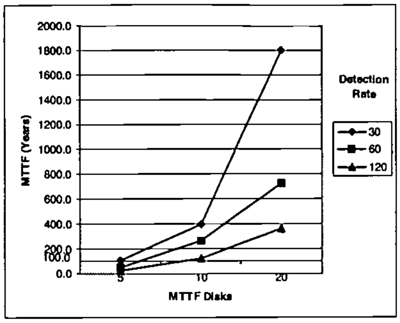


Figure 9: MTTF Evaluation

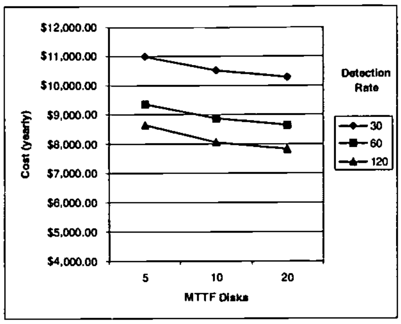


Figure 10: Cost Evaluation

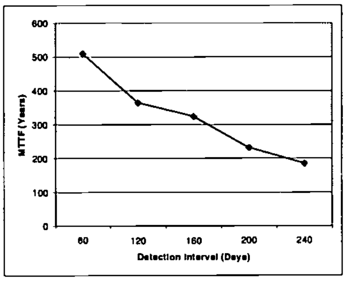


Figure 11: MTTF Sensitivity Analysis (Detection Interval)

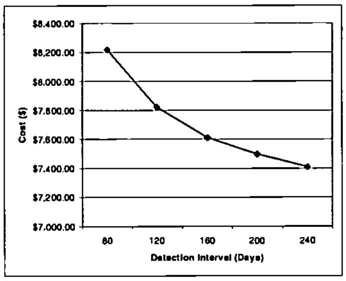


Figure 12: MASC Sensitivity Analysis (Detection Interval)

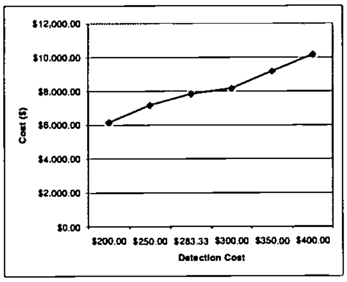


Figure 13: MASC Sensitivity Analysis (Detection Cost)

disks and the yearly operational cost of the disks as critical uncertainties and the rest, as variables. Using this setup, we use ArchSim/C to fully simulate the AR and obtain its reliability and cost.

In Figure 9 we see the MTTF of the AR for different configurations. From the figure, we conclude that we need disks with a MTTF of either 10 or 20 years and detection intervals of 30 to 120 days to achieve our target MTTF of 100 years.

In Figure 10 we see that the least expensive alternative is the one with a detection interval of 120 days. Consistently with the preliminary simulation, here again the cost decreases when using more reliable disks. Therefore, the best alternative is one that uses disks with MTTF of 20 years and has a detection interval of 120 days. Such an AR will have a MASC of $7,822 and a MTTF of 364 years. Note the critical role that costs played in reaching this decision: if we had ignored costs we could have easily selected a design that achieves the desired MTTF but in a much more expensive way!

5.8 Appraising Cost Decisions

In this final phase, we revisit our assumptions by running sensitivity analysis of the critical and non-critical uncertainties. We again illustrate the process via our case study. Recall that our proposed design was an AR with disk with a MTTF of 20 years and a detection interval of 120 days. Using ArchSim/C we ran sensitivity analyses for all uncertainties, but due to space limitations, we only present the results for two: the Detection Interval (a critical uncertainty) and the cost of detecting failures (a non critical uncertainty).

In Figure 11 we perform a sensitivity analysis for the detection interval. We want to find out the impact of a small change in the suggested 120-day interval. In the figure, we can see that the smaller the detection interval, the higher the MTTF of the AR. In particular, an AR with a detection interval of 240 days will have a MTTF of 184 years. So, we can double the value of the detection interval and we still achieve the target archival guarantees. If doubling the value of the detection interval had caused an important decrease in cost, then we would need to reassess our recommendation.

In Figure 12 we see the AR cost for different detection intervals. As expected, larger detection intervals decrease costs. For instance, increasing the interval to 240 days, causes a reduction of cost of $400 or about 6%. We now have to decide if it is worth considering new alternatives given a potential saving of $400. If this is the case, we should return to the formulation phase and add alternatives with detection intervals in the 120 to 720 days range. Notice that we cannot make a new recommendation based only on the sensitivity analysis, because the variable we want to change may interact in unexpected ways with the reliability and cost metrics. Concretely, in this case, the cost associated with the detection interval might not be a continuous function, so we may need to revisit our cost estimates and re-run the simulations.

Let us turn now our attention to the sensitivity analysis of the cost of detecting failures. This variable has a different nature than the detection interval, as it does not affect the MTTF of the AR. Additionally, we may not be able to change the value of this variable (e.g., the cost of detecting failures may be determined by the market). Therefore, a sensitivity analysis here rather than validating or invalidating our proposal, gives us an idea of how much the cost of the AR may increase (or decrease) if our estimate of the value of this variable was erroneous.

Figure 13 shows the AR cost for different detection costs. As expected, the figure shows higher costs when the detection cost increases. The important observation here is that costs are increasing almost linearly with a very small slope. An increase of 100% in the detection cost (from 200 to 400), only results in an increase of 33% in the AR cost. This means that a small error in the estimate of the detection cost will not affect the AR much. Unfortunately, it also means that efforts in reducing the detection cost will have small payoffs.

6. CONCLUSIONS

In this paper we have studied how to make cost-driven decisions about archival repositories. We presented a framework that improves the efficiency and effectiveness of the AR design process. We described a powerful simulation tool, ArchSim/C, for evaluating the reliability and cost of ARs and the available archival strategies. We described how ArchSim/C can efficiently perform large simulations involving many components and very long simulated periods. We believe our design framework and ArchSim/C can help librarians and computer scientists make rational and economical decisions about preservation, and help achieve better archival repositories.

7. REFERENCES

[1] C. Borgman, S. Chen, H. Garcia-Monlina, K. Thibodeau, , and G. Wiederhold. *NSF Workshop on Data Archival and Information Preservation.* National Science Foundation, March 1999. At http://cecssrv1.cecs.missouri.edu/NSFWorkshop/.

[2] Y. Chen, J. Edler, A. Goldberg, A. Gottlieb, S. Sobti, and P. Yianilos. A prototype implementation of archival intermemory. In *Proceedings of the Fourth ACM International Conference on Digital Libraries*, 1999.

[3] B. Cooper, A. Crespo, and H. Garcia-Molina. Implementing a reliable digital object archive, 1999. Submitted for publication to ACM DL 2000.

[4] A. Crespo and H. Garcia-Molina. Modeling archival repositories for digital libraries. In *Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libaries (ECDL)*, 2000.

[5] A. Crespo and H. Garcia-Molina. Taking cost-driven decisions about archival repositories. Technical report, Stanford University. At http://www-db.stanford.edu/ crespo/papers, 2001.

[6] Gartner Consulting. *TCO Analyst. A White Paperon GartnerGroupś Next Generation Total Cost of Ownership Methodology*, 1997.

[7] A. Goldberg and P. Yianilos. Towards an archival intermemory. In *Advances in Digital Libraries*, 1998.

[8] J. Halpern and C. Lagoze. The Computing Research Repository: Promoting the rapid dissemination and archiving of computer science research. In *Proceedings of the Fourth ACM International Conference on Digital Libraries*, August 1999.

[9] R. Howard. The science of decision-making. In *Readings on the Principles and Applications of Decision Analysis*, volume 1, 1964.

[10] R. Howard. Decision analysis: Practice and promise. In *Management Science*, volume 34, 1988.

[11] H. Levy. *Stochastic Dominance: Investment Decision Making under Uncertainty.* Kluwer Academic Publishers, 1998.

[12] C. S. von Holstein. Assessment and evaluation of subjective probability distributions. Technical report, Economic Research Institute, Stockholm School of Economics, 1970.

Hermes – A Notification Service for Digital Libraries

D. Faensen, L. Faulstich, H. Schweppe, A. Hinze, A. Steidinger
Institute for Computer Science
Freie Universität Berlin
hermes@inf.fu-berlin.de

ABSTRACT

The high publication rate of scholarly material makes searching and browsing an inconvenient way to keep oneself up-to-date. Instead of being the active part in information access, researchers want to be notified whenever a new paper in one's research area is published.

While more and more publishing houses or portal sites offer notification services this approach has several disadvantages. We introduce the Hermes alerting service, a service that integrates a variety of different information providers making their heterogeneity transparent for the users. Hermes offers sophisticated filtering capabilities preventing the user from drowning in a flood of irrelevant information. From the user's point of view it integrates the providers into a single source. Its simple provider interface makes it easy for publishers to join the service and thus reaching the potential readers directly.

This paper presents the architecture of the Hermes service and discusses the issues of heterogeneity of information sources. Furthermore, we discuss the benefits and disadvantages of message-oriented middleware for implementing such a service for digital libraries.

Categories and Subject Descriptors

H.3.6 [**Information Storage and Retrieval**]: Library Automation; H.2 [**Information Systems**]: Database Management

General Terms

Digital libraries

Keywords

alerting services, digital libraries, selective dissemination of information (SDI)

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

1. INTRODUCTION

The traditional way of using electronic publications is searching and browsing. The user must become active to find information. A much more appropriate usage pattern for a digital library would be the notification of readers whenever new material of interest becomes available. A service that provides such a value is called notification or alerting service (AS). The increasing number of scholarly publications emphasizes the need for sophisticated filtering capabilities of such an alerting service to prevent the drowning of users in a flood of irrelevant publications.

Some information providers such as publishing houses do offer alerting on their publications. This is definitely beneficial for end users. However, this bilateral 'user-provider' approach has significant disadvantages. Users have to know all relevant providers. They have to deal with a variety of interfaces, and must maintain their profiles at many sites. For privacy reasons users might not want to disclose their profiles of interest to arbitrary information providers. Duplicate notifications (e.g., occurring in the case when user receives notifications from both a publishing house and an abstracting and citation service) cannot be avoided. The capabilities to express users' interests (profiles) are poor. Typically, the user can just select a set of journals and then receives the table of contents by E-mail whenever a new issue of one of these journals appears. This leads to low precision and recall. Notifications from different providers clutter the mailbox since one rarely has the possibility to specify a notification schedule, and notifications are not integrated. Finally, small information providers (small publishers, universities, etc.) usually do not offer an alerting service.

An integrative alerting system that integrates the variety of providers and publication types and hides their specifics from the users avoids most of these problems. It should offer a unified interface and sophisticated profile expression capabilities including filtering as well as user-defined notification schedules. It must be open and easy to join for other providers which can then feed their bibliographical data into the alerting service and advertise their materials directly to an interested target group. Finally, it must be scalable to support millions of user profiles and tens of thousands of publications daily.

Today, huge amounts of valuable scientific literature are available on the Web. While users can retrieve those data from the publishers' Web sites it is unlikely that each provider is capable and willing to install or support the interfaces of an integrative alerting service. Instead, appropriate wrappers for such Web sites can be employed to make their material available for profile-filtered notification.

Despite its obvious usefulness, only a small number of integrative alerting services have been developed for digital libraries.

Individual alerting services are offered by several publishing houses, such as Springer Link Alert[1] and Elsevier Contents Direct[2]. However, these services underly serious restrictions: they are mostly based on the publications offered by the particular publishing house only. Profiles can only be defined in a basic way, their definition is restricted to the selection of certain journals, no full-text retrieval is possible. Services by secondary publishers as Swets (service SwetsScan[3]) naturally cover a wider but still restricted selection of materials.

Some abstracting and citation services also offer notification services, such as the ISI services[4], Catchword[5], or UnCover Reveal[6]. These services are also restricted to the material offered by the hosting service. Thematically focused portal sites such as Neuroscion[7], BioMedNet[8] underly similar limitations. Most of these are commercial sources. We are convinced, however, that all metadata, including abstracts, will be freely available to the scientific community in the near future.

There is also a number of services specialized in particular types of publications, such as Technical Report Servers (ArXiv[9], REPEC[10], NCSTRL[11]) that offer notification about their documents. Here, the focus of notification is, by definition, restricted.

Some alerting services are offered by libraries (e.g., CISTI Source[12]). Here, the material has a large spectrum covering several publishing houses and different types of publications. Nevertheless, these services are restricted to the material offered by the library.

When subscribing to more than one alerting service, the problem of duplicate notifications due to overlapping coverage arises. In particular, this applies to alerting services of different libraries. The problem of duplicate notifications can be addressed by an intermediate integrative alerting system.

In summary, there are currently a number of services with restricted material, accessible mostly by profiles with poor expressiveness. Even if single providers offer sophisticated profiles to their users they are still restricted to the provider's material.

The scientific research covering Internet-based alerting services that could be used for a digital library environment has a longer tradition. One of the earliest systems developed was SIFT [18], a tool for wide-area information dissemination, that is now commercially operated as InReference. SIFT was a monolithic service that did not support distribution.

Alerting services may be implemented in many ways. Since notification about events is a useful paradigm in different kinds of applications several distributed notification infrastructures have been developed in order to evaluate scalability, examples are JEDI [7], Elvin [16], Siena [4], and NiagaraCQ [5]. The profile definition languages supported by JEDI, Elvin and Siena are too restricted and not appropriate to support sophisticated full-text retrieval. NiagaraCQ supports XML-QL queries that select subtrees in an XML document. In contrast to such tree-based queries, filters in the Hermes application domain typically focus on 'flat' structures like bibliographical references.

The Continual Queries Project [15] investigates update monitoring problems. In this project, the Continual Queries (CQ) language was developed, which has been implemented in several prototypes. This language supports a sophisticated and detailed definition of profiles. CQ is a system that could be used as a basis for an alerting service for digital libraries, as well as other message-oriented middleware.

The academic projects on alerting services introduced above are not all fully implemented. None of these systems is actually used in a digital library context.

Many publishers offer free access to their bibliographic data, usually by sending them via E-mail or by making them available on their Web sites. Unfortunately, the format of this data is not standardized. The problem of extracting bibliographic data from publishers' Web sites can be addressed using methods from the field of semistructured data management [1]. *Wrappers* must be created that allow queries against the data offered by a Web site while hiding layout and structure of this site. Answering such queries means collecting data from multiple linked HTML documents and separating it from the HTML code in which it is embedded. Since HTML is a layout-oriented language, a-priori knowledge about the structure of a Web site and its documents must be employed in the construction of a wrapper. Several approaches such as [2, 3, 11] use regular-expression matching on HTML documents. However, regular expressions are not very robust with respect to layout variations and structural changes that occur frequently in Web sites. Rule-based approaches such as YAT [6] or HyperView [9] that operate on syntax trees of HTML documents provide a higher robustness. The HyperView system that is used in the Hermes project supports the maintenance of wrappers by using a multi-layered approach to separate the concerns of data *extraction* and data *conversion*.

The main contribution of this paper is an architecture for an alerting service which overcomes the heterogeneity problem and – from the user point of view - integrates all kinds of providers into a single source.

The paper is structured as follows: We discuss the issues of implementing an integrative alerting system in the following section. Section 3 presents the architecture of the Hermes system. Section 4 discusses the problems we encountered during development and propose some solutions. The paper concludes with a short outlook.

[1] Springer Link Alert, http://link.springer.de/alert

[2] Elsevier Contents Direct, http://www.elsevier.nl

[3] SwetsScan, offered by Swets http://www.swets.nl/

[4] ISI: Alerting Services, formerly Research Alert Direct, http://www.isinet.com/prodserv/rad/radp.htm

[5] Catchword: Internet Publishing Services, http://www.catchword.com/

[6] UnCover Reveal, http://uncweb.carl.org/reveal/

[7] Neuroscion, http://www.neuroscion.com/

[8] BioMedNet, http://www.bmn.com/

[9] arXiv.org e-Print archive, http://www.arxiv.org/

[10] Research papers in economics, http://www.repec.org/

[11] Networked Computer Science Technical Reference Library, http://www.ncstrl.org/

[12] Canadian Institute for Scientific and Technical Information, http://www.nrc.ca/cisti/source/

Table 1: Provider types

	cooperative	*non-cooperative*
active	sends notification with well-defined metadata	sends human-readable email
passive	allows AS download of well-defined metadata	makes metadata available at its Web site

2. ISSUES

The main problems that have to be solved when building an open integrating alerting system are

1. coping with the heterogeneity of information providers,
2. specifying the users' information needs (user profiles), and
3. efficiently matching the incoming notifications with the user profiles (filtering).

2.1 Heterogeneity

First we define the different types of information providers an integrative alerting system has to handle.

2.1.1 A Classification of Information Providers

Information providers are any suppliers of scholarly information that make metadata (e.g., bibliographical data) available to the alerting service (AS).

We distinguish different types of providers as introduced in [12]. Providers can be either *active* or *passive*. Active providers offer their own AS. For instance, users of the Darwin AS[13] can subscribe to receive an E-mail notification whenever a new issue of a particular journal appears. Passive providers do not offer such a service and have to be queried for new material in a scheduled manner. An algorithm to optimize the query scheduling is introduced in [9].

Additionally, providers can be *cooperative* or *non-cooperative*. Cooperative providers provide their information in one of the standard formats that can be handled automatically by Hermes, and they implement at least one of the protocols supported by Hermes. Non-cooperative providers offer their information in a proprietary format. Specialized wrappers have to be written in that case. Possible combinations of provider types are shown in Table 1.

The distinction between *active* and *passive* is not an exact equivalent to *push vs. pull*. Mixed types occur as well. Consider, for instance, a provider that notifies (push) about new metadata at its FTP server (pull). A similar case is a passive provider together with an active external alerting system such as Darwin that polls the provider on a regular basis and sends notifications containing the URLs of new issues at the provider's Web site.

Passive, non-cooperative providers are most critical for an alerting service. They basically offer an HTML interface.

Active cooperative providers can join the Hermes system by registering and regularly submitting bibliographical metadata in a standardized format.

2.1.2 Heterogeneous Formats

Except for cooperative providers, the alerting system has to deal with various metadata formats. The degree of structure varies from mostly unstructured E-mail notifications in ASCII format over semistructured HTML pages on publisher Web sites to well-structured formats such as XML or the various formats used by citation management software.

2.2 User Profiles

Users express their information needs in *profiles*. A profile is an aggregate of a *query* or *filter*, and a *notification policy*. While the query part defines *which* information the user wants to receive the notification policy specifies *how* it is to be delivered, for instance how often (e.g., daily or weekly), by which protocol (e.g., E-mail), and in which format (e.g., XML, HTML, plain text, BibTeX ...).

A query consists of a *Boolean* filtering expression and a *ranking* part. Both parts are optional. The Boolean query is an SQL-like simple attribute-value matching, e.g., "author=Smith" or "title LIKE '%alerting%'". The set of attributes are the bibliographical metadata fields delivered by the information provider.

Bibliographical data that match the Boolean query can further be ranked according to the ranking part of the query. Such a ranking query is a text-retrieval-like term list (including phrases, proximity operators, or term weights) and a relevance threshold. Documents that are scored by Hermes with a relevance above that threshold are delivered to the user. Alternatively, the user can request the delivery of the top n relevant documents published during the notification period. If the Boolean query part is missing any incoming document is scored according to the ranking query. In this paper we focus on Boolean queries.

Users can create and maintain their profiles using a Web interface. Software clients such as reference managing tools access the service using a dedicated API.

2.3 Filtering Scalability

Incoming bibliographical metadata have to be matched against the user profiles. Obviously, scalability is an issue. While the frequency of incoming notifications in the application domain of digital libraries is usually low, the number of profiles can be very high. Since several 10,000 profiles can be deposited at the alerting service, an efficient matching of profiles and notifications is a crucial task.

3. ARCHITECTURE

Hermes consists of three main components as shown in Figure 1: The Observer, the Filter, and the Notifier.

3.1 Observer

The first task of an integrative alerting system is to collect information from a number of information providers and to produce a combined stream of events that can be filtered against the users' profiles. In the Hermes architecture, the component responsible for this task is called the *observer*.

[13] http://darwin.inf.fu-berlin.de

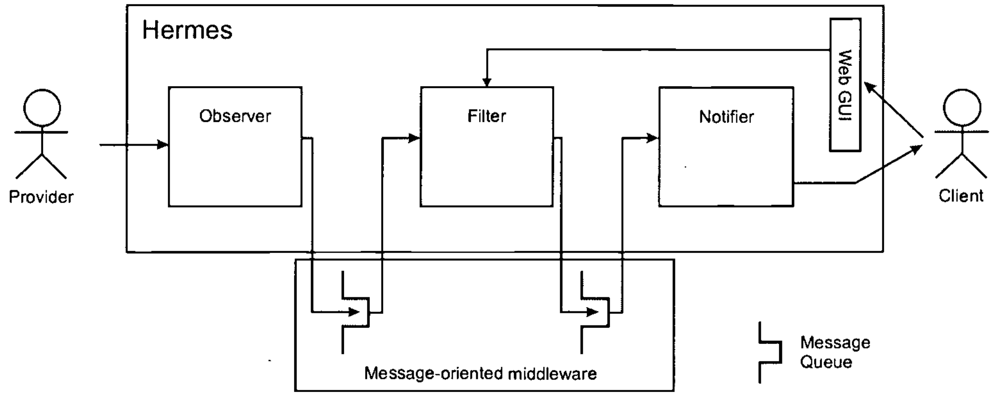


Figure 1: Architecture of the Hermes system

3.1.1 Observer Architecture

The architecture of the observer component (Fig. 2) treats the different types of information providers discussed above in a unified way. Each information provider is handled by a dedicated *wrapper* component. A wrapper consumes a stream of notifications in a provider-specific format and produces a stream of citations in the internal format used by Hermes.

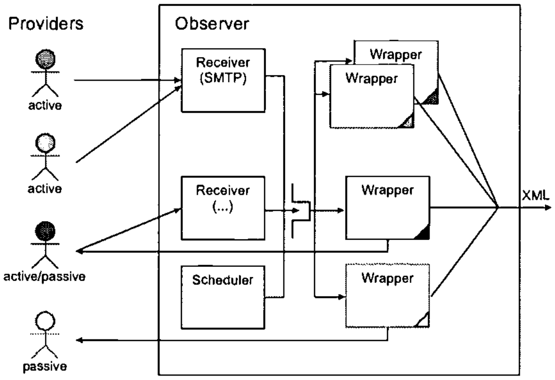


Figure 2: Architecture of the Observer.

In the case of a *cooperative* provider, the wrapper just forwards the incoming notifications since they are already in the internal format. For *non-cooperative* providers the wrapper has to translate the notification into the internal format.

For active providers, the information contained in a notification is sufficient to generate citations. In the case of mixed active/passive types of information providers the wrapper retrieves the material advertised in the notification from an external server of the information provider. For completely passive providers, a synthetic notification is generated by the scheduler component within the observer that causes the wrapper to retrieve new material from the provider.

Since notifications can be sent by the providers using different protocols, the observer contains a component called *receiver* for each protocol. For instance, there exists a receiver in the Hermes system that handles E-mail sent using the SMTP protocol and forwards the content of an E-mail message to the wrappers.

All incoming notifications including those generated by the scheduler form a single stream. Each wrapper subscribes to the notifications originating from the provider it is responsible for (not shown in Fig. 2). This can be easily implemented on top of a message-oriented middleware.

3.1.2 Wrappers

Bibliographical material that can be obtained from non-cooperative providers is typically in a form that resembles a journal's table of contents (TOC). For instance, some publishing houses allow users to browse the TOC of their journals on their Web sites. Others send E-mails containing the journal TOC in plain text. Since the format of the TOC varies between providers, dedicated wrappers are necessary to extract the metadata.

The Hermes system currently uses two types of wrappers: individually coded converters for E-mail notifications of several active providers, and a generic wrapper for the HyperView system [9] that in turn contains several rule-based wrappers for different publisher Web sites.

The first type of wrappers converts E-mails containing tables of contents into citations. Since most publisher's E-mail notifications are highly unstructured, every wrapper has to be implemented individually using a general purpose programming language.

Since HTML pages are more structured than typical E-mail notifications, a higher-level approach based on rules is used for extracting bibliographical data from publisher Web sites: The generic HyperView wrapper consumes XML notifications from the Darwin alerting system that contain references to new journal issues. Each reference contains a URL pointing to the table-of-contents (TOC) page of this journal issue at the publisher's Web site.

The HyperView wrapper responsible for the particular publisher Web site loads the TOC page from the given URL and applies a set of graph-transformation rules to the syntax graph of the page. If necessary, hyperlinks on the page are followed by loading additional pages. As a result, a graph is created that contains the pure data extracted from the analyzed pages. This graph has a source-specific structure.

At the second stage, another set of rules is applied to this graph to transform it into a source-independent representation of the extracted table of contents. Finally, a third set of rules rewrites the bibliographical data of each article into the syntax tree of an XML document that contains the citation in the internal format used by Hermes. This citation is then returned by the HyperView system and posted by the generic HyperView wrapper to the outgoing stream of citations.

Compared to a hand-coded wrapper in a general-purpose programming language, a HyperView wrapper is smaller and easier to understand and maintain since HyperView rules are on a much higher conceptual level.

3.2 Filter

The Filter's responsibility is the comparison of the bibliographical data with the query part of the user profiles (see 2.2). For simple queries using only Boolean filtering Hermes can use a message-oriented middleware (MOM).

For the first filtering step it is not necessary to respect document structure. Therefore, the filterable attributes are extracted from the XML representation using a SAX parser and assigned to filterable message header fields. For further processing the XML representation is included in the message.

3.3 Notifier

The Notifier retrieves the messages buffered for a user according to the user-defined schedule as specified in the notification part of the user profile. The XML payload of the retrieved messages is transformed to the format preferred by the user. Transformation to various formats is performed using XSL stylesheets.

The transformed notification is delivered to the client via the client's preferred protocol. Currently, only E-mail delivery is implemented.

3.4 Message-Oriented Middleware

The communication between the components is performed using a message-oriented middleware (MOM). This has the advantage that components are relatively independent of each other. Another benefit is that profile matching can be performed by the MOM. The MOM is capable of filtering messages according to a selector that is applied against the message header fields. The query parts of the user profiles are stored as *subscriptions* at the MOM. The MOM filters messages and buffers them until the next notification is due.

The MOM is accessed by the Java Message Service (JMS) API [14].

4. PROBLEMS AND SOLUTIONS

In this section we discuss some of the issues encountered during the design and implementation of Hermes in more detail, namely heterogeneity and scalability.

4.1 Wrappers *vs.* Cooperative Providers

Wrappers for information sources are based on a-priori knowledge of the structure and formats of these sources. This knowledge is typically incomplete because in most cases it must be obtained by a reverse-engineering process. Hence, there is no guarantee that a wrapper can handle all data available from an information provider. Moreover, this knowledge can rapidly become outdated since the information provider is autonomous and can choose to change its structure at every moment without notice.

E-mail notifications by publishers are intended for human users rather than for machine input, so they are often unstructured. This requires heuristics like counting the percentage of single letters in a line to distinguish the list of authors from the title of an article. It is a matter of fact that such heuristics are unsharp and can produce incorrect results.

In the case of information extraction from Web sites, one has to deal with semistructured HTML documents that typically provide more hints to find and separate different data items. The HyperView system used in the Hermes system allows to write rules that are robust to some extent with respect to structural variations such as inserts, deletes, or reorderings on a page. This holds as long as the navigation path used to access the data on a page is not affected by the changes. Unfortunately, major changes in the layout will break a wrapper in most cases.

Since wrappers can break, they have to be monitored permanently. Heuristics can be used to detect, e. g., if a wrapper produces no citations for a new journal issue. An alerting system must support recovery, i.e., after fixing a wrapper it must be possible to feed a notification or parts of it to this wrapper again.

Although our wrappers turned out to be quite stable over several months, it is preferable to cooperate with information providers. In fact, the Hermes project has an agreement with a major scientific publisher to deliver bibliographical data as a cooperative provider.

4.2 Bibliographical Formats and Interoperability Protocols

Cooperative providers offer their metadata in a format that can be automatically processed by Hermes. The delivered data set should at least contain all the information against which the queries can be defined, and a link to the document's full-text. For journal papers this data format is the Majour Header DTD [8].[14]

Active cooperative providers submit their data by sending an E-mail or using the HTTP POST method. Data from passive cooperative providers is loaded by FTP or HTTP GET. In the case of passive providers scheduling of Observer activation must be configured. Since in the domain of electronic journals changes on the provider site occur in relatively stable intervals the Observer activation intervals can be precisely tuned. As mentioned above (3.1) the Observer can easily be extended supporting additional protocols (e.g., Z39.50 or CORBA). Currently, no error handling in the data exchange protocol is implemented.

Different protocols have been proposed for delivery of bibliographic data. One of them is defined by the CrossRef project[15], an initiative to provide a centralized source to obtain object identifiers of electronic publications. The idea is to support publishers or authors to link their bibliography entries directly to the reference's full-texts – even if the full-text is located at a different provider's site. CrossRef defines

[14] The Majour Header is an SGML DTD. Hermes transforms SGML documents to conform to XML. Some minor changes are made to the Majour Header, e.g., the addition of a tag <aloc> containing the location (URL) of the article's full-text.

[15] CrossRef, `http://www.crossref.org`

a protocol for submission of bibliographical data that can be easily adopted by Hermes. Bibliographic data are formatted according to the `doibatch.dtd`, an XML DTD for batch submission. Transfer of the formatted data is done via HTTP POST. An HTML-formatted diagnostic message is returned. A detailed failure report is sent by E-mail. A DTD for this diagnostic error message is available.

The CrossRef format allows to submit a minimal set of bibliographical data that is sufficient to identify a document. Unfortunately, this data is not sufficient to provide personalized filtering. For example, including an article title is optional, and the DTD contains no elements for keywords and abstract. However, many publishers deliver data to CrossRef. Therefore, Hermes will support the CrossRef protocol in the near future to provide an efficient way to let publishers join the service. The only modifications that need to be made are the addition of elements to carry the document's abstract and keywords.

Recently, the Open Archive Initiative (OAI)[16] published a protocol for metadata harvesting [17]. It allows *data providers* like technical report servers, publishing houses, libraries etc. to make their data accessible by *service providers* that build value-added services for these data (e.g., alerting services). The OAI protocol is a *pull* protocol that allows service providers to request metadata records from the data providers. It is based on HTTP and XML and is designed to be so simple that an experienced developer can implement it "within a day of work". Metadata (e.g., bibliographical data) are usually delivered in a Dublin Core format (domain-specific schemata can be defined). Requests allow a simple selection by *set* (semantics of sets are not defined in the protocol), and/or by *date*.

It can be expected that the OAI protocol will be implemented by a significant number of data providers. We will therefore apply the protocol as a means to integrate *passive cooperative* providers.

4.3 Filtering

There are two alternatives for matching document metadata against user profiles: (i) implement the matching algorithms yourself, or (ii) make use of existing infrastructures. We compared two implementations of filtering. The first one uses a message-oriented middleware (MOM). The second is based on tables in a relational database system.

A MOM has the capability not only to perform basic message filtering according to a message selector but includes transaction support as well as message buffering. Therefore, Hermes is built on top of such a MOM. In addition to the filtering capabilities the application of a MOM results in a loose coupling of the components and increases the stability, as failure of a single component does not affect the whole system.

However, the application of a MOM has a number of disadvantages:

- Filtering is restricted to Boolean queries (message selectors). Ranking is not supported.
- The number of subscriptions (i.e., queries) is limited.

The latter is attributed to the fact that incoming messages are assigned immediately to the subscribers that are interested in it to avoid a long delay. This approach does not scale well (for instance, in one product the number of subscribers is limited to 1024).

In [10] an implementation of messaging on top of a relational database is proposed. The solution handles the subscription rules as tuples of the relations, and the publication of a message as a trigger that performs a `select` ... `from` ... `where` statement to find all interested subscribers. While this approach is convenient for most applications of messaging systems it restricts the subscription rules to make use of only a small set of comparison operators. However, substring search, which is most important in the field of digital libraries, cannot be supported.

We therefore experimented with a simple implementation of a message queue based on a relational database system that takes advantage of some knowledge of the application domain. The structure of the messages is known in advance, the publication frequency is low, and a message delay up to a few hours is tolerable for users.

A message queue consists of three database tables: a *message* table containing messages (bibliographical data), a *subscriber* table that stores message selectors, and the actual *queue* table where messages are assigned to the subscribers. The message selectors are conditional expression strings that can be applied in the SQL `where` statement to select messages from the *message* table.

Message publication is the insertion of message data into the *message* table. To update a queue two alternatives are conceivable. The most obvious one is the implementation of a trigger for each subscriber that is executed after each insertion of a message. The trigger would match the message with the subscriber's selector and insert a tuple of message id and subscriber id in the *message queue* table (or directly notify the client). Obviously, this approach does not scale. Even with a low frequency of 1 message per second the execution of more than 10,000 triggers is beyond the scope of current database implementations. The alternative of having a single trigger execute a `select` for each subscriber's selector does not solve the problem. Instead, we take advantage of the fact that updating the queue can be deferred for hours. Once in a while (e.g., once a day, which is enough for notification on scientific publications) the *queue* is updated by iterating through all subscribers, selecting the messages that are of interest for the subscriber, and inserting the resulting tuples of message id and subscriber id into the *queue*. Selection and insertion can be performed by a single SQL statement within a short execution time. The time required to update the queue for 10,000 messages and 10,000 subscribers with a message selectivity 0.1 % is in the order of 3 hours (Oracle 8.1.6, Sun Enterprise 450 with 2 processors and 1 GB main memory). A subscriber receives its messages by performing a natural join on the *message* and *queue* tables and selecting the messages that are dedicated to the subscriber (as indicated by the subscriber id in the *queue*).

An API to access the queues that conforms to a subset of the JMS API can easily be implemented.

Existing infrastructures like database systems or message queuing systems allow the easy implementation of simple Boolean filtering. However, for text documents (bibliographical data can be seen as text documents regarding title and abstract) this kind of query leads to low precision and recall. More sophisticated filtering is therefore necessary.

[16] `http://www.openarchives.org/`

4.4 Ranking

As mentioned above simple attribute-value matching as it is performed by the available message-oriented middleware products is far from satisfying the users' needs. Since most sources of bibliographical data provide document abstracts in addition to author, title, etc., one can achieve much higher precision by applying classical information retrieval methods.

In information retrieval document relevance is usually measured by two parameters. The term frequency tf, an indicator of how often the query terms occur in the document, positively influences the relevance scores. The inverse document frequency *idf* indicates the relevance of the search terms with respect to the document collection (in how many documents of the collection does each search term occur?) and influences the score negatively. The problem is that in an alerting service documents are 'transient events' and therefore no document collection exists. To mimic such a collection a 'virtual' collection can be built: All incoming documents are inserted into that virtual collection. The *idf* is computed for that virtual collection. Since access to older documents is not required it is not necessary to keep the actual documents in that collection. Instead it is sufficient to store the term statistics. Topics covered by an alerting service can evolve over time. It may therefore improve the filtering effectiveness to let the term statistics of the virtual collection evolve as well.

5. CONCLUSIONS AND OUTLOOK

Alerting systems offer a necessary means to cope with the ever-increasing amount of scholarly literature. Existing alerting systems in this field are mostly proprietary and often of limited coverage and functionality. Scholars need an open integrative alerting system that (i) combines bibliographical metadata from various sources to maximize coverage and that (ii) offers sophisticated filtering and notification facilities to achieve a high selectivity and usability.

In this paper we have introduced an open integrative alerting system, the Hermes alerting service for digital libraries. An instance of the service is available to the public at `http://hermes.inf.fu-berlin.de`. This instance is currently covering several hundred scientific journals from major publishers, including our project partners Springer and Nature, as well as the technical report servers NCSTRL and ArXiv. Typically, such an alerting system would be operated by a scientific library as a service for its users.

The main issues that are addressed by the Hermes project are the heterogeneity of the information providers, and information filtering according to the user's need.

For the provider integration Hermes follows two alternative approaches. For providers that disclose their bibliographic metadata but not in a well-defined format, Hermes provides appropriate wrappers. Since wrappers are subject to failure due to format changes, Hermes allows providers to join the alerting service with little effort by implementing a simple protocol for data delivery.

Filtering can be implemented with low effort on the basis of message-oriented middleware or relational database systems, but this is restricted to simple Boolean filtering.

In the future we will focus on the following issues: improvement of filtering quality, scalability by distributed execution, duplicate elimination, and relevance feedback.

Improving the filtering quality can be achieved by applying the methods of text retrieval to alerting.

Hermes is designed for scalability. Its loosely coupled components can each be deployed in multiple instances and thus share their work. Related projects [4] have shown that cooperating alerting services can improve scalability. Consider an instance of an alerting service that receives metadata from publishers and propagates it to topic-specific instances. Further research is necessary to find methods to compute the *covering relations* [4] of profiles in a digital libraries.

Duplicate elimination will become necessary once we add information providers with overlapping coverage. Due to different metadata formats, heuristics have to be used to identify different citations of the same publication.

Relevance feedback will allow users to achieve a higher selectivity by grading and returning notifications to the alerting systems. For this purpose we plan to adapt existing approaches in classical information retrieval to relevance feedback.

6. ACKNOWLEDGMENT

The Hermes Project is funded by the German Research Ministry (BMBF) within the Global Info initiative.

7. REFERENCES

[1] S. Abiteboul. Querying semi-structured data. In *ICDT*, volume 6, pages 1–18, 1997.

[2] N. Ashish and C. Knoblock. Wrapper generation for semi-structured internet sources. In *Proc. Workshop on Management of Semistructured Data*, Tucson, 1997.

[3] P. Atzeni and G. Mecca. Cut & paste. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 12–15, Tucson, Arizona, 1997.

[4] A. Carzaniga. *Architectures for an Event Notification Service Scalable to Wide-area Networks.* PhD thesis, Politecnico di Milano, Milano, Italy, Dec. 1998.

[5] J. Chen, D. DeWitt, F. Tian, and Y. Wang. NiagaraCQ: A scalable continuous query system for internet databases. In *Proc. of the ACM SIGMOD Conf. on Management of Data*, 2000.

[6] S. Cluet, C. Delobel, J. Siméon, and K. Smaga. Your mediators need data conversion. In L. M. Haas and A. Tiwary, editors, *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA.* ACM Press, 1998.

[7] G. Cugola, E. Di Nitto, and A. Fuggetta. Exploiting an event-based infrastructure to develop complex distributed systems. In *Proceedings of the 20th International Conference On Software Engineering (ICSE98)*, Kyoto, Japan, Apr. 1998.

[8] European Workgroup on SGML. MAJOUR Header DTD, version 1.1. Software.

[9] L. C. Faulstich and M. Spiliopoulou. Building HyperView wrappers for publisher web-sites. *International Journal on Digital Libraries*, 3(1):3–18, 2000.

[10] J. Freytag, F. Leymann, D. Roller, and M. Stillger. Publish/subscribe functions based on object-relational features. Humboldt University Berlin, Computer Science department, 2000, in preparation.

[11] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo. Extracting semistructured information from the web. In *Proc. Workshop on Management of Semistructured Data*, Tucson, 1997.

[12] A. Hinze and D. Faensen. A unified model of internet scale alerting services. In Hui and Lee [13].

[13] L. C.-K. Hui and D. L. Lee, editors. *Internet Applications. 5th International Computer Science Conference, ICSC'99, Hong Kong, China, December 1999*, volume 1749 of *Lecture Notes in Computer Science*. Springer, 1999.

[14] *Java Message Service API*. `http://www.javasoft.com/products/jms/`.

[15] L. Liu, C. Pu, and W. Tang. Supporting internet applications beyond browsing: Trigger processing and change notification. In Hui and Lee [13].

[16] B. Segall and D. Arnold. Elvin has left the building: A publish/subscribe notification service with quenching. In *Proceedings of the Australian Unix and Open Systems User Group Conference (AUUG97)*, 1997.

[17] H. Van de Sompel and C. Lagoze. The Open Archives Initiative protocol for metadata harvesting. Protocol Specification 2001-01-21, Open Archives Initiative, Jan. 2001. `http://www.openarchives.org/OAI/openarchivesprotocol.htm`.

[18] T. W. Yan and H. García-Molina. SIFT - a tool for wide-area information dissemination. In *USENIX 1995 Technical Conference on UNIX and Advanced Computing Systems, Conference Proceedings*, pages 177–186, New Orleans, Louisiana, Jan. 1995. USENIX Association, Berkeley, CA, USA.

An Algorithm for Automated Rating of Reviewers

Tracy Riggs
Robert Wilensky
Division of Computer Science
UC Berkeley
Berkeley, CA 94720
{tracyr,wilensky}@CS.Berkeley.EDU

ABSTRACT

The current system for scholarly information dissemination may be amenable to significant improvement. In particular, going from the current system of journal publication to one of self-distributed documents offers significant cost and timeliness advantages. A major concern with such alternatives is how to provide the value currently afforded by the peer review system.

Here we propose a mechanism that could plausibly supply such value. In the peer review system, papers are judged meritorious if good reviewers give them good reviews. In its place, we propose a collaborative filtering algorithm which automatically rates reviewers, and incorporates the quality of the reviewer into the metric of merit for the paper. Such a system seems to provide all the benefits of the current peer review system, while at the same time being much more flexible.

We have implemented a number of parameterized variations of this algorithm, and tested them on data available from a quite different application. Our initial experiments suggest that the algorithm is in fact ranking reviewers reasonably.

Keywords

collaborative filtering, recommender systems, electronic publishing

1. MOTIVATION AND BACKGROUND

One of the goals of our project [5] is to explore how technology may be exploited to enable alternative models of dissemination of scholarly information. In the traditional system for scholarly information dissemination, with its roots in paper-based documents, authors submit their papers to journals or conferences, where they are reviewed. The papers deemed worthy are then published in journal volumes or conference proceedings, perhaps with some modifications suggested by reviewers or editors. The volumes are then made available to readers. To a first approximation, members of the same scholarly community write, review, and read these papers, with publishers insuring quality control, distribution and, perhaps, editing services. By most accounts (see for example [17]), journal subscription costs are rising exponentially. Hence, the current scholarly publication system in effect manages to sell content produced by scholarly communities back to those communities at exponentially increasing cost. In addition, it imposes long delays, as papers are made available only at the end of the often lengthy review and distribution process. These and other problems with peer review have been addressed by several researchers, for example those in the medical community [16].

Digitalization per se provides no significant improvement. Digitization can improve the distribution process slightly, but the bulk of the delay is in the reviewing, not in the shipping. Moreover it cannot alter the basic cost structure of the system: Publishers are economically motivated, and hence will license electronic versions of journals in accordance with some charging model that lets them recover at least the same amount. Thus, a significant improvement, we maintain, requires a radical change in the basic structure of scholarly information dissemination.

Fortunately, such a radical change is possible, because the originators of scholarly content are not economically motivated. Writers of research papers receive no direct financial incentive for publishing in journals and conferences.[1] Instead, they generally seek wide circulation and recognition of the merit of their views, discoveries, etc., for other reasons, be they altruistic, or in hopes of obtaining academic rewards, such as tenure, promotion, and the esteem of one's peers.

Abstractly, we can characterize information dissemination generally as "Publication = Distribution + Filtering". Filtering is the function provided by the review system (or, more generally, by whatever mechanism a publisher uses to decide what to publish); distribution is the circulation or availability of the content subsequently. Thus the current system filters first, and distributes later, a sequence that makes sense when distribution costs are relatively high, as they are in paper-based systems. Technology, of course, makes distribution simple and cheap–most authors can post

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

[1] Of course, other, related forms of scholarship, e.g., authoring textbooks or popular publications, generally have an appreciable financial motive. While the work presented here is still applicable in these contexts, going to a self-distributed system doesn't obviously provide cost benefits if the user is charged.

a paper on web server at little or no cost, thus making it available to the world. Indeed, the scarcest resource is generally attention, so it makes economic sense to distribute first, and filter later [19].

The problem, then, is to provide a filtering mechanism that is at least as good as that provided by the current peer review system, but which can operate in the context of self-distributed publications.

There are several parts to this problem. One is that reviewers sometimes make detailed comments on submitted papers, and these comments are often relayed back to authors. Providing finer-grain capabilities for distributed annotation of electronic documents would further this goal, and we have been developing the document technology to do so, which is described elsewhere [13].

While such comments are an important part of the scholarly process, the reviewer's primary task is to indicate some rating of the merit of the submitted work. There is considerable variation of rating systems from one venue to the next, but all have the following components: Reviewers rate submissions along some scale or scales; the highest ranked submissions are published, with editors adjudicating mixed reviews or other controversies.

Thus a system in which the rating of articles is made available to readers would allow the readers to select papers that have the highest aggregate review ratings. Such a system should in theory provide the equivalent of the current peer review system, although, of course, it would be much more flexible. Readers might be interested in seeing papers in which reviewers disagree, or in looking at papers that might be quite good, but beneath the arbitrary cutoff of a journal or conference proceeding. Such functionality is not easily accommodated in the traditional journal system, but could easily be done in the system we propose. In other words, the current peer review system should be approximated by an appropriate collaborative filtering system, which would also be capable of offering additional value.

So far as we are able to discern, the primary value that journals claim to provide is quality control, in the form of the quality of the reviewers that they use. It may indeed be the case that the better journals manage to secure the services of better reviewers (and, perhaps, authors then self-select their publications, so that better journals receive better submissions as well). The bottom line, then, is judging the reviewers. That is, readers want not the papers with the best reviews, but the papers deemed best by the best reviewers. Thus, we need a collaborative filtering system that will automatically provide simultaneous quality filtering of both papers and reviewers.

There has been considerable work on collaborative filtering and recommender systems [1, 4]. The vast majority of this work relies on the principle of finding users with similar affinities. For example, Tapestry [9] provided a filtering system for e-mail messages; the GroupLens [10] project initially targeted Usenet and movies, more recently extending its scope to include general information filtering algorithms; PHOAKS [18] supports recommending and annotating Usenet messages; and Siteseer [14] and Fab [7] perform filtering on World Wide Web pages. Affinity-based recommender systems are also gaining popularity in E-commerce [15].

In contrast, we are attempting to perform collaborative *quality* filtering, based on the principle of finding the most reliable users. We would categorize as collaborative quality filtering work such as [12], which supports automatically assessing reputations in the context of E-commerce transactions. However, the problem of automatically rating reviewers seems unaddressed. Here we provide an algorithm to do so. The algorithm is applicable to any collaborative filtering scenario in which reviewers rate items along some scale. Indeed, our initial tests of the algorithm are in quite a different domain, because of the availability of reviews and reviewer ratings.

2. ALGORITHM

The general idea of the algorithm is that good reviewers are those whose reviews predict the ultimate consensus review of an item. We assume that the average rating of an item is the closest measure we can obtain to the true "value" of that item. Thus, any reviewer who consistently ranks items near their ultimate average can be considered to be a reliable reviewer.

2.1 Basic Algorithm

The basic algorithm is quite simple. First, we assume that each reviewer's rating, and each item's rating, can be translated into a score between 0 and 1. The following is a general outline of the algorithm:

```
while (not converged)
   compute item rating as weighted average
   compute reviewer score based on how close
        to average reviewer rates items
```

This iterative algorithm is similar to the web-searching algorithm proposed by Kleinberg [11]. In Kleinberg's algorithm, web pages are labeled as *hubs* and/or *authorities*. The hubs are pages which point to many authorities, and the authorities are pages that are pointed to by many hubs. An iterative algorithm is used to compute hubs and authorities, in which the "hub score" and "authority score" of a set of pages are alternately computed until convergence is reached. (A related idea is to apply Kleinberg's algorithm to research papers by using the citation graph; this feature is offered by Citeseer [3].) Note that in Kleinberg's algorithm, if the hub score of page X increases, then that increases the contribution of X to the authority scores of all the pages to which X points. The algorithm proposed here is similar–if the score of reviewer Y goes up, then that increases the contribution of Y to the ratings of items that Y has reviewed.

Kleinberg's algorithm does not apply directly here, as Kleinberg deals with a symmetric matrix of items versus items, whereas we have a set of reviewers versus a set of items. Furthermore, a reviewer should not benefit from giving an item a high score, but rather should benefit from giving an item a score that is close to the item's weighted average. This leads to a nonlinear iterative algorithm, as opposed to Kleinberg's linear algorithm. Kleinberg offers a proof that his algorithm will always converge. We do not yet have such a proof for our algorithm (and doubt that a straightforward proof exists), but in practice, we have found that it converges rapidly.

A related idea has been proposed separately by Canny [8]. Canny's algorithm is based on a consensus model; a score is assigned to each reviewer based on how closely that reviewer's ratings vector correlates with the vectors of other reviewers. Our algorithm is based on a similar idea, but also incorporates the item averages in the iteration.

The formula for computing the item rating is simple; it is just a weighted average:

$$a_j = \frac{\sum_{i \in R_j} w_i r_{ij}}{\sum_{i \in R_j} w_i}$$

where a_j is the rating for item j, R_j is the set of reviewers that have reviewed j, w_i is the score of reviewer i, and r_{ij} is the rating that reviewer i gave to item j.

Computing the reviewer scores is slightly more involved. We compute the average difference between r_{ij} and a_j, which is simply the Manhattan distance divided by the number of items reviewed.

$$w_i = 1 - \frac{\sum_{j \in S_i} |a_j - r_{ij}|}{n_i}$$

Here S_i is the set of items that user i has reviewed, and n_i is the cardinality of S_i.

2.2 Additional Factors

There are a number of additional factors that one may or may not want to incorporate in the algorithm. We define three of these factors as α, β, and γ, which are incorporated into the formula for calculating reviewer score as follows:

$$w_i = \alpha_i \left[1 - \frac{\sum_{j \in S_i} \gamma_{ij} \beta_j |a_j - r_{ij}|}{\sum_{j \in S_i} \gamma_{ij} \beta_j} \right]$$

We now discuss each of these factors in turn.

2.2.1 Number of items reviewed

If a reviewer has rated one item close to the average, it would seem unwise to conclude that he or she deserves to be ranked among the top reviewers. Instead, we might want to discount inexperience (or lack of data). The factor α is to compensate for such a lack of data, and is defined as:

$$\alpha_i = 1 - \frac{1}{n_i}$$

2.2.2 Number of reviews of an item

Another consideration is the number of reviews available for an item. If a reviewer rates an item that has very few reviews, then, without any adjustment, that review will greatly influence the overall rating of the item, and, consequently, suggest that the reviewer is highly reliable. In contrast, a review of an item that has been reviewed by many reviewers will not influence the score of that item much, and hence, have a much smaller effect on the subsequent assessment of the reviewer, despite the fact that the reviewer provided the same value in both cases. Thus, the factor β is used to give more weight to those items that have received more reviews:

$$\beta_j = 1 - \frac{1}{m_j}$$

Here m_j is the number of reviews that item j has received.

2.2.3 Time of review

Consider the time at which a reviewer rates an item with respect to other reviewers. If a reviewer is an early reviewer, and is close to the subsequent average, then that reviewer has in fact predicted the average. In contrast, if a reviewer has available to him or her the benefit of many previous reviews, that reviewer could influenced by those reviews, a concept known as "herding" [6]. It is reasonable, then, to give more credit for reviews of an item for which fewer reviews are available than reviews for which more reviews are available. Doing so is the point of the factor γ, defined as:

$$\gamma_{ij} = 1 - \frac{t_{ij}}{m'_j}$$

where m'_j is the number of available reviews and t_{ij} is the *rank* of reviewer i with respect to item j (the number of reviews that were available when reviewer i rated item j, plus one). Here we assume that a reviewer for which no reviews are available has a rank of 1; the last reviewer of an item has a rank of m'_j (when all previous reviews are available, $m'_j = m_j$).

2.3 Exploiting Undue Influence

A reviewer could attempt to unduly influence the system as follows: He rates many items in which he has no great interest at their known average, to eventually obtain a high reviewer rating; then he rates a few items in which he has a great interest as he desires, in an attempt to have a greater influence on their average. Such spoofing could be used to advance "cliques", or groups of people that would like to promote each other's work.

The parameter γ, which uses the rank order of the reviewer's rating, could alleviate this problem somewhat. However, the parameters α and β could exacerbate it.

Our belief is that this vulnerability is in fact an intrinsic problem of peer review, rather than a problem with the algorithm per se. Indeed, traditional peer review processes try to filter potential reviewers for conflicts of interest in a variety of ways: asking reviewers to name their students and advisors, or presenting papers to be reviewed without authorship in evidence. Of course, each of these measures can be implemented in our collaborative filtering scenario. However, both in the traditional case and in our proposal, such measures will be at best superficial. Indeed, it is hard to discern the difference, in terms of the patterns of reviews, between cliques of malicious spoofers and affinity groups of scholars with deeply held differences of opinion.

We believe our algorithm is not exceptionally vulnerable to this intrinsic problem, and may indeed provide some help. For example, with a data base of reviews available, it may be possible to automatically detect spoofers, or affinity groups of scholars, and adjust the weighting of a review in accordance with such an affinity. We leave this problem for future work.

2.4 A Variation on the Algorithm: Assessing Reviewer Expertise

An additional aspect we may incorporate into the algorithm is that a reviewer may have multiple areas of interest, but may not necessarily have the same level of knowledge in all areas. Thus, the reviewer may be more skilled at judging papers in one research area than another. We have proposed an enhancement to the algorithm that accounts for this detail.

There are various ways to approach the addition of this feature. One possibility is to categorize all of the documents and to give the reviewer a score for each category. However, classifying documents in this manner is limiting, as papers generally overlap several categories. We chose a method based on using pairwise similarity among documents. Two documents can be compared to one another, for example by computing the cosine of the angle between the word vectors of the documents, thus resulting in a similarity measure between them.

In the enhanced algorithm, the rating for each reviewer is a vector rather than a scalar, so a reviewer has a different score for each item - a measure of his or her "expertise" on rating that item.

$$a_j = \frac{\sum_{i \in R_j} w_{ij} r_{ij}}{\sum_{i \in R_j} w_{ij}}$$

Here w_{ij} is the reviewer i's expertise rating for item j. In this variation of the algorithm, we compute a weight for each reviewer-item pair. The idea is based on the following principle: if a reviewer has rated many items similar to item j and has given those items accurate ratings, then he or she has a high level of "expertise" on item j. Let s_{jk} be the similarity of items j and k. Then we compute w_{ij} for reviewer i and item j as follows:

$$w_{ij} = \frac{\sum_{k \in S_i} s_{jk}(1 - |a_k - r_{ik}|)}{\sum_{k \in S_i} s_{jk}}$$

The additional factors discussed in the paper may also be incorporated in the enhanced algorithm. The following equation incorporates these factors:

$$w_i = \frac{\alpha_i \sum_{k \in S_i} \gamma_{ik} \beta_k s_{jk}(1 - |a_k - r_{ik}|)}{\sum_{k \in S_i} \gamma_{ik} \beta_k s_{jk}}$$

3. AN EXPERIMENT

The basic algorithm and its parameterized variations were tested on data gathered from Epinions.com, a web site designed for consumers to share product reviews with other consumers. The Epinions.com data was chosen for several reasons: (1) the data are usable for testing the algorithm because members give items numerical ratings, (2) it is a popular website and therefore contains a large amount of data, and (3) the Epinions.com assessment of member "reliability" may be used as a metric by which to measure the performance of the algorithm.

We have not yet tested the variation of the algorithm that includes assessment of reviewer expertise, but intend to conduct a similar experiment using Epinions.com data. The items on Epinions are arranged in a taxonomy, allowing us to use item proximity in the graph as a similarity measure.

3.1 About Epinions.com

Members of Epinions.com submit reviews for any item in a finite set of items maintained by Epinions.com. The member rates the item using a score of 1 to 5 (5 being the best) and also offers a written review. Other members may then rate the review in terms of whether or not they would recommend the review to others. Furthermore, a member may read several reviews by another member and then decide to either "trust" or "distrust" that member. The result is that some members end up being "highly trusted" or have "highly recommended" opinions, while others are "not trusted" or have "not recommended" opinions.

3.2 Metrics

The following metrics were used for assessing the algorithm's performance:

- The number of members that trust a reviewer. The more a reviewer is trusted, the more reliable we can expect her reviews to be. However, a reviewer who has written more reviews can be expected to have more trusters, simply by virtue of being more visible in the community. Therefore, the number of trusters is normalized by dividing by the number of reviews written.

- The average "recommendation level" of a reviewer's reviews. We assign a score to each possible rating of a reviewer's review: "highly recommended" = 3, "recommended" = 2, "somewhat recommended" = 1, and "not recommended" = 0. If we take all of a reviewer's reviews and average the numerical value assigned to them, that should be a reasonable measure of the reliability of that reviewer.

4. RESULTS

We have run the algorithm on a set of 100,000 reviewers from the Epinions.com community. Figure 1 shows the results of the algorithm measured against the number of "trusters", and Figure 2 shows the results measured against the average recommendation level.

The graphs may be interpreted as follows: each point on the horizontal axis represents the group of reviewers who fell into a given score range. For instance, if a reviewer's score was 0.64, the reviewer is included in the group 0.6-0.8. Within each group, the average number of trusters per review (Figure 1) and the average recommendation level (Figure 2) were computed. The five bars within each group correspond to five different variations of the algorithm. For the bars labeled "none", $\alpha = 1$, $\beta = 1$, and $\gamma = 1$. For the bars labeled "α", $\gamma = 1$, $\beta = 1$, and α is computed as described above, and so forth. For each variation of the algorithm, a single-factor ANOVA test showed that the five groups were significantly different at a 99% confidence interval ($p < 0.01$).

In general, the graph shows that the ratings given by our algorithm tend to increase as the ratings given by the Epinions.com metrics increase. The most dramatic results are seen when the parameter α is used. This is unsurprising, because we would expect to see high reliability among the active members of Epinions.com. Interestingly, the algorithm appears to correlate with Epinions.com data even when no additional factors are used. The factors β and γ have a less dramatic effect.

5. DISCUSSION

We believe that these initial results suggest that (some variations of) our proposed algorithm provides a plausible way to automatically assess the reliability of reviewers, and

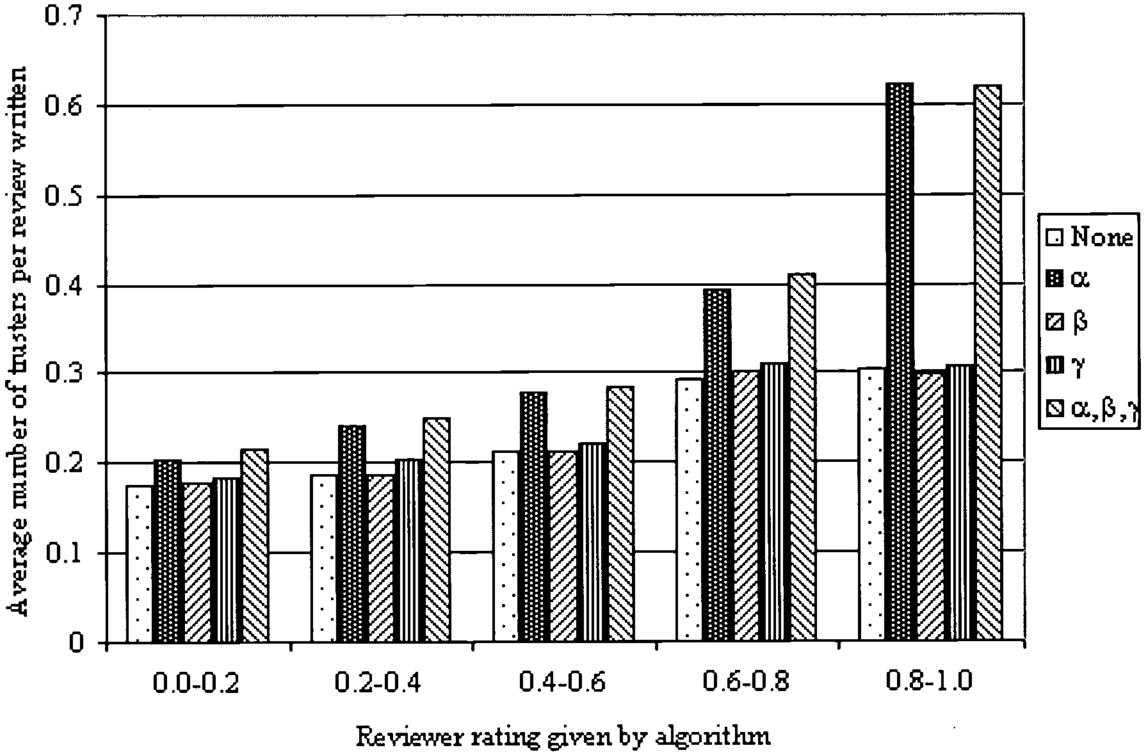


Figure 1: Epinions Trusters vs. Algorithm Rating

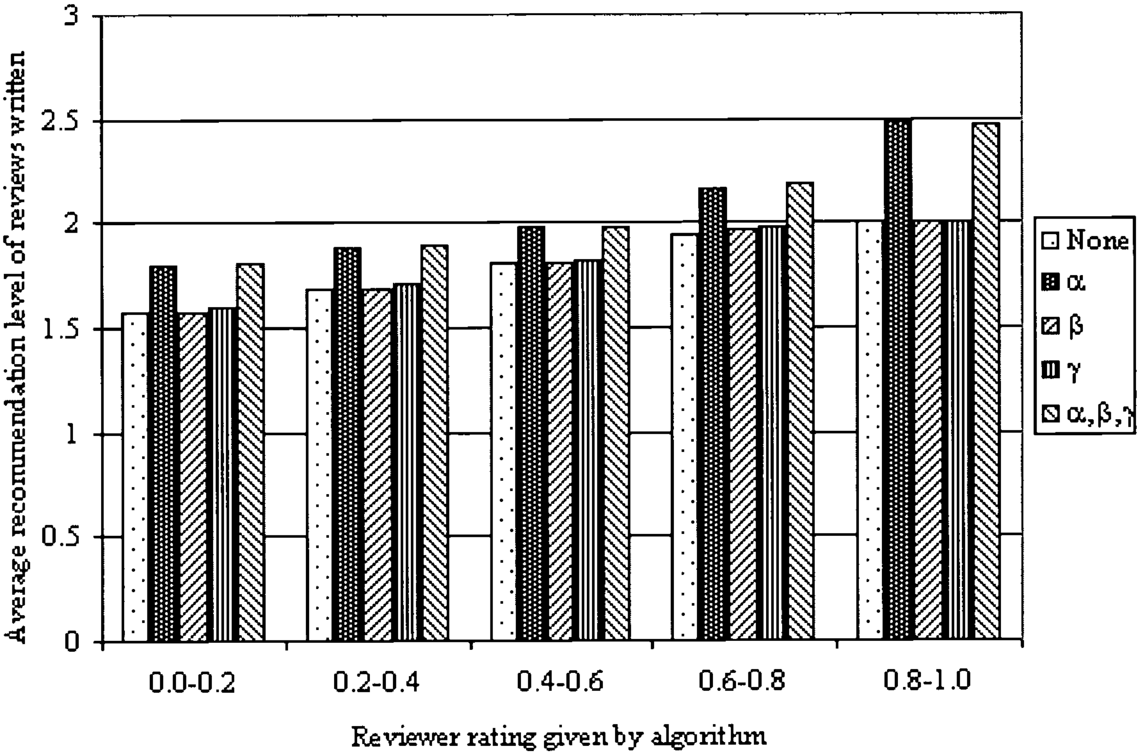


Figure 2: Epinions Recommendation Level vs. Algorithm Rating

hence, may serve the purpose of its design, namely, to supply the value of peer review in a self-distributed system of scholarly information dissemination. We must be tentative about our conclusions, of course, since we intend the algorithm to be used for scholarly information dissemination, and it is difficult to judge its efficacy using the consumer-oriented Epinions.com data. One reason that the Epinions.com metrics are not entirely ideal is that members are generally rated on the *quality* of their written review, rather than on the *accuracy* of their numeric rating. We expect there to be some correlation between the two, but have no way to verify this conjecture. However, we suggest that the Epinions.com data provide a reasonable starting point. A more definitive test would involve deploying the algorithm in a context for which it was designed, which we plan to do.

6. FUTURE WORK

One feature of the current algorithm is that it conflates confidence with quality. Specifically, one of our parameters discounts the rating of a reviewer based on the number of reviews he or she has done; another discounts the rating based on the number of reviews contributing to the item rating. However, the reviewer's quality may be excellent to begin with; it is only our confidence in his or her work that is increasing. Thus, separating the assessed quality of a reviewer from the system's confidence in that quality may be desirable, although we are uncertain that doing so will affect the algorithm's bottom line.

We mentioned above that current review systems often ask reviewers to rate papers along more than one dimension. The algorithm described here could easily be applied to multi-dimensional reviewing strategies, simply by applying it independently to each rating dimension. Indeed, it would be particularly useful, at least in some fields, to rate papers along a "correctness" dimension and an "importance" dimension, as an interesting theory may ultimately turn out to be false, but still be important, and indeed, highly referenced by discrediting work, and skeptical reviewers would have a means to express a "positive disagreement", i.e., lower correctness but high importance. Of course, rating papers along multiple dimensions also opens the possibility of rating reviewers along these same dimensions.

Separating out an importance and a correctness dimension allows for another, substantial addition to the algorithm: This is to regress on author citations. That is, the number of citations to a work is some measure of the importance of the work. Thus, reviewers whose previously "highly important"-rated articles ended up with large numbers of citations should also be considered good reviewers (insofar as importance is concerned).

Along this line, there are many other parameters one may suspect are correlated with the reliability of a review, such as the length of commentary, the institution with which a reviewer is affiliated, and so forth. With such a rating system in place, we might be able to find out if our intuitions about such items have empirical merit.

There are a number of variations of the algorithm that may be worth exploring:

- Use a different measure of distance from the average (a change to the "basic algorithm"). Superlinear distance measures will have the effect of penalizing one big "error" in a review more than the same about of error distributed over many reviews. We believe doing so is undesirable. But perhaps some other measure would prove valuable.
- Use different values for the parameters α, β, and γ. For instance, α is a such a major factor in the reviewer rating, we may wish to reduce its influence.
- There are other factors that could be used and have not been mentioned in this paper. For example, one might try to solicit a "degree of confidence" from the reviewer, i.e., a self-rating of the reviewer's own confidence in his or her review. This would be helpful if the reviewer wanted to spend only a short amount of time on the paper, or questioned his own expertise, etc. (This could not be tested using an Epinions.com metric.)

As mentioned above, it may be possible to automatically detect spoofers, or affinity groups of scholars. For example, reviews by reviewers that give each other mutually excessively positive reviews could be discounted. Alternatively, one could use affinity groups for paper recommendations. In this use, the proposed system would act more like other collaborative filtering systems, in which users simply use reviewers that they like to filter for them.

We also proposed a variation on the algorithm that includes an enhancement for assessing the reviewer's expertise in a given research area. This variation has not yet been tested, but we are currently in the process of experimenting with the algorithm using the Epinions.com data.

Of course, the algorithm should be tested in a real system where it can be judged by actual users. Performing such a test is an important future step. There are many practical and sociological issues that need to be addressed to deploy such an algorithm in a realistic context. One is to motivate individuals to review papers, and to review them accurately. While we do not believe the sociology of reviewing is well-understood, we believe that practices found effective in both traditional reviewing and other collaborative filtering work can be applied here. For example, [6] suggests that keeping early reviews unavailable is effective in both soliciting subsequent reviews and preventing "herding".

We suggest that a collaborative filtering scheme such as we propose may not only provide the same value as supplied by peer review, but may ultimately provide additional value. Journal editors believe they know who the good reviewers are, but such knowledge is apparently largely anecdotal–perhaps algorithms such as this one will provide a more objective assessment. Similarly, the value of reviews can be tracked, and one's prowess as a reviewer measured, so that the rewards currently associated with such activity may be better calibrated. Finally, entirely new motivational schemes are possible. For example, The Berkeley Electronic Press [2] has established an "authors and reviewers' bank", in which authors must review other authors' papers in order to receive reviews for their own.

The inclination to use a system such as we propose is likely to vary from discipline to discipline. In some fields, authors are very careful not to leak results pending very careful reviews; obviously, such scholarly communities would be less interested in self-distribution and quality filtering. It is an interesting challenge to see if mechanisms such as the ones we propose can be applied to disciplines with such different sociologies.

7. ACKNOWLEDGEMENTS

Paul Resnick pointed out to us the interesting difficulties of reviewer motivation. Hal Varian made valuable comments on a previous version of this paper.

This research was supported by the Digital Libraries Initiative, under grant NSF CA98-17353.

8. REFERENCES

[1] ACM-SIGIR 1999 Workshop on Recommender Systems: Algorithms and Evaluation. http://www.csee.umbc.edu/ ian/sigir99-rec/summary.html.

[2] The Berkeley Electronic Press. http://www.bepress.com/.

[3] CiteSeer. http://citeseer.nj.nec.com.

[4] Collaborative filtering. http://www.sims.berkeley.edu/resources/collab/.

[5] The UC Berkeley Digital Library Project. http://elib.cs.berkeley.edu.

[6] C. Avery, P. Resnick, and R. Zeckhauser. The Market for Evaluations. *American Economic Review*, 89(3):564–584, 1999.

[7] M. Balabanovic and Y. Shoham. Fab: Content-Based Collaborative Recommendation. *Communications of the ACM*, 40(3):66–72, March 1999.

[8] J. Canny. Personal communication with the authors.

[9] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM*, 35(12):61–70, December 1992.

[10] J. L. Herlocker, J. A. Konstant, A. Brochers, and J. Riedl. An Algorithmic Framework for Performing Collaborative Filtering. In *Proceedings of the 1999 Conference on Research and Development in Information Retrieval.* ACM-SIGIR, August 1999.

[11] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[12] G. Z. A. Moukas and P. Maes. Collaborative Reputation Mechanisms in Electronic Marketplaces. In *Proceedings of the 32nd Hawaii International Conference on System Sciences*, 1999.

[13] T. A. Phelps and R. Wilensky. Multivalent Documents: Anywhere, Anytime, Any Type, Every Way User-Improvable Digital Documents. *Communications of the ACM*, 43(6), June 2000.

[14] J. Rucker and M. J. Polanco. Siteseer: Personalized Navigation for the Web. *Communications of the ACM*, 40(3):73–75, March 1997.

[15] J. B. Schafer, J. Konstan, and J. Riedl. Recommender Systems in E-Commerce. In *Proceedings of the ACM Conference on Electronic Commerce*, November 1999.

[16] R. Smith. Opening up BMJ peer review. *BMJ*, 318:23–27, 1999.

[17] C. Tenopir and D. W. King. Trends in scientific scholarly journal publishing in the United States. *Journal of Scholarly Publishing*, 28:135–170, 1997.

[18] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter. PHOAKS: A System for Sharing Recommendations. *Communications of the ACM*, 40(3):59–62, March 1997.

[19] H. R. Varian. The Future of Electronic Journals. *Technology and Scholarly Communication*, 1999.

HeinOnline: An Online Archive of Law Journals

Richard J. Marisa
Cornell Information Technologies
Cornell University
Ithaca, NY 14853
+1-607-255-7636
rjm2@cornell.edu

ABSTRACT

HeinOnline is a new online archive of law journals. Development of HeinOnline began in late 1997 through the cooperation of Cornell Information Technologies, William S. Hein & Co., Inc. of Buffalo, NY, and the Cornell Law Library.

Built upon the familar Dienst and new Open Archive Initiative protocols, HeinOnline extends the reliable and well-established management practices of open access archives like NCSTRL and CoRR to a subscription-based collection. The decisions made in creating HeinOnline, Dienst architectural extensions, and issues which have arisen during operation of HeinOnline are described.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Collection, System issues.*

General Terms

Management, Design, Experimentation.

Keywords

Dienst, digital library, document structure, law journals, metadata, system design.

1. INTRODUCTION

Law is grounded in the past, in the decisions and reasoning of generations of lawyers, judges, juries, professors. Ready access to this history is vital to solid legal research, and yet, until 2000, much of it was buried in vast collections of aging paper journals.

Enter HeinOnline, an ambitious project to make complete runs of US law school journals available on the Web, and then to expand this to include an array of other classic legal materials.

HeinOnline is a collaboration among William S. Hein and Co. Inc., the world's largest distributor of legal periodicals, Cornell Law Library and Cornell Information Technologies. The collection currently contains more than 55 journals, comprising over 1.4 million pages (as of 4/1/01) and is growing at over 150,000 pages a month.

Before HeinOnline, searching these materials was often difficult. Researchers depended on the journals being on the shelf when they needed them, and the only tool for finding articles was paper-based indexes — often constructed with outdated legal terminology and lacking topics of contemporary interest.

Browsing the collections and searching for specific information are both supported, which means researchers don't have to trade the convenience of online access for the ability to "flip" through the pages of journals. Another feature is being able to enter a standard citation and instantly pull up the article.

Cornell's involvement began in 1997 when Hein was investigating ways to put its collections on the Web and Cornell University Library's *Making of America*[1] project caught its attention. *Making of America* is a collection of 19th-century periodical literature which serve as primary sources for American social history. Like HeinOnline, it relied on a digital library protocol called Dienst [1]. Cornell's Office of Information Technology had provided technical support for the *Making of America* project, and became technical lead for HeinOnline.

2. DIENST

Dienst is a framework for implementing digital library systems. The Dienst architecture specifies a set of distributed services which allows access to documents, components of documents and aggregations of documents. Dienst was attractive because it offered a clearly defined set of document *services* which together comprise a capable repository, an open, proven *protocol* [5] to communicate with those services, upon which to build a user interface and application layer, and it encompassed a *document model* which was applicable to law journals.

Dienst was developed as part of the Arpa-funded CS-TR project, which was undertaken to make computer science research available over the Internet and to undertake basic research in digital libraries [2]. It formed the basis of the Networked Computer Science Technical Report Library [4], and is used by

- CoRR, the Computing Research Repository[2],
- the Open Archives Initiative[3],
- ETRDL, the ERCIM Technical Reference Digital Library[4], and

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001ACM 1-58113-345-6/01/0006...$5.00.

[1] http://cdl.library.cornell.edu/MOA

[2] http://xxx.lanl.gov/archive/cs/intro.html

[3] http://www.openarchives.org

[4] http://www-ncstrl.inria.fr/Dienst/htdocs/

- the Cornell University Library Historical Math Book Collection[5].

2.1 Services

Services defined within Dienst include:

- a *Repository Service* which stores digital documents according to the defined document model, each of which has a unique name and may exist in multiple versions, each with different components and formats,
- an *Index Service* which accepts queries and returns lists of documents identifiers matching those queries,
- a *Collection Service* which provides information on how a set of services interact to form a logical collection., and
- a *User Interface Service*, through which human interaction with the other services and their protocols is mediated.

2.2 Protocol

Dienst protocol requests are expressed as URLs embedded in HTTP requests A typical implementation uses a standard Web server, such as Apache, that is configured to dispatch Dienst URLs to the appropriate Dienst service.

Responses to protocol are formatted as HTTP responses. The content type of the response will vary according to the type of reply. For example:

- `text/xml` is used for responses that contain structured information (such as the protocol request for the internal structure of a digital object), and
- content specific types such as `image/gif` and `application/postscript` are employed for disseminations from digital objects.

2.3 Document Names

Documents in Dienst Repositories are named by assigning them a *handle* [3]. The Handle System is a comprehensive system for assigning, managing, and resolving persistent identifiers for digital objects and other resources on the Internet. Every handle consists of two parts: its naming authority (the "prefix"), and a unique local name under the naming authority (the "suffix"). The naming authority and local name are separated by the ASCII character "/". Naming authorities are defined in a hierarchical fashion resembling a tree structure.

2.4 Document Model

The Dienst document model allows for the storage and dissemination of documents in multiple forms. While documents in some other collections are "born digital" and have primary representations in Postscript or other electronic formats, HeinOnline journal volume documents are stored as sets of high resolution scanned images and as text files derived from those images. Derivative representations of the page images are disseminated at different resolutions and in different image formats for viewing and for printing and as text for search procedures.

The Dienst service responsible for document storage and dissemination is the *Repository*. The Repository service processes commands ("verbs" in Dienst parlance) to deposit documents, discover their structure, and to provide disseminations of documents.

The *Structure* verb allows the user to discover the logical structure of a document, the *Formats* verb returns the list of formats ("content-types") which may be requested from a document, and the *Disseminate* verb allows a client to request a dissemination of the document by specifying a structural component and content-type.

The Dienst Repository service supports other concepts, such as "versions" of a document, which are not employed in the HeinOnline system.

2.5 Document Structure

Dienst maintains structural and descriptive metadata – information about the tables of contents, indexes, chapters, and so on – along with descriptive (cataloging) information about each structural component. The cataloging information is used to facilitate discovery and location of the subcomponents, for example, finding articles by author and/or title.

The only requirement of the descriptive metadata used by Dienst is that it is transported (formatted) in XML. The Dienst Protocol Specification offers examples of descriptive metadata using rfc1807, Dublin Core [7], and OAMS (Open Archive Metadata Set) elements. HeinOnline formats descriptive metadata using Dublin Core elements.

As represented in the Dienst Structure file, a document may contain zero or more views. Each view is an alternative expression or structural representation of the content encapsulated in the digital object. Two examples from the Dienst Protocol Specification illustrate the notion of alternate views:

A digital object representing a musical work may contain three views:

- the audio view of the music
- a textual view of the lyrics
- a video view of a performance of the musical piece.

A digital object representing a scholarly paper may contain two views:

- the complete content (body view) of the paper, including text and tables
- a table view that only provides access to the tables in the paper.

A Dienst view may then be hierarchically structured using nested divisions ("*divs*"). Two examples illustrate the purpose of a div hierarchy:

- A book view may contain a hierarchy indicating sections and chapters
- A scholarly journal view may contain a hierarchy containing issues and articles.

Each div may contain descriptive metadata and may then contain one or more *terminal elements* that are individually disseminable components of the document instance. At present Dienst only supports a single terminal element, *pageimage*, which represents

[5] http://cdl.library.cornell.edu/math.html

an individual page of text. Future versions of the protocol may support other terminal elements such as frames in a movie or samples in a digital audio format.

3. HEINONLINE DESIGN

3.1 Dienst Repository Server

A Dienst Repository server was written in Perl for HeinOnline according to the Dienst Protocol specification. This new implementation, distinct from the Cornell Computer Science/ NCSTRL implementation of Dienst, is a CGI program which works with the Apache web server. The Dienst server is used under MS Windows operating systems (NT, 2K, 98) for internal users; the production server for external users runs under Solaris 2.6 on an UltraSparc 10.

3.2 Document Names

The documents in the HeinOnline law journals collection are assigned handles using the naming authority `hein.journals`. The local name part of the handle specifies a specific document in the collection, i.e., a journal volume. Thus, the handle for Texas Law Review, volume 50, may be `hein.journals/tlr50`.

3.3 Document Structure

The law journal volume view used in HeinOnline consists of a sequence of page images organized into a hierarchy of issues, articles, indexes, cases, and so on. The volume and each level of the hierarchy (the divs) may have an associated metadata record, which is marked up using Dublin Core elements. Listing 1 shows the beginning of the structure file for Cornell Law Review volume 85 (1999), including the Dublin Core record for the volume as a whole. Listing 2 shows a fragment from the middle of the file, including an article and its Dublin Core record and some of its pageimage elements.

Listing 1. Beginning of Structure File

```
<?xml version="1.0" encoding="UTF-8"?>
 <Structure>
  <meta-formats>
   <dc xmlns:dc=
     "http://purl.org/dc/elements/1.0/">
     <dc:Title>
        Cornell Law Review
     </dc:Title>
     <dc:Series>85</dc:Series>
     <dc:Date>1999-2000</dc:Date>
     <dc:Identifier>
        citstring:Cornell L. Rev.
     </dc:Identifier>
   </dc>
  </meta-formats>
  <view type="volume">
    <div id="misc1" desc="titlepage">
      <meta-formats>
         <dc xmlns:dc=
     "http://purl.org/dc/elements/1.0/">
           <dc:Description>
             Title Page
           </dc:Description>
           <dc:Identifier>
         citation:85 Cornell L. Rev. []
           </dc:Identifier>
         </dc>
      </meta-formats>
    <pageimage id="1" native="[]"/>
    <pageimage id="2" native="[]"/>
    </div>
...
</Structure>
```

Listing 2. Fragment of Structure File

```
...
<div id="misc7" desc="article">
  <meta-formats>
    <dc xmlns:dc=
     "http://purl.org/dc/elements/1.0/">
      <dc:Title>
   Efficiency of Managed Care Patient
   Protection Laws: Incomplete
   Contracts, Bounded Rationality,
   and Market Failure
      </dc:Title>
      <dc:Creator>
         Korobkin, Russell
      </dc:Creator>
      <dc:Identifier>
         citation:85 Cornell L. Rev. 1
                  (1999-2000)
      </dc:Identifier>
    </dc>
  </meta-formats>
  <pageimage id="9"  native="1"/>
  <pageimage id="10" native="2"/>
  <pageimage id="11" native="3"/>
  <pageimage id="12" native="4"/>
  <pageimage id="13" native="5"/>
 ...
</div>
...
```

3.4 User Interface

The HeinOnline user interface was developed at Cornell Information Technologies, with input from Daniel Rosati, Senior Vice President of William S. Hein Co., Claire Germain, professor of law and Edward Cornell Law Librarian, and her colleagues at Cornell Law Library, and law school librarians at several other institutions.

JSTOR, Lexis-Nexis and Westlaw — other popular online research tools — were used as the comparison standards, and second- and third-year Cornell law students in advanced legal research courses tested system prototypes.

HeinOnline displays the exact image of a page. This protects the integrity of the original document and ensures its authenticity. Additionally, since users have access to text derived from the images, they can copy text and paste it into their notes and papers. Both metadata and full text are searchable.

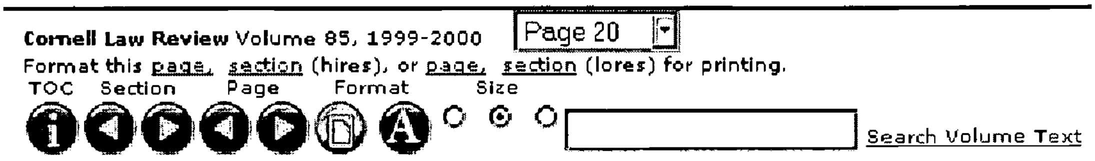

Cornell Law Review Volume 85, 1999-2000
Page 20
Format this page, section (hires), or page, section (lores) for printing.
TOC
Section
Page
Format
Size
Search Volume Text

Figure 1. HeinOnline Document Browsing Control

3.4.1 Hunter Routines

The HeinOnline system employs a suite of Perl and JavaScript routines, collectively known as "Hunter", to navigate the journal volumes as structured Dienst documents. When a volume is opened, a server CGI routine issues a Dienst request for the associated structure file. After analyzing that file, JavaScript objects representing the page data and hierarchic structure of the volume are generated. These are sent with the Hunter JavaScript routines to the client web browser. The client can now access any of the named (numbered) pages of the volume, and browse from section to section (e.g., article to article) via direct Dienst requests to the HeinOnline repository. This design relieves the server of maintaining state information on the document the user is accessing and of re-parsing a representation of the document structure for every client interaction.

Hunter client functions include:

- next page
- previous page
- next section
- previous section
- go to (named) page
- format page for printing
- format section for printing
- select image size
- display page image / display OCR text of page (toggle)

The HeinOnline end-user interface to access these function is shown in Figure 1.

The "go to (named) page" drop-down menu is used to allow the user to browse the page sequence as it is bound in the printed volume without knowing the page number sequence. Because law school journals are often published by students, page numbers are occasionally non-sequential and even repeat within a volume. Navigating a list of actual page labels allows the reader to choose any page in spite of such anomalies.

3.4.2 Citation Access Widget

One of the most time-consuming aspects of writing and reviewing legal papers is the checking of numerous citations. A citation-based navigation widget was one of the most requested features in early prototypes of the system. The citation widget allows a user to enter a citation in a standard Bluebook[6] format, and the system opens the journal volume to the requested page.

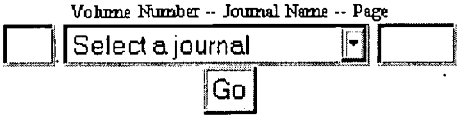


Figure 2. Citation Navigation Widget

Law librarians often receive requests from patrons for copies of journal articles. Satisfying these requests has been labor intensive, involving retrieving the paper volume (perhaps recalling it), copying the article, then generally faxing it to the patron. Libraries using HeinOnline report that they use the citation widget to access the article, format it, save it as a web page, and e-mail the result to the patron, with great labor and copying savings.

The printing formatter complements the citation widget by placing the standard page citation on each printed page. This simplifies record keeping for researchers while they are collecting references.

3.5 Preservation

Digital libraries are increasingly being considered by libraries as a way to ease shelf-space pressures. Now that HeinOnline has put historical law journals online, law libraries have the option of keeping only single copies of those journals on their shelves, putting them in storage or discarding them, depending on their preservation policies.

Hein is a republisher of historical legal materials both in print and microfiche and, as such, has long been interested in preservation issues. Since 1920, William S. Hein & Co., Inc. has specialized in locating rare and out-of-print collections, reprinting government documents and periodicals, converting archive collections to microforms, and preserving legal classics. Hein's holdings include over 75 million pages.

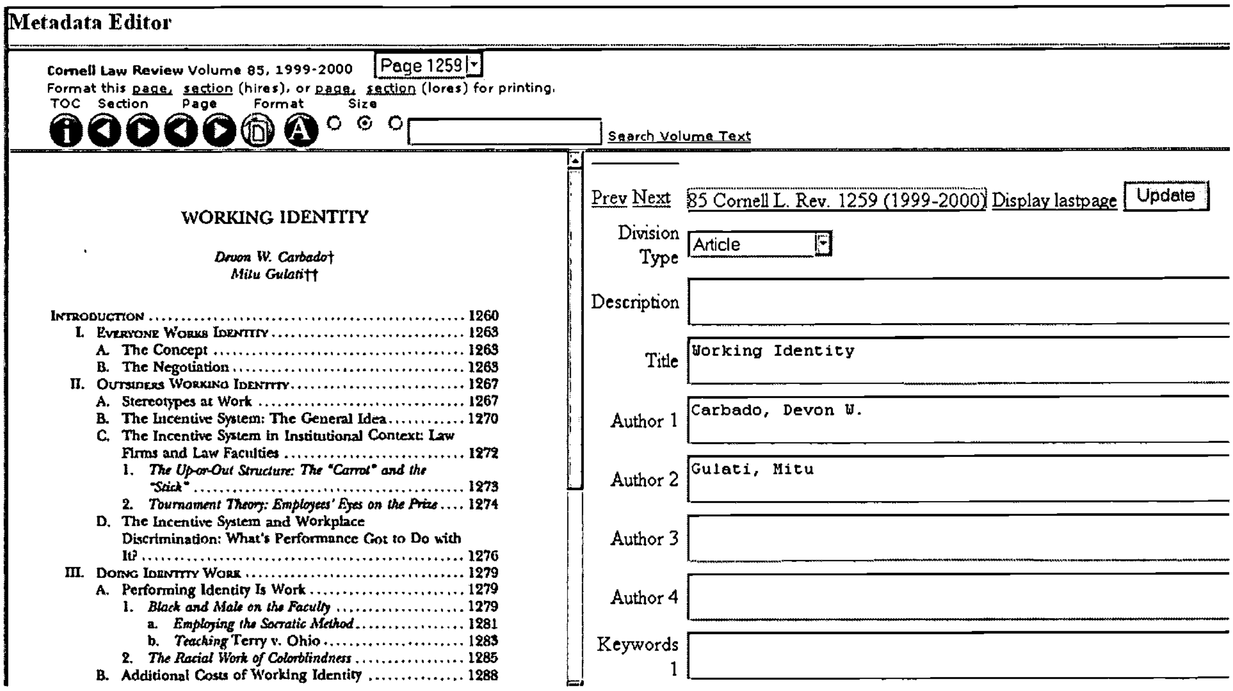
Print and microfiche archives of source materials are maintained in the firm's Buffalo, NY and Littleton, CO facilities. Electronic images of pages, generally bi-tonal 300-dpi, group-4 compressed TIFF format images on optical media are maintained in two separate Buffalo facilities. These images, originally created to enable print reproduction, are the masters for the HeinOnline system.

4. CREATING A WORKING LIBRARY

4.1 Work Flow

Hein scans pages of unbound journal volumes using a 2-sided scanner with a straight-line paper path – a must for old, and often brittle, pages. Page images are collected and organized using the Xerox Digipath[6] scanning system.

[6] http://www.xerox.com



Metadata Editor

Cornell Law Review Volume 85, 1999-2000 Page 1259

Format this page, section (hires), or page, section (lores) for printing.

WORKING IDENTITY

Devon W. Carbado†
Mitu Gulati††

85 Cornell L. Rev. 1259 (1999-2000)

Division Type: Article

Description:

Title: Working Identity

Author 1: Carbado, Devon W.

Author 2: Gulati, Mitu

Author 3:

Author 4:

Keywords 1:

Figure 3. HeinOnline Metadata Editor Screenshot

A Digipath operator records the "native" page numbers (those which appear on the page image) and identifies the beginning and end of each journal section (article, index, editorial, etc.). This information is exported from the Digipath system as a PDF file, which contains the page number and structure information as "bookmark" data.

The PDF file is then processed to extract the bookmark and page image data to non-proprietary formats using ISIToolBox[7] and Ghostscript[8]. A locally developed Perl script converts the bookmark data to an XML-encoded Dienst structure file. The text of the TIFF page image is extracted using an OCR ("optical character recognition") program, ScanSoft TextBridge[9]. A word occurrence index for each journal volume is generated from the OCR text files.

4.1.1 Entering Metadata

Fashioned from the Hunter suite and an additional server-side script, a web-based metadata editor (Figure 3) allows a production operator to enter or update the descriptive Dublin Core metadata for any division of the structure file hierarchy. Additional controls allow the operator to navigate the structural hierarchy and display the associated pages, for example accessing the last page of a section (where in some journal styles, the author's name is printed).

4.1.2 Quality Control

When scanned pages are collected and organized, every page image is viewed to be sure that it is present, readable and not excessively skewed. Similarly, the output of each operation (OCR, image extraction) is inspected to verify that the operation completed successfully.

The digital library is dependent on the quality of its metadata. As descriptive metadata is entered into the metadata editor, an operator checks the structural metadata for consistency by reference to a bound copy of the journal. After metadata is entered, a second operator copyedits the descriptive metadata record for each division in the structure file, using the metadata editor to display images of the original pages.

A style guide was developed for entry of article titles, author names, and other descriptive entries. Division types are assigned by selecting from a controlled vocabulary of types on a pull-down menu.

4.1.3 Placing Content into Production

The TIFF images, OCR text, structure file and index comprise the primary data served by Dienst. These are archived at Hein locations on optical media and copies are shipped on CDs to Cornell University, which houses the external server. The CDs are copied onto spinning magnetic media and integrated into the collection using automated scripts to build searchable metadata and full text databases, as described below.

4.2 Derivative Data

Page images, such as multi-tonal PNG images formatted for display by web browsers, are generated on the fly from the bi-tonal TIFFs by the Dienst Repository server. Derivative image

[7] http://www.imagesolutions.com/isi_software.htm

[8] http://www.cs.wisc.edu/~ghost/

[9] http://www.scansoft.com/products/tbpmill/

generation is done at user-selectable resolutions to accommodate various client configurations and to aid the visually impaired.

While OCR text data could, in principle, be generated on the fly, the latency time for generation and the need to access many pages in single search operations led us to store a text file of each journal volume page.

4.3 Subscription Management

The initial set of subscribers to HeinOnline are primarily law libraries in universities or state and federal government offices. HeinOnline is made available to their patrons within local facilities or across campuses or other facilities.

HeinOnline downloads a JavaScript application to client web-browsers to enable browsing of journal pages. The application makes direct requests to the Dienst Repository for content, therefore subscription enforcement must be integrated directly into the Dienst Repository server.

The initial implementation of subscription enforcement in HeinOnline relies on IP address restrictions. The Dienst Repository service checks a database of allowed IP address ranges which correspond to subscriber campuses or facilities. The database is administered by Hein, which does all subscription servicing. This strategy works well for these customers which are often assigned blocks of static IP addresses. Customers with dynamically assigned addresses, but which use an IP proxy server, can also be serviced with an IP address enforcement scheme because the proxy has a fixed IP address.

However, enforcement of subscriptions by IP address does not work for individuals who are not affiliated with subscribing institutions or who do not have fixed addresses (e.g., dial in or cable modem customers), so an alternate method for authentication and authorization is necessary. For this reason, a cookie-based session management system was built into the subscription enforcement mechanism. This system is currently only used to facilitate guest access to the system, and will be expanded to general subscriber use as system usage grows.

4.4 Full Text Searching

4.4.1 Uncorrected OCR

The text generated from the page images usually contains recognition errors, which result in misspelled or unrecognizable words. These errors are more frequent in pages with ornamented or antique fonts, or in volumes produced with less precise printing technologies. Broken letters characteristic of early printing processes introduce some systematic errors.

For financial reasons, the generated text used in HeinOnline is not corrected or edited. Experience has shown that OCR text errors have minimal effect on search function ability to find relevant articles, thanks to the redundancy of word usage in English prose.

4.4.2 Index Files

Two levels of index files are created to facilitate full-text searching. For each volume, the set of derivative text pages is read to determine a list of unique words and the list of page images on which each word occurs. This list is alphabetically sorted and written as a text file (a volume index), one word (and list of page images) per line.

A second index is created, this time by using the indexes previously generated to create a list of all words which occur in the entire collection. The list of volumes in which each word occurs is recorded, and the alphabetized list of words and volumes is written as another text file (a collection index).

When a user enters a set of search terms, the collection index is consulted to determine which volumes contain the referenced terms. Then the volume indexes for those volumes are consulted to determine if the user's terms occur together (or in a user specified combination) on the same page. The list of selected volumes is then displayed to the user, with an indication of how many pages in each volume may be of interest.

When the user selects one of the relevant volumes, the derivative text files for the specific pages containing the search terms are accessed, and the lines of text containing the terms in the user's search request are displayed in the user's browser, along with links which will take the user to the specified page.

This simple structure has several advantages. It performs well with a large number of pages. It is easy and quick to update the indexes as the collection grows, although care must be taken to use efficient algorithms when processing a million-plus pages of text. It is easy to limit searching to specific titles or volumes.

4.4.3 Metadata Harvesting

To enable searching by article title and author, the descriptive metadata in the Dienst structure files is harvested into an SQL database table. This is done automatically by issuing a Dienst Repository List-Contents command to determine the set of documents currently in the repository, and then sequentially requesting the structure file for each volume. Each structure file is parsed and analyzed and the SQL table is updated appropriately.

4.4.4 Open Archives Initiative Server

This same metadata harvesting technique may be used to update a database which supports an OAI (Open Archives Initiative) server.

The OAI protocol[10] is an application-independent interoperability framework for *metadata harvesting*. *Data Providers* like HeinOnline use the OAI protocol as a means of exposing metadata which describes their content. *Service Providers* may issue OAI protocol requests to data providers for the metadata (the "harvesting") and use the metadata as a basis for building value-added services.

HeinOnline participated in the alpha-test of the OAI protocol in 2000-2001.[11] Volume level and/or article level metadata from HeinOnline may be served via OAI, per the subscription policy.

5. PRODUCTION EXPERIENCE

Working two shifts, six days per week, HeinOnline has been in full production since mid-2000. Since coming online in July, 2000, the number of subscribing institutions has grown to over 125.

[10] http://www.openarchives.org/OAI/openarchivesprotocol.htm

[11] http://www.openarchives.org/OAISC/alpha-testing-press-release.htm

The HeinOnline collection has grown to over 1,600 volumes, encompassing over 40,000 articles. Titles include *Cornell Law Review*, *Texas Law Review*, *Harvard Journal of Law and Technology*, *University of Pennsylvania Law Review*, *Tulane Law Review*, among many others. Some titles include full runs from inception to the current volume; for other titles only the earliest volumes have been processed.

6. FUTURE PLANS

6.1 Collections

Following the initial focus on law journals, plans are underway to add additional collections to HeinOnline including:

- *International Documents* which will include items such as the Nuremburg Trials and Classics of International Law,
- *Case Law* which will include exact reproductions of the first one hundred volumes of U.S. Reports, and
- *Legal Classics* which will initially include collections such as Blackstone's Commentaries and Elliott's Debates.

6.2 Repositories

The Dienst architecture is designed to work seamlessly with collections of materials which are distributed across the Internet.

To accommodate the growing body of material, the HeinOnline system will be split across multiple Dienst servers in 2001. In addition, we have plans to duplicate the collection in multiple locations to insure reliability and performance. To implement and manage this configuration, we expect to use the Dienst Collection service.

The Collection service maintains a registry of Repository servers (as well as other data), and allows clients to discover which servers are operating and which hold documents under the various naming authorities.

6.3 Full Text Searching

The full text search capability in HeinOnline is basic, but performs its main objective well, that is, to locate the set of articles relevant to the users' search criteria.

We have experimented with fuzzy text matching which overcomes some OCR recognition errors at the expense of additional false positives. The fuzzy match technology also assists matching over singular and plural terms and over words with the same word-stem.

We anticipate adding more functionality to the full text search modules as the archive grows.

7. ACKNOWLEDGEMENTS

Thanks to Steve Worona for launching and shepherding the formation of the HeinOnline project at Cornell. Thanks to Carl Lagoze for many discussions on the Dienst protocol. Thanks to Beth Goelzer Lyons for contributions to the Introduction.

The HeinOnline project was supported by the William S. Hein & Co., Inc. and Cornell Information Technologies. The Cornell Law Library staff contributed invaluable suggestions and insight. Support for the Dienst protocol work came from the National Science Foundation through grant IIS-98174416.

8. REFERENCES

[1] Cornell Digital Library Research Group. Dienst Overview and Introduction. http://www.cs.cornell.edu/cdlrg/ - dienst/DienstOverview.htm

[2] Corporation for National Research Initiatives. CSTR Computer Science Technical Reports. http://www.cnri.reston.va.us/home/cstr.html

[3] Corporation for National Research Initiatives. Handle System. http://www.handle.net

[4] Davis, James R., Lagoze, Carl. The Networked Computer Science Technical Report Library. http://cs-tr.cs.cornell.edu/Dienst/UI/1.0/Display/ - ncstrl.cornell/TR96-1595

[5] Davis, J., et. al. Dienst Protocol Specification. http://www.cs.cornell.edu/cdlrg/dienst/protocols/ - DienstProtocol.htm

[6] Harvard Law Review Association. The Bluebook: A Uniform System of Citation.

[7] OCLC. Dublin Core Metadata Initiative. http://purl.org/DC

Designing a Digital Library for Young Children: An Intergenerational Partnership

Allison Druin, Benjamin B. Bederson, Juan Pablo Hourcade,
Lisa Sherman, Glenda Revelle, Michele Platner, Stacy Weng
Human-Computer Interaction Lab
University of Maryland
College Park, MD 20742 USA
+1 301 405 7406
allisond@umiacs.umd.edu
http://www.cs.umd.edu/hcil/searchkids/

ABSTRACT

As more information resources become accessible using computers, our digital interfaces to those resources need to be appropriate for all people. However when it comes to digital libraries, the interfaces have typically been designed for ·older children or adults. Therefore, we have begun to develop a digital library interface developmentally appropriate for young children (ages 5-10 years old). Our prototype system we now call "SearchKids" offers a graphical interface for querying, browsing and reviewing search results. This paper describes our motivation for the research, the design partnership we established between children and adults, our design process, the technology outcomes of our current work, and the lessons we have learned.

Categories and Subject Descriptors

H.5.2 **[Information Interfaces and Presentation]**: User Interfaces – *graphical user interfaces, interaction styles, screen design, user-centered design*. H.3.7 **[Information Storage and Retrieval]**: Digital Libraries – *user issues*. H.3.3 **[Information Storage and Retrieval]**: Information Search and Retrieval – *query formulation*. D.2.1 **[Software Engineering]**: Requirements/Specifications – *elicitation methods*.

General Terms

Design, Human Factors.

Keywords

Children, digital libraries, information retrieval design techniques, education applications, participatory design, cooperative inquiry, intergenerational design team, zoomable user interfaces (ZUIs).

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

1. THE NEED FOR RESEARCH

A growing body of knowledge is becoming available digitally for adults and older students. Far less, however, has been developed with interfaces that are suitable for younger elementary school children (ages 5-10 years old). Children want access to pictures, videos, or sounds of their favorite animals, space ships, volcanoes, and more. However, young children are being forced to negotiate interfaces (many times labeled "Appropriate for K-12 Use") that require complex typing, proper spelling, reading skills, or necessitate an understanding of abstract concepts or content knowledge that are beyond young children's still-developing abilities [13, 18, 20]. In recent years, interfaces to digital libraries have begun to be developed with young children in mind (e.g., Nature: Virtual Serengeti by Grolier Electronic Publishing, A World of Animals by CounterTop Software). However, while these product interfaces may be more graphical, their digital collections tend to be far smaller than what is available for older children or adults.

A common trend over the past decade in children's digital libraries interfaces has been to use simulated books as metaphors for traversing hierarchies of information on the screen. One such well-known example in the library community was the Science Library Catalog (SDL) developed in the mid 1990s led by Professor Christine Borgman at UCLA [20]. While this system didn't necessitate keyboard input, it did require reading keywords on the sides of graphical books and reading lists of content results. This system exemplified technologies that were created for older elementary school children (ages 9-12) where reading skills are an important part of the interface. An early example of visual interfaces for accessing digital libraries is BookHouse [14]. This system used icons and a spatial metaphor for searching, but used text lists to show the results.

Novel work in the HCI community has also produced numerous alternative approaches to visualizing searches and their results. One such approach is the "Dynamic Queries" interface developed at the University of Maryland [1]. It enables the user to drag sliders to specify the range of each query element, select from check boxes or radio buttons, or type for string search. Colored and size coded markers for each item represent search results. This approach works well with ordered data that can be filtered by a linear range, for categorical values that can be selected one-by-one, and for nominal values that can be string searched. For young children however, this interface may be cognitively challenging. It is somewhat abstract to connect the idea that changes to the query criteria on the side of the screen result in changes to the visualization of the query results.

On the other hand, a somewhat more concrete approach is "NaviQue," developed at the University of Michigan as a part of their Digital Libraries initiative [10]. With this system, there is no separate space for query results; any object can be used to launch a query. A user simply selects one or more objects and that becomes the query. Then by dragging that data set over another collection of objects, a similarity-based search is launched. The results of the query are highlighted in the data set. While the interaction for this system is deceivingly simple, the abstraction used to query is surprisingly difficult for children to grasp. This system, while extremely flexible, needs more concrete labeling for young children to understand what question they are asking in the query.

Another approach is the idea of "Moveable Filters" based upon the work done at Xerox PARC on lenses [9]. With this graphical query interface, transparent boxes or filters are dragged over a scatter plot of data. Each filter contains buttons labeled for Boolean query operations (e.g., "and", "or"), and a slider that controls the threshold for numeric data. When two filters overlap each other, their operations combine. The results of the query are immediately highlighted. For children, the difficulty in this system lies in the need to understand Boolean query concepts.

Another approach to presenting Boolean searches is to use Venn-like diagrams [12]. Developed by the University of Waikato in New Zealand, "V-Query" is a system where users drag circles around query terms. A new term is created by typing it into the workspace. Depending on the placement of the circles, an "and", "or", "not" query can be created. Each time, a dynamic result of digital resources is displayed. This system while somewhat simple to manipulate, still asks users to type keyword terms and read lists of results, both difficult for young children.

While there are many more researchers focusing on graphical direct manipulation interfaces for querying, the handful of examples just discussed shows promising possibilities. However, there are definite limitations to these systems when young children are the users. To address these limitations, we have begun our own research in developing a graphical direct manipulation interface for searching, browsing, and viewing query results of digital libraries. Supported by a 3-year DLI-2 National Science Foundation grant, we began our research in September 1999. Content provided by the Discovery Channel and the U.S. Department of the Interior's Patuxent Wildlife Research Center, has enabled us to develop a digital library prototype devoted to multimedia information on animals. The technologies and teaching strategies we are developing are not limited to this content area, but that is our starting point.

2. THE ROLE OF CHILDREN AND TEACHERS IN THE DESIGN PROCESS

We believe children can play an important role in creating new technologies for children [6, 7]. Therefore, we have established an interdisciplinary, intergenerational team of researchers that include computer scientists, educational researchers, visual artists, biologists, elementary school children (ages 5-11) and classroom teachers. Throughout the research process, we have looked for methods that make use of our diverse points of view and enable each voice to be heard in the design process. During our research activities, not only have we come to understand the impact children can have on the design of children's digital libraries, but we have also come to understand how these new technologies can impact children as users.

These understandings have developed as we have worked with children in two different ways in two different locations. In our HCI lab, we have collaborated with a team of seven children ages 7-11 years of age as "Design Partners." At the same time, we have worked in a local elementary school with almost 100 children 7-9 years old in 2nd and 3rd grades as informants. We saw the design partner children in our lab as having a critical role in the initial brainstorming experiences that would set directions for our digital libraries research. On the other hand, we saw the children in school as informants in helping us to understand if our ideas were generalizable among a diverse population of children. As a team, we have not previously made use of both roles for children in a large-scale research study. In addition, the integration of teachers as design partners in our lab was something new to our group. In the sections that follow, each role will be described in regards to methods, context, and challenges.

2.1 Design Partners

The role of *design partner* for children includes being part of the design process throughout the experience [6]. With this role, children are equal stakeholders in the design of new technologies. While children cannot do everything that adults can do, we believe they should have equal opportunity to contribute in any way they can to the design process. For the past three years, our research team has been developing new technology design methodologies to support children in their role as design partners (Figure 1).

This strategy of working with children as partners is something we have come to call *Cooperative Inquiry* [6]. It combines and adapts the low-tech prototyping of participatory design [11, 17], observation and note-taking techniques of contextual inquiry [5] and the time and resources of technology immersion [7]. Children and adults alike gather field data, initiate ideas, test, develop new prototypes, and reflect by writing in journals. Together we pursue

Figure 1: Children and adults collaborating as design partners in our HCI lab

projects, write papers, and create new technologies [2, 7]. In a subsequent section of this paper (The Design Process), we will discuss in more detail the specific design methods we used in brainstorming our digital libraries technologies.

The current design partner team includes two faculty members, one graduate student, two undergraduate students, two staff members, three teachers, and seven children (ages 7-11 years old). The disciplines of computer science, education, psychology, biology, and art are represented. Members of the team meet two afternoons a week in our lab or out in the field. Over the summer we meet for two intensive weeks, six hours a day.

When we began our digital libraries research in the fall of 1999, we added to our design team three elementary school teachers (one 2nd grade teacher, one 3rd grade teacher and one technology coordinator for the school). The children on our team did not come from the school of those teachers. In addition, the children had already been with the lab team working with University researchers on other projects for a minimum of six months. We did not meet at the teachers' school when we began, but rather in our HCI lab environment. Thanks to this process of introduction for the teachers, the children in some sense became mentors for the teachers who had never before considered developing new software. As one teacher pointed out, "At first I was bit worried that I wouldn't know how to contribute to the team. What did I know about research labs? But the children made it easy. They knew what they were doing. And since I'm not their teacher, I wasn't worried I'd look too foolish." (Teacher Journal, November, 1999).

One of the challenges of this kind of design partnership is that adults are not in charge, but neither are children. Design partners must negotiate team decisions. This is no easy task when children are accustomed to following what adults say, and adults are accustomed to being in charge. Children must learn to trust that adults will listen to their contributions, and adults must learn to elaborate on children's ideas, rather than merely listening passively or not listening at all [2]. This idea-elaboration process takes time to develop, but is something that we have found to be extremely important to work towards in a design partnership. We have found however, that it can take up to 6 months for an intergenerational design team to truly develop the ability to build upon each other's ideas (regardless of who originated the idea). Due to this challenge, the development process can take more time than expected.

On the other hand, a strength of the design partnering experience is that there is no waiting to find out what direction to pursue. A continuous relationship with children can offer a great deal of flexibility for design activities. If researchers know that children will always be available at certain times, then less formal schedules need to be made. Another strength of this partnership is that all members of the design team can feel quite empowered and challenged by the design partner process. Children for example have so few experiences in their lives where they can contribute their opinions and see that adults take them seriously. When a respect is fostered, we have found that it does change how children see themselves [2]. As one child shared with us, "My idea helped the team today. The adults saw we don't need books on the screen. I was cool" (8-year old Child Journal, December, 1999).

2.2 Informants

In our lab's previous research [19], we attempted to adapt the design partner experience to school settings in Europe. What we found is that the parameters of the school day and the existing power structures between teachers and students, made it quite difficult to develop a true design partnership. Very little time could be devoted to the necessary activities in building a partnership. Therefore, in looking to involve more children and teachers in the technology development process, we chose to integrate the role of *informant* in our research. This role became more clearly defined in the late 1990s by Scaife and Rogers from the University of Sussex [16]. They described the notion of "informant design" and questioned when children should be a part of the design process. Before this time, numerous researchers were including children in the design process, but not making a distinction of when. Were children testers at the end of the design process? Were children partners contributing throughout the process? Were children informants helping the design process at various critical times?

With this role of informant, children play some part in informing the design process. Before any technology is developed, children may be observed with existing technologies, or they may be asked for input on paper sketches. Once the technology is developed, children may again offer input and feedback. With this role, young people can play an important part in the design process at various stages, but not continuously as is the case in a design partner experience.

For our digital libraries research, we found this method of working with children much easier to negotiate in a school setting. We had the opportunity to work with an ethnically diverse population of children, yet we minimally disrupted their busy school day. We learned from these children how our digital libraries technologies should be changed to make them more useable by children with a wide variety of backgrounds and styles.

In all, 100 children have been working with our research team as informants. 50% of the children are males and 50% females. 52% are Caucasian, 36% African American, and 22% are either Asian or Hispanic. To work with our team, same-sex pairs of children were pulled out of their regular schedule for no more than one-hour at a time, for no more than three times over the school year. The children worked with one to two university researchers for a session. While this may seem quite minimal in time contribution, it did complement quite well the on-going research efforts of our design team back at the lab. Since the children we work with at the school are taught by the teachers who are also our design partners, we have run into much less resistance to changes in the school day than one might expect. The teachers have taken ownership of the technologies we are developing, since they too are designing them in partnership. Yet this partnership minimally impacts their busy school day. For details of the methods we used as informants and design partners, see the section that follows.

3. THE DESIGN PROCESS

We began our digital libraries research with what we call a "low-tech prototyping" session. Before the teachers or children looked at any other systems, we thought it was important for them to brainstorm without consideration to previous work. We felt that this would encourage a feeling that anything was possible. The team was split into three groups consisting of 2-3 children, 1 teacher, and 1-2 university researchers. Each group was asked to

Figure 2: Note from children's journals on what an animal digital library should look like

design a digital library of the future that contained all of the animal information they ever wanted know. To do this, each group used low-tech prototyping materials (the children call "bags of stuff") containing paper, clay, glue, string and more. From this brainstorming session, three low-tech prototypes were developed that generated ideas for digital libraries (e.g., the interface did not have to look like a book, the interface should be specific to the content area—in our case animals, the interface should use graphical representations as queries).

Following this experience, the team spent some time using and critiquing various children's digital libraries systems that contained animal content: *The Magic School Bus Explores the World of Animals* by Microsoft, *Amazing Animals Activity Center* by DK Multimedia, *Premier Pack: Wildlife Series* by Arc Software, The National Zoo (www.si.edu/natzoo), and Lincoln Park Zoo, Chicago (www.lpzoo.com).

We had two children use a particular technology and one teacher and one university researcher observe their use. While the children were using the technologies the adults were writing down what the children were saying and doing during the session. Meanwhile the children were also taking notes. They wrote on "sticky notes" three things they liked about what they were using and three things they did not. When the sessions were over we collated the sticky notes on the board and looked for frequency patterns in likes and dislikes. Two overwhelming conclusions that came out of these sessions were: (1) there needs to be a purpose for the search and something needs to be done with the information once it is collected; (2) the use of animated characters to tell a child what to do were extremely annoying to the children. At the beginning of our "sticky note session," the adults on the team were quite baffled by numerous sticky notes with comments such as, "It doesn't do anything" "I was bored at looking" "Nothing happens" (Researcher notes, November 1999). As it turned out the children were explaining that it just wasn't good enough to search for things, they wanted to use them to make something. The one application that did allow them to do something with their images, the children found particularly annoying due to the use of an animated character that kept telling the children what to do. After the session, the adults on the team compared their notes, and found that their observations were very much the same as the children's.

The team then spent a few sessions brainstorming and drawing in their journals (Figure 2). From this experience, a few critical ideas crystallized for the team. One idea the team particularly liked was the metaphor of going on a journey. One of our 8-year old design partners explained that "Finding things is like going on a trip, so you should go with friends" (Researcher notes, December 1999). She thought that these friends shouldn't be "pushy" like the character we saw, but should give kids a reason for wanting to find things. Another idea that emerged was that the interface should be based on animals "the thing you're looking for." The notion of dragging animal parts that represented things you wanted to search for came out in a number of journals. So instead of a text question of "what do animals eat," a picture should be dragged into a "mixing space" that represents that question. Other ideas that emerged had to do with the questions that the children wanted answered about animals. These included: (1) what do they eat; (2) how do they move; (3) where they live; (4) what animal family are they part of. One additional area of information that an 11-year old design partner wanted to know more about was "what waste products do animals make?" Even though the children loved this idea, it was decided that the information would be so hard to find, that this would have to wait for version 2.

Other ideas that emerged from the teachers were also critical in structuring our approach to digital libraries. One teacher pointed out that in the youngest grades, the children learn about animals grouped by "pets at home" or "farm animals." While older children learn about animals by where they might come from geographically (e.g., Australia, Africa, etc.). Therefore, various ways to browse for animals were needed, so that children at different grade levels could take advantage of the library. As the teachers pointed out, there are big differences between what a 2nd grade teacher needs to cover as compared to a 3rd grade teacher, even though this represents only one year's difference in the children's ages.

Soon after this set of sessions, three members of our team began working with 50 elementary school children in our local school. We realized that as a team we knew very little about how young children actually searched for animals, and how complex their queries could actually be. To understand this, we conducted an empirical study at the school to develop an understanding of how children searched based on what we had already learned in the lab [15]. We developed a set of hierarchically nested envelopes based on the four categories of information our child design partners were interested in (e.g., habitat, food, movement, and animal taxonomy). The children in the school were asked to search within those envelopes for pictures of animals.

From observing the children's behavior in this situation, we learned that the children appear to search very differently depending on gender. For example, we found that boys tended to dump all the envelopes on the floor (with little thought of putting things back) in search of the animal they wanted. On the other hand, the girl teams tended to be quite careful in their search style, but at times seemed to be more interested in browsing the pictures rather than finding the exact animal in question. This led us to the notion that the application should fully support both structured searching and browsing as equally valid and efficient methods of accessing information.

Our next step back at the lab was to begin designing an "interactive sketch". By this we mean something that could begin to help us get a feel for some of the ideas that had emerged in our previous design sessions. For this we used KidPad, a zoomable authoring tool for children [4, 8]. The group's artist began sketching with this tool, and as she sketched, the team refined its ideas. The notion of how to use characters became clearer to the team. These were not characters that told you to do things, but

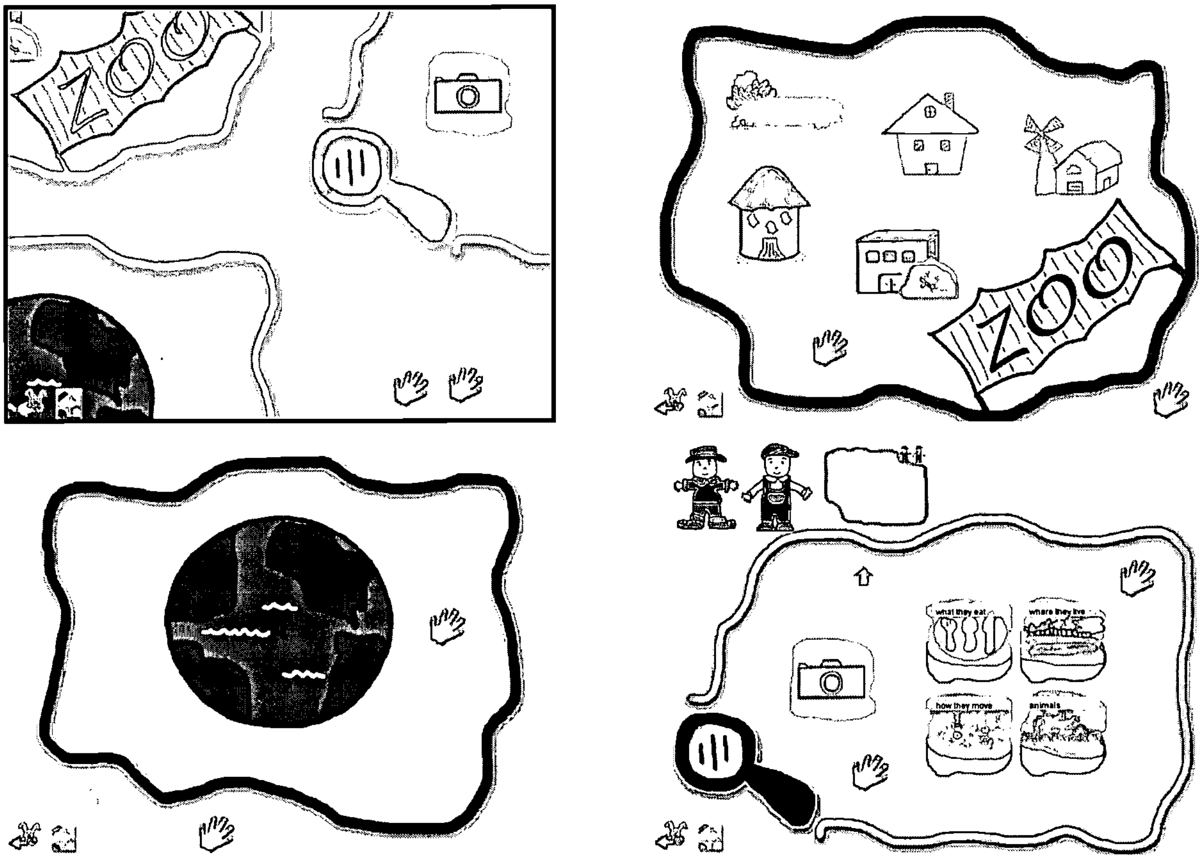


Figure 3: From left to right, and top to bottom:
The prototype's initial screen, the zoo area, the world area, and the search area.

rather, they represented the query as it was being formulated. The characters held onto the search criteria a child wanted to use. Also the notion of "doing something" with the search results began to take form. Since the team was already helping to develop KidPad (www.kidpad.org), it made sense to link the digital libraries application with an authoring tool. Ultimately this meant building our first interactive prototype on top of KidPad. In addition to these ideas, the concept of having three different areas to look for animals evolved. This took the form of the zoo (with a farm house, a pet house, a bird house, and more), the world, and the search area.

As our technical team was developing the first functional prototype, we continued to refine its interface by using paper chips to represent the search criteria and people to represent the kids. We also populated, in consultation with our team biologist, a Microsoft Access database with metadata on animal images contributed by our content partners. At one point, however, one of our child design partners insisted our biologist had "gotten it all wrong for gorillas" about what they ate, so this 8-year old spent the afternoon looking up on the web what gorillas ate to prove his point (he was quite correct and the metadata was fixed). When our first interactive prototype was far enough along to be usable by someone besides the design team, it was brought back into the school to be used with our informant children. Fifty of them who had not previously taken part in exploring the paper prototype were asked to offer feedback on the computer prototype. This study is reported in detail in [15]. In the section that follows a full description of our current prototype is presented.

3.1 Today's Prototype

As previously discussed, our initial interactive prototype we now call SearchKids is built upon KidPad, a real-time continuous zooming application that our lab originally developed in partnership with researchers at the Royal Institute of Technology, Sweden, the Swedish Institute of Computer Science and the University of Nottingham for the purpose of children's collaborative storytelling [4, 8]. KidPad and SearchKids make extensive use of Jazz, a Java toolkit we developed for research in Zoomable User Interfaces (ZUIs) [3]. SearchKids accesses metadata about images of animals from the Microsoft Access database mentioned in the previous section.

The prototype follows a few interface design principles that we have learned through years of work with our design partners. We made the interface very visual, avoiding the use of text as much as possible and therefore reducing the cognitive load. We also made interactions with the mouse as simple as possible by using a one-click interface (i.e. no dragging, no double-clicking) with all mouse buttons having the same functionality. The fact that we are using a one-click interface makes SearchKids easy to use on touch screens therefore avoiding the problems many young children have controlling mice.

The current version of SearchKids consists of three areas through which users can look for media about animals. Figure 3 shows the prototype's initial screen and the three areas. Users may navigate between areas and within areas by clicking on their destination or making use of the "home" and "back" icons that are always at the bottom left of the screen.

The zoo area provides a way of browsing the contents of our animal database in a familiar setting. When entering the zoo area, users see the map of a virtual zoo. By zooming into parts of the zoo, children can find representations of animals and through them, access media. For example, to access media about lizards, children can zoom into the reptile house and click on a representation of a lizard.

The world area provides a way for children to browse the animal database by looking for animals geographically. It presents children with a globe they can spin and zoom into. By zooming into a region of the world they can find representations of the animals that live in that part of the world and though them, access media. For example, to access media about polar bears, children could zoom into the North Pole and click on a representation of a polar bear. The world area is currently not fully implemented.

The search area gives users the ability to visually specify and manipulate queries. It also provides previews of query results. The initial look of the search area is shown in the right-most picture of Figure 3. The query region makes up most of the search area. The chips in this region are the components from which queries can be formed. The chips on the left side of the region represent the types of media available through the database. Currently, only images are available and a camera represents them. The chips on the right side of the region represent the hierarchies under which the animals in our database have been classified. They enable children to look for media about animals based on what they eat, where they live, how they move, and a biological taxonomy.

To explore these hierarchies, users can click on the shadows under the chips. For example, clicking on the "what they eat" chip brings into focus three chips representing animals that are carnivorous, herbivorous, and omnivorous. To move up in a hierarchy, users can click on the up arrow to the left of the hierarchical chips. Figure 4 shows the search area after the shadow under the biological taxonomy chip has been clicked on. The chips on the right side of the figure are the children of the biological taxonomy chip and represent the types of animals present in the database.

When a chip (media or hierarchical) is clicked on, it zooms towards one of the characters on the top-left corner of the screen (Kyle and Dana, the "search kids"), to hold around their neck. This chip becomes part of the query criteria. Media chips zoom to Kyle while hierarchical chips go to Dana. Clicking on a chip that is on Kyle or Dana makes it go back to its original location therefore removing it as one of the criteria for the current query.

The chips on Kyle and Dana visually represent the queries children formulate. Our prototype returns an intersection of the media items represented by chips selected from different categories and a union of the media items represented by chips selected from the same category. This approach, while somewhat limiting expressive power, successfully enables children to specify their desired queries and does so without requiring them to explicitly distinguish between unions and intersections. Figure 5 shows a series of screenshots that demonstrate how children may pose a query.

The red region to the right of Kyle and Dana shows the results of the current query. Children can zoom into the region by clicking on it. By seeing the results of their queries as they pose them, children can quickly tell whether the database has any items that correspond to their query criteria. The items shown in the results area can be zoomed into (this feature was not available during testing at the school). We are currently working on a mechanism that will allow children to transfer media of their choice into KidPad.

This prototype has been used with our child informants in school and the results have been encouraging. The differences by gender the children displayed in their searching disappeared when they used this prototype [15]. In addition, children were able to

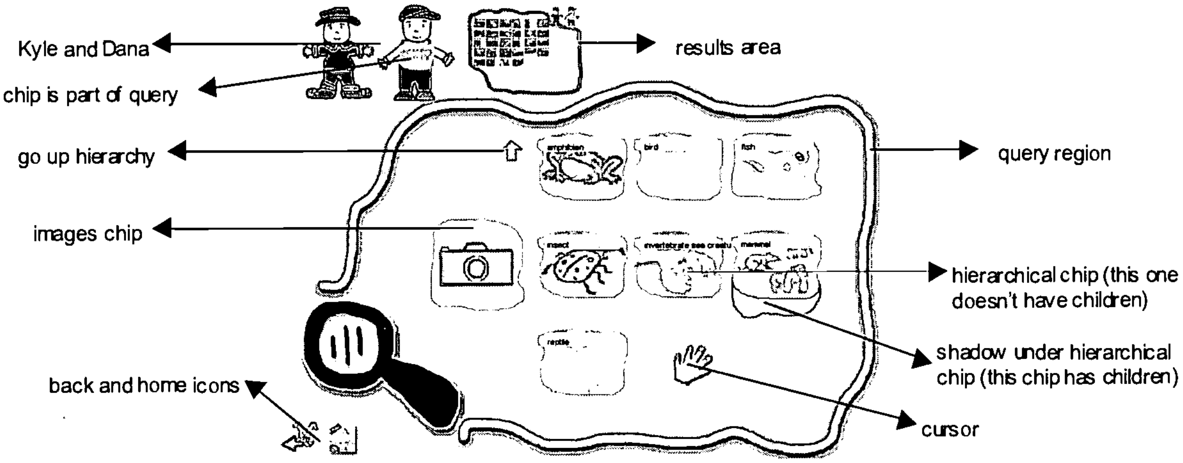


Figure 4: Components of the search area when exploring biological taxonomy.

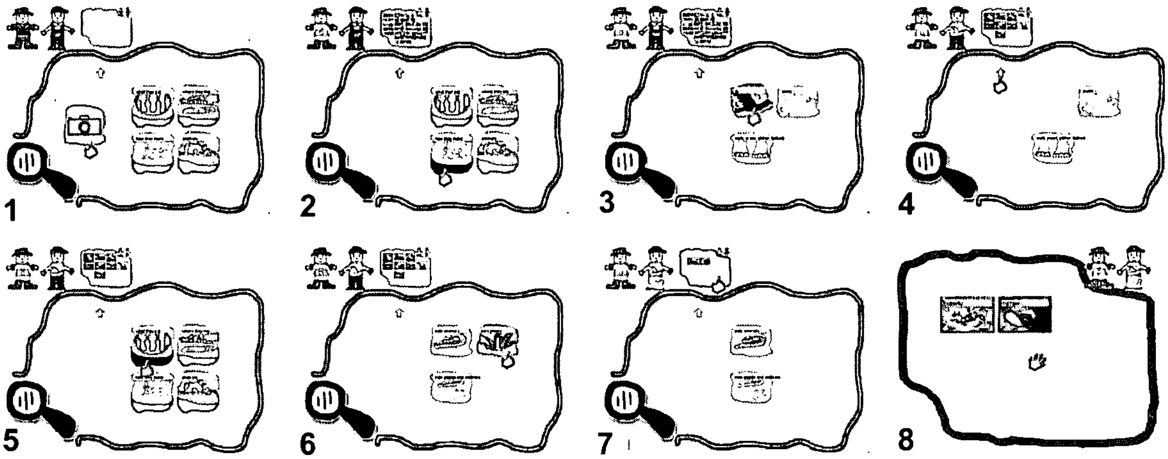


Figure 5: Process of querying for images of animals that fly and eat plants.

1. **Child clicks on the item representing images.**
2. **Child clicks under "how they move" category (notice the thumbnails in the results area, and the camera on top of Kyle).**
3. **Child clicks on "fly" item.**
4. **Child clicks on up arrow to go up in the hierarchy. The query at this point is asking for images of animals that fly. Notice there are less thumbnails in the results area.**
5. **Child clicks under "what they eat" category.**
6. **Child clicks on "eats plants" item. This completes the specification of the query.**
7. **Child clicks on results area.**
8. **Child browses results in results area.**

construct more complex queries with SearchKids than with the paper prototype. However, most of the children did encounter some difficulty with the size of the images in the results screen, and the size of the navigational controls for up and back, but that has already begun to be addressed in later versions of the prototype.

4. LESSONS LEARNED

While our project is only in its second year, we have learned a number of lessons in regards to design process as well as digital library technologies. In terms of the design process, the combination of children as design partners in the lab and children as informants in the school helped considerably. We were able to quickly brainstorm possibilities with children, yet minimally disrupt school schedules or renegotiate power-structures between children and teachers. What we did come to understand was that without a design partner experience, child informants in the school could merely offer feedback on ideas presented to them, as opposed to elaborate or build upon ideas as was the case in our lab.

Another lesson learned in our design process concerned the teachers. By introducing the teachers the way we did with a delay and with children they did not teach, we helped to equalize the footing between child and adult. We found the teachers learning from the children in the group and the children not treating the "teachers" as they might normally. Yet thanks to this partnership, the teachers quickly embraced the technology as their own, and helped a great deal in contributing to the design and content structure of the digital library, as well as facilitating our work in the school.

In terms of lessons learned concerning the technology, one of the most interesting was that children don't want to just search for information, they want to use it too. They want a reason to search or browse for items (besides some adult saying to look for it). This led us to a firm belief that our work is also in developing a connection between our digital library and authoring tools.

In addition, the notion of a content specific interface also emerged quite strongly. Needless to say, if we were developing an interface for a digital library containing all forms of plants, it would not make sense to have a zoo browsing area. But it does make sense that a content specific metaphor is critical for children. To some degree they see the digital library as not a library with books, but as a place to wander about looking for different kinds of information.

5. FUTURE DIRECTIONS

In terms of future directions, we look forward to exploring the possibilities of multi-user navigation and searching. Since our application is built upon KidPad, we have the functionality built right in to have multiple mice work at the same time. We are exploring what can happen when children collaborate as they navigate information.

In addition, we are enhancing the database content by adding video, sound, and text items. We are also developing a direct connection from SearchKids to KidPad. With these major additions to our prototype interface, we expect further empirical studies will be needed, especially those with younger children (ages 5-6).

6. ACKNOWLEDGMENTS

This work could not have been accomplished without our child design partners in the lab Alex, Abby, Jack, Joe, Lauren, Rebecca and Thomas and our 100 informants in the second and third grade at Yorktown Elementary School. We are also indebted to our teachers led by Technology Coordinator Carol Poundstone and Principal Libby Thomas. On-going inspiration and intellectual discussion has come by way of Ben Shneiderman, Catherine Plaisant, Anne Rose, Joseph Ja'Ja', and the KidStory research project supported by the i3, ESE. Financial support has come from the National Science Foundation's DLI-2, Discovery Channel, Patuxent Wild Life Refuge, and the Baltimore Learning Community project.

REFERENCES

[1] Ahlberg, C., Williamson, C., & Shneiderman, B. (1992). Dynamic Queries for Information Exploration: An Implementation and Evaluation. *In Proceedings of Human Factors in Computing Systems (CHI 92)* ACM Press, pp. 619-626.

[2] Alborzi, H., Druin, A., Montemayor, J., Sherman, L., Taxén, G., Best, J., Hammer, J., Kruskal, A., Lal, A., Plaisant, S. T., Sumida, L., Wagner, R., & Hendler, J. (2000). Designing StoryRooms: Interactive Storytelling Spaces for Children. *In Proceedings of Designing Interactive Systems (DIS 2000)* ACM Press, pp. 95-104.

[3] Bederson, B. B., Meyer, J., & Good, L. (2000). Jazz: An Extensible Zoomable User Interface Graphics Toolkit in Java. *In Proceedings of User Interface and Software Technology (UIST 2000)* ACM Press, pp. 171-180.

[4] Benford, S., Bederson, B. B., Akesson, K., Bayon, V., Druin, A., Hansson, P., Hourcade, J. P., Ingram, R., Neale, H., O'Malley, C., Simsarian, K., Stanton, D., Sundblad, Y., & Taxén, G. (2000). Designing Storytelling Technologies to Encourage Collaboration Between Young Children. *In Proceedings of Human Factors in Computing Systems (CHI 2000)* ACM Press, pp. 556-563.

[5] Beyer, H., & Holtzblatt, K. (1998). *Contextual Design: Defining Customer-Centered Systems*. San Francisco, CA: Morgan Kaufmann.

[6] Druin, A. (1999). Cooperative Inquiry: Developing New Technologies for Children With Children. *In Proceedings of Human Factors in Computing Systems (CHI 99)* ACM Press, pp. 223-230.

[7] Druin, A., Bederson, B. B., Boltman, A., Muira, A., Knotts-Callahan, D., & Platt, M. (1999). Children As Our Technology Design Partners. A. Druin (Ed.), *The Design of Children's Technology* (pp. 51-72). San Francisco: Morgan Kaufman.

[8] Druin, A., Stewart, J., Proft, D., Bederson, B. B., & Hollan, J. D. (1997). KidPad: A Design Collaboration Between Children, Technologists, and Educators. *In Proceedings of Human Factors in Computing Systems (CHI 97)* ACM Press, pp. 463-470.

[9] Fishkin, K., & Stone, M. C. (1995). Enhanced Dynamic Queries Via Movable Filters Papers: Information Visualization. *In Proceedings of Human Factors in Computing Systems (CHI 95)* ACM Press, pp. 415-420.

[10] Furnas, G. W., & Rauch, S. J. (1998). Considerations for Information Environments and the NaviQue Workspace. *In Proceedings of International Conference on Digital Libraries (DL 98)* ACM Press, pp. 79-88.

[11] Greenbaum, J., & Kyng, M. (Eds.), (1991). *Design at Work: Cooperative Design of Computer Systems*. Hillsdale, NJ: Lawrence Erlbaum.

[12] Jones, S. (1998). Graphical Query Specification and Dynamic Result Previews for a Digital Library. *In Proceedings of User Interface and Software Technology (UIST 98)* ACM Press, pp. 143-151.

[13] Moore, P., & St. George, A. (1991). Children As Information Seekers: The Cognitive Demands of Books and Library Systems. *School Library Media Quarterly, 19*, pp. 161-168.

[14] Pejtersen, A. M. (1989). A Library System for Information Retrieval Based on a Cognitive Task Analysis and Supported by and Icon-Based Interface. *In Proceedings of Twelfth Annual International Conference on Research and Development in Information Retrieval (SIGIR 89)* New York: ACM, pp. 40-47.

[15] Revelle, G., & Druin, A. (2001). Young Children's Search Strategies and Construction of Search Queries. *In Proceedings of Human Factors in Computing Systems (CHI 2001)* ACM Press, p. (submitted).

[16] Scaife, M., & Rogers, Y. (1999). Kids As Informants: Telling Us What We Didn't Know or Confirming What We Knew Already. A. Druin (Ed.), *The Design of Children's Technology* (pp. 27-50). San Francisco: Morgan Kaufman.

[17] Schuler, D., & Namioka, A. (Eds.), (1993). *Participatory Design: Principles and Practices*. Hillsdale, NJ: Lawrence Erlbaum.

[18] Solomon, P. (1993). Children's Information Retrieval Behavior: A Case Analysis of an OPAC. *Journal of American Society for Information Science, 44*, pp. 245-264.

[19] Taxén, G., Druin, A., Fast, C., & Kjellin, M. (2000). KidStory: A Technology Design Partnership With Children. *Behaviour and Information Technology (BIT)*, pp. (in press).

[20] Walter, V. A., Borgman, C. L., & Hirsh, S. G. (1996). The Science Library Catalog: A Springboard for Information Literacy. *School Library Media Quarterly, 24*, pp. 105-112.

Dynamic Digital Libraries for Children

Yin Leng Theng, Norliza Mohd-Nasir, George Buchanan, Bob Fields, Harold Thimbleby,
and *Noel Cassidy
School of Computing Science, Middlesex University, London, N11 2NQ, +44 208 362 6926
*St-Albans School, Abbey Gateway, St Albans, AL3 4HB
{y.theng, norliza1, george10, b.fields, h.thimbleby}@mdx.ac.uk;
*Noel.Cassidy@mail.btinternet.com

ABSTRACT

The majority of current digital libraries (DLs) are not designed for children. For DLs to be popular with children, they need to be fun, easy-to-use and empower them, whether as readers or authors. This paper describes a new children's DL emphasizing its design and evaluation, working with the children (11–14 year olds) as design partners and testers. A truly participatory process was used, and observational study was used as a means of refinement to the initial design of the DL prototype. In contrast with current DLs, the children's DL provides both a *static* as well as a *dynamic* environment to encourage active engagement of children in using it. Design, implementation and security issues are also raised.

Keywords

Design process, design partners and testers, participatory design, collaborative writing, observational study, ethnography.

1. INTRODUCTION

The design of systems, including DLs, is often inspired by what technology makes possible. In user-centered design, design emphasizes users, their tasks and needs. This paper shows how observational and participatory work with children as users resulted in the design of a DL with novel — and useful — features. Beyond summarizing the design itself, a main contribution of the paper is making explicit the relationship between design and observational study, in particular video analysis, that inspired the refinement of the initial design of the dynamic component of a children's DL. The paper also mentions implementation and security, and discusses directions for future work.

We will argue that if DLs are to be popular, they need to be easy-to-use and empowering for users both as *readers* and as *authors*. DLs should provide both static as well as dynamic features.

1.1 Static vs dynamic DLs

The history of DLs is rich and varied because the "digital library" is not so much a new idea as an evolving conception of contributions from many disciplines. In recent years, there has been an emergence of subject-based DLs on the Web. Many people have contributed to the idea, and everyone seems to have something different in mind! The metaphor of the traditional library is both empowering and constraining [8]: empowering, because DLs automate and extend opportunities offered by traditional libraries, as well as harnessing opportunities not possible on the anarchic web; constraining, because the metaphor evokes certain legacy impressions, many originating in arbitrary physical constraints.

Because DLs mean different things to different people, the design of the DLs is, therefore, dependent of the perceptions of the purpose/functionality of DLs.

To the library science community, the roles of traditional libraries are to [13]: (a) provide access to information in any format that has been evaluated, organized, archived and preserved; (b) have information professionals that make judgements and interpret users' needs; and (c) provide services and resources to people (students, faculty, others, etc.). Others think that DLs may mean carrying out functions of libraries in a new way (e.g., [8; 18] etc.). It may be encompassing new types of information resources, new methods of storing and preservation, and new approaches to classification and cataloguing.

To the computer science community, DLs may refer to (e.g., [8]; etc.) a distributed text-based information system, a collection of distributed information services, a distributed space of inter-linked information system, or a networked multimedia information system.

Levy and Marshall [12] argue for DLs to be broadly-construed so that "the design of DLs must take into account a broader range of materials, technologies, and practices," and they emphasize the importance of access and use of the collection by a community. Miksa and Doty [14], however, argue for a narrowly-construed definition of DLs, emphasizing the role of collection and intellectual access to it.

Hence, we have the difficulty of precisely classifying DLs. In this paper, we will group them according to their collections and whether there are features provided in the DLs to allow user-initiated activities (for example, annotations, reviews, etc.) to append additional data to the organisational memories of the collections. Organizational memories may include, for example, informal information, time-sensitive information or bulletin board mechanisms, and they tend to be more dynamic [2]. We agree with Levy and Marshall [12] that DLs should contain "transient as well as permanent documents, fluid as well as fixed materials, paper as well as digital technologies, and collaborative as well as individual practices" (p. 163).

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

In this paper, we call these kinds of DLs *dynamic*, in that the organizational memories of the collections can be modified through user-initiated actions, and the environment provides a social space for collaborative and individual practices. In contrast, DLs that permit only browsing and retrieval are termed as *static*. Of course, in both static and dynamic DLs, the collections generally grow over time, but the emphasis in a dynamic DL is that the authors are primarily in control of the collection.

Table 1 identifies various features of a representative sample of DLs that have been implemented. The list is not exhaustive (and not intended to be) but gives a flavor of the nature of current popular DLs for academic and commercial purposes designed for adults and children. (Owing to constraint of space, we will not describe the collections contained within these DLs but we have given the URLs for those who want to find out more about them.) Except for **Wiki Web** (which some might argue is not a proper DL), the majority of current DLs are static, that is they contain mainly repositories of information that can be retrieved using the search and browse facilities, but the collections are built and maintained by specialists — typically modifying the collection may mean an interruption to user services. However, Wiki does not have any security to ensure the quality of its collection, and this really disqualifies it as anything better than a "departmental" or private library. Although **Stories from the Web** allows children to submit stories and write reviews, it does not permit full-text search.

Table 1. Digital libraries and their features

	Static		**Dynamic**		
DL	*Search*	*Browse*	*Annotate*	*Review*	*Create*
For academic purposes					
NZDL	Y	Y	N	N	N
NCSTRL	Y	Y	N	N	N
ETD	Y	Y	N	N	Y
CDL	Y	Y	N	N	N
For public					
BL	Y	N	N	N	N
LIC	Y	Y	Y	Y	Y
Wiki Web	Y	Y	Y	Y	Y
For commercial purposes					
IDEAL	Y	Y	N	N	N
ACM	Y	Y	N	N	N
For kids					
Stories from the Web	Y	Y	N	N	Y
Story Place	N	N	N	N	N

ACM: ACM Digital Library (http://www.acm.org/dl/)
BL: British Library (http://portico.bl.uk)
CDL: California Digital Library (http://cdlib.org)
ETD: Electronic Thesis and Dissertation (http://etd.vt.edu)
IDEAL: IDEAL On-Line (http://www.idealibrary.com)
LIC: Library of Congress (http://lcweb.loc.gov)
NCSTRL: Networked Computer Science Technical Report (http://cs-tr.cs.cornell.edu/)
NZDL: New Zealand Digital Library (http://www.nzdl.org)
Stories from the web: (http://www.storiesfromtheweb.com)
Story Place: (http://www.storyplace.org)
Wiki: (http://wiki.org/wiki.cgi?WikiWay)

1.2 Design Philosophy

In this section, we briefly revisit our previous work so that its methods and findings can provide a background for the body of this paper and the issues explored within it. The theoretical motivations, commitments and assumptions that have shaped the design and development of a children's DL prototype of stories and poems written by and for 11-14 year olds are also described.

Most contemporary DLs are not designed for children [6]. Using a concrete example to demonstrate our design philosophy and research approach, a DL of stories and poems for children aged 11 to 14 has been built. The work was carried out as part of a project funded by the UK Engineering and Physical Sciences Research Council (EPSRC) in collaboration with a secondary school, St. Albans School (UK).

From the start, we wanted our project to be a thoroughly collaborative endeavor as we wanted to design the DL with and for children. We invited a class of 23 boys and their English teacher to be our design partners. These children were selected because they were competent web users and would be able to give more informed comments on the efficiency and effectiveness of DLs, compared to say, novice users. (Also, one of this paper's authors is a Governor of the school.)

Two separate sessions were conducted during a 70-minute English lesson between November and December 1999 to carry out participatory design, engaging children as *design partners*. At the end of the second session, the children developed a list of requirements:

- DL should be like a "traditional" library providing efficient search facilities to retrieve relevant materials;
- DL should be more game-like;
- DL should offer opportunities to children to submit materials;
- DL should give recognition for good stories submitted by listing the top ten books/authors;
- DL should be fun to use; and
- DL should provide opportunities to chat with and get feedback from other readers.

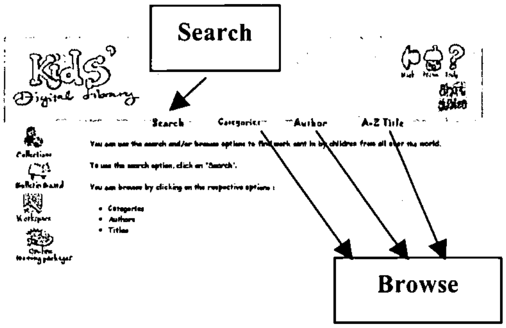
A third session was conducted in February 2000 to carry out *participatory evaluation*, engaging children as testers. The aim of the third session was to get a quick impression of the children's responses to the "look-and-feel" of three different interface designs, prototyped in the meantime. When we started with this project, we were uncertain as to the likes and dislikes of this age group (11–14 years old). From initial evaluation, their preferences were for fun and interest, as well as functional. Further details of the requirements gathering and initial evaluations can be found in [23, 24].

1.3 Design Choices

The children's DL was built using the Greenstone DL software, an open source system for the construction and presentation of information collections [26]. Our children's DL collection provides effective full-text searching and metadata-based browsing facilities, as offered by Greenstone. The collection is easily maintainable and rebuilt entirely automatically. Because special features are required, customised plug-ins have been developed.

Figure 1 shows the horizontal navigation bar, which contains the usual browse and search facilities provided by *static* DLs. Users can browse the DL by category: the stories and poems are classified according to twelve categories by author and title, in alphabetical order. The stories and poems are contributions from authorised children users, explained in more details when we describe the dynamic features provided in this children's DL.

Users can perform simple search by typing in the search terms as well as restricting the search space to specific collections.



You can use the search and/or browse options to find work sent in by children from all over the world.

To use the search option, click on 'Search'.

You can browse by clicking on the respective options :

- Categories
- Authors
- Titles

Figure 1. Static features of the children's DL.

2. DYNAMIC DLs

Although subject-based DLs are beginning to emerge on the Web, and promise opportunities we never had with traditional libraries or even the Web. DLs in general have not taken up in a "big" way compared to the Web (or, indeed, many other applications such as word processors or even e-books). One reason for the Web becoming popular almost overnight was the introduction of Mosaic, a graphical user interface which made it "very easy" for anyone to explore information. In contrast, because the majority of current DLs are mainly repositories of conventional-media information, users' experience in DLs is passive and less engaging compared to the Web. Furthermore, many DLs remove social exchange and interaction, focusing narrowly on the technical mechanisms of information access [1; 5].

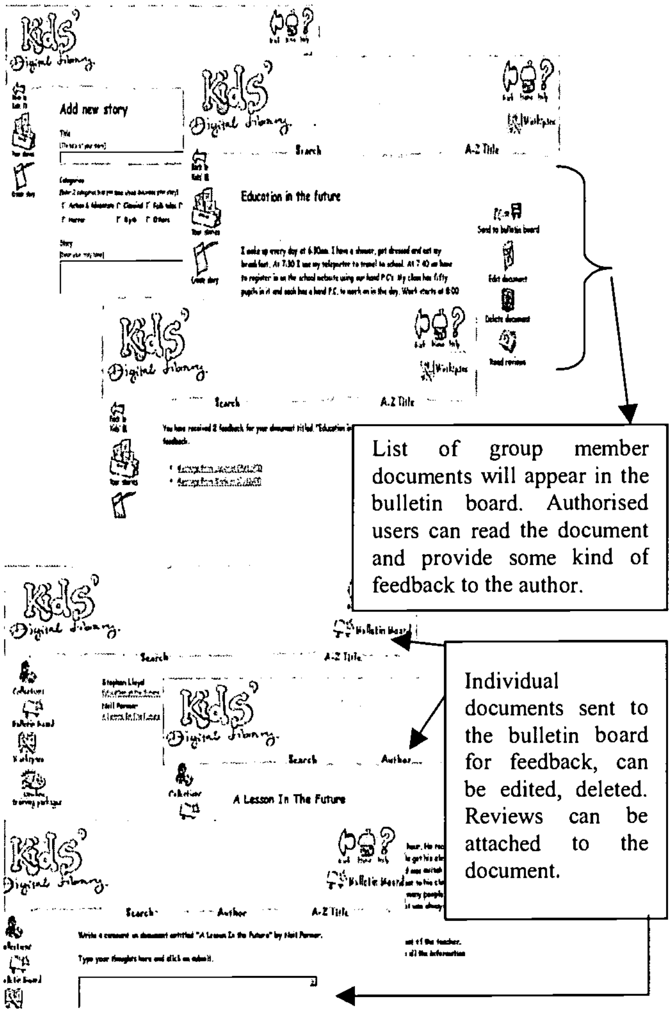
In an excellent review of the field of computer-supported cooperative work (CSCW) with respect to DLs, Nichols and Twidale [25] urge designers not to be captivated by new technologies but to learn from librarians who have been doing something analogous for years in a well-laid out library with carefully designed signage, access points to cataloguing and indexing sources using physical media, such as paper and index cards. They suggest that careful analysis of the design and evolution of these physical artifacts and conventional face-to-face collaborative interactions may be useful to inform the design of DLs. This knowledge *combined* with new technological opportunities presents many possibilities in supporting different kinds of information retrieval to support the usability, usefulness and acceptability of DLs.

2.1 Initial Design Choices

Work carried out to develop educational applications of DLs across all disciplines ranging from primary school through graduate school include [e.g. 3; 5; 11; etc.]. One of our main interests is how the use of DLs can promote collaborative writing among children.

Presently, we have implemented one kind of collaborative writing — collaborative review — allowing children to create and submit their own stories and poems to a workspace and permitting others such as their teachers and peers to read and give feedback by sending their comments via email to the children authors. Thus, in contrast to other work [e.g., 11; 16; etc.], a distinctive feature in our children's DL is the opportunity for children to *create* their own stories/poems and upload them into the bulletin board (see Figure 2) for reviews from their teachers and peers, before submitting to the permanent DL. Only material approved by the teachers can be submitted to the 'core' DL, thus ensuring the quality of the documents. To encourage *collaboration*, children can query and browse stories and poems written by other children. They can read stories, give reviews, read other children's reviews on stories and email authors for other comments.

The DL environment also provides a display of the top ten stories/poems; information about the authors, and a message board to post and discuss ideas.



List of group member documents will appear in the bulletin board. Authorised users can read the document and provide some kind of feedback to the author.

Individual documents sent to the bulletin board for feedback, can be edited, deleted. Reviews can be attached to the document.

Write a comment on document entitled "A Lesson In the Future" by Neil Parmer.

Type your thoughts here and click on submit.

Figure 2. Dynamic features implemented in children's DL.

2.2 Collaborative Writing Behavior

While children can be extremely honest in their feedback and comments, much of what they say needs to be interpreted carefully within the context of concrete experience [7]. While there is a great deal of variability subsumed within the practice of ethnographically-based studies, most practitioners agree that there should be a commitment to studying activities in the natural settings in which they occur, and to focus on what people actually do, not simply on their own accounts of behavior [15].

Observational study was used as a means of refinement of the initial design of the dynamic component of the DL. We hoped to gain insights from the observational study on how the DL collaborative writing environment should be conceptualized not just from what children said they wanted or what we thought they wanted, but on what they actually did. Our priorities were:

- To understand how children carry out collaborative writing in their natural classroom setting; and
- To draw insights from analyzing the children's natural writing behaviors, to refine the initial design of the collaborative writing environment.

Our study was inspired by Robertson's work on structuring the results of a field study in such a way that they might bridge, or reduce, the gap between the description of the work and the design of the technology to support the work [17]. Video recording of people working, talking about their work is viewed by Suchman and Trigg [21] as a valuable resource in later analysis and reflection. In our study, results of the video analysis will be used to understand the relation between collaborative writing and visual conduct to give us insights into the design of our DL that might support collaborative writing over distance. The analysis of the video recording is based on the taxonomy defined by Robertson [17], to connect backwards to our study the children's collaborative writing behaviors and forwards to the refinement of the initial design of the dynamic component of the children's DL.

2.2.1 Experimental protocol

Three sessions were conducted with twenty-three Year 2 (Class 99/00) boys at St. Albans School (UK). Their roles were that of design partners and testers. To carry out in-depth analysis of the design of the children's DL, we worked with a smaller group of six children, encouraged by established researchers in the design of children technology [9].

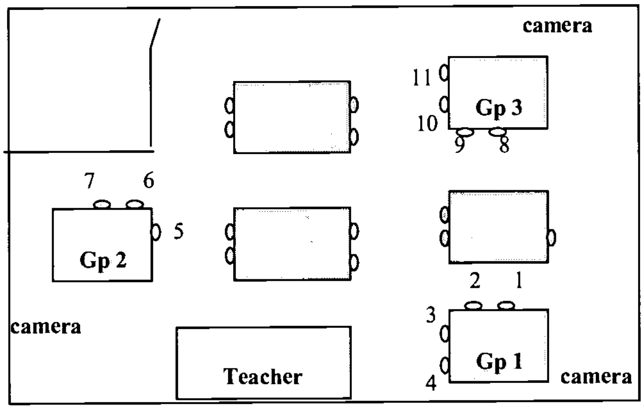
To evaluate the children's DL, we invited another class of twenty-four Year 2 (Class 00/01) boys at the same school. In contrast with the first batch of boys, these boys have not been introduced to the concept of DLs. The same English teacher is teaching this class.

The observational study was conducted during an English lesson in November 2000. The main objective was to observe behaviors in collaborative writing within the classroom. Results were used to verify whether what the first batch of boys (Class 99/00) wanted in the collaborative environment matched with the collaborative writing behaviors of the second batch of boys (Class 00/01).

Prior to the English lesson, Class 00/01 boys were asked to read the chapter "The Black rocks of Brittany" in the book *The Road to Canterbury* by Ian Serraillier. The session began with the teacher explaining what the task was: to discuss which character in the story was the most generous. Next, they were to write in their exercise books the reason(s) for their choice. At any time, the group could exchange drafts and comment on each other's writing. The class would convene for the last fifteen minutes for discussion.

Figure 4 shows the seating arrangement and the positions of the video cameras. The boys were divided and seated in groups of three or four, forming a total of six groups. (Two boys were absent during the video-taping session.) Video recordings were made on three groups working together throughout the 45-minute session.

Figure 5 shows a video segment of Group 1's activities during period 10-15 minutes (pre-writing). Because the teacher was called away from the class to attend to some urgent matters, the class was left unmanned for about 30 minutes during the periods of writing and reviewing. This explained the absence of the teacher's interactions with the groups during these periods.



Gp 1:	Gp 2:	Gp 3:
Boy 1 - Bhavin	Boy 5 - Neil	Boy 8 - Alex D.
Boy 2 - Lucas	Boy 6 - Nicky	Boy 9 - William
Boy 3 - Alex	Boy 7 - James F.	Boy 10 - James T.
Boy 4 - Ravi		Boy 11 - James E.

Figure 4. Seating arrangement and positions of video cameras

2.2.2 Results and analyses

Situated action, proposed by Suchman [20], is a term to "underscore the view that every course of action depends in essential ways upon its material and social circumstances (p. 50)". Suchman went on to explain that rather than attempting to abstract action away from its circumstances and represent it as a rational plan, the approach is to study how these people use their circumstances to achieve intelligent actions. Robertson [17] put a case for human embodiment as the fundamental consideration for designing systems that support people working over distance. She used "embodied action" to name the publicly available, purposeful and meaningful actions that people rely on to interact with others and their environment.

Using the taxonomy proposed by Robertson [17], we categorise embodied actions into group and individual activities. In the first viewing of the tapes, we were interested to simply observe what the groups were doing in general. The group activities were constituted by individual embodied actions. These actions define shared activities in a shared physical space. Robertson's taxonomy is modified to describe the embodied actions observed in the boys' collaborative writing behaviors and they include:

- *Conversing/discussing/arguing.* These are actions or activities that describe face-to-face interactions between the children within the groups. They can involve either maintaining a single conversation/discussion/argument involving the whole group or maintaining more than one conversation involving different individuals but within the same space.

- *Looking together for an answer (book).* This activity involves looking at the textbook from which the story is taken.
- *Focusing attention.* This action is generally initiated by an individual resulting in the group re-orienting its attention.
- *Breaking up/interrupting.* 'Breaking up' is an action generally initiated by an individual for the group to move into an individual activity such as writing.
- *Reforming.* This action is brought about by the teacher or an individual in the group to come together.
- *Questioning/clarifying.* These are actions that members of the group take advantage of other members' understanding and knowledge of the task at hand.
- *Doing something else/uninterested.* Individuals occasionally did something other than the group activity, while remaining in the same physical space. These individuals can get back to the group's activity by changing their spatial position and orientation.
- *Writing.* This activity is carried out by an individual.
- *Listening to teacher.* This action can either be initiated by the teacher or by individuals who want the teacher to clarify things.
- *Affirming/listening.* These activities are essential in encouraging whether members of the group are doing the task right.
- *Reading a book.* This action involves the individual reading the book by Ian Serrailler.

Tapes were viewed again from the perspective of what the individual boys were doing when they were carrying out embodied actions in relation to:

- *Teacher.* The actions involve individuals raising hands to attract the attention of the teacher as well as clarifying, discussing and listening to the teacher.
- *Other group members.* Activities involve exchanging and reading draft, giving and getting feedback. Other actions include pointing, shifting gaze and initiating change.
- *Class.* These actions include reading draft, affirming and giving feedback. They can also include moving round within the workspace. Sometimes these actions contribute to the current class activity.
- *Physical artefacts.* These include affordances of the physical environment, for example, paper, pen, book, table, etc.

A map of the embodied actions of the three groups throughout the 45-minute sessions at intermittent intervals is shown in Figure 6, which provides examples of how the embodied actions, defined in the taxonomy, were performed over time by the children both individually and as part of the group activities. The activity of the class is recorded in the top row. The other rows record the embodied actions of the three groups. Within each group, individual embodied actions are recorded. Each of the categories defined in the taxonomy is allocated a symbol. A single symbol means that the individual, whose actions are represented in that row, performed the action. (The duration of the action is not to scale by the length of the symbols used to represent them.) Some rows seem empty, because individuals are participating in the group or class activities.

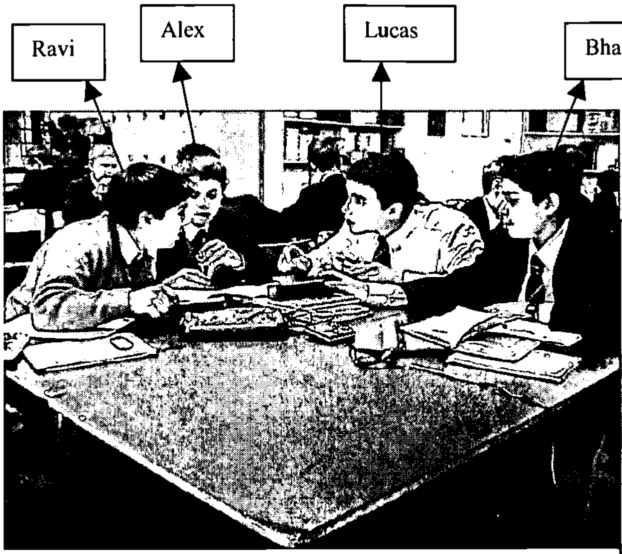
Figure 5, for example, shows a video segment of the boys in Group 1 during the period 10-15 min. Lucas and Ravi were engaged in a discussion, so are mapped to Figure 6, given a ☺ symbol. Bhavin was listening, indicated by ☼ . Alex seemed uninterested and he was playing with his pen, and this is indicated by ⧗ .

Consistent with Robertson's findings [17], the mappings show individuals perform a number of different actions during different stages (starting, pre-writing, writing, reviewing and reporting), as identified in the writing process. From the observations of the three groups' collaborative writing behaviors, the transition through the stages was not distinct. However, there are certain patterns defining these stages (see Figure 6): for example, in the first 5 minutes, almost all three groups were engaged in group activities such as conversing/discussing/arguing or attention/listening. Between 10–15 minutes, Groups 2 and 3 were engaged in writing. The boys in Group 1 were trying to settle down to do their work (in fact they were playing between 10-15 minutes instead of doing some kind of pre-writing). Writing became the main activity in all three groups between 20–25 minutes. Note that though the teacher was called away for the period between 20–35 minutes, the boys continued working on their drafts. The Group 2 boys, for example, seemed to be doing well, exchanging drafts, giving feedback to each other.

2.2.3 From observation to refinement

One of the major challenges confronting those who believe ethnography has something to offer system design is how to bring descriptions and analysis of work practices to bear on the design of new technologies [15]. Shapiro [19] says that ethnographers should embrace the problems of design, and try to link observations to design implications. Although some studies have been conducted to address this challenge [e.g., 21; etc.], the transition from ethnographic study to design remains complex and difficult. There is no simple relation between the findings of ethnographic study and design specifications [15].

How does one structure the results of an observational study in a way that might bridge the gap between the study of children's natural collaborative writing behaviors and the design of a dynamic collaborative environment within the children's DL? How could collaborative writing be done if it were to be done remotely, over a network? What would happen if the physical artefacts are replaced by the computer? Of course one needs to be aware that in remote collaboration, a shared workspace is not a shared physical space, but one made possible by the computer system and communication technology.



This video segment shows the boys in Group 1 during the period 10-15 min.

Lucas and Ravi were engaged in a discussion, so is given a ☺ Symbol. Bhavin was listening, indicated by ✲ .

Alex seemed uninterested and he was playing with his pen, and this is indicated by ⧗ .

Figure 5. Video segment of Group 1's activities during period: 10-15 minutes

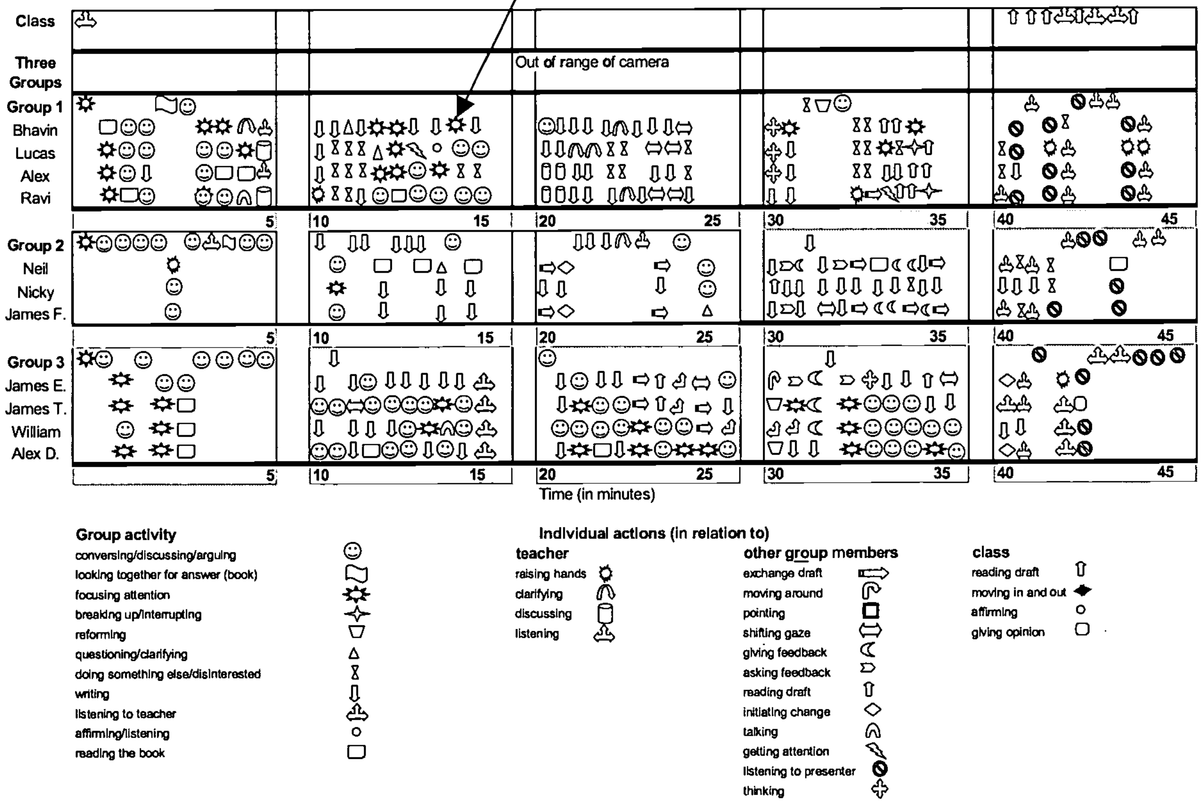


Figure 6. Overview of 45-minute session - children engaging in collaborative writing for periods: 0-5 min; 10-15 min; 20-25 min; 30-35 min and 40 -45 min

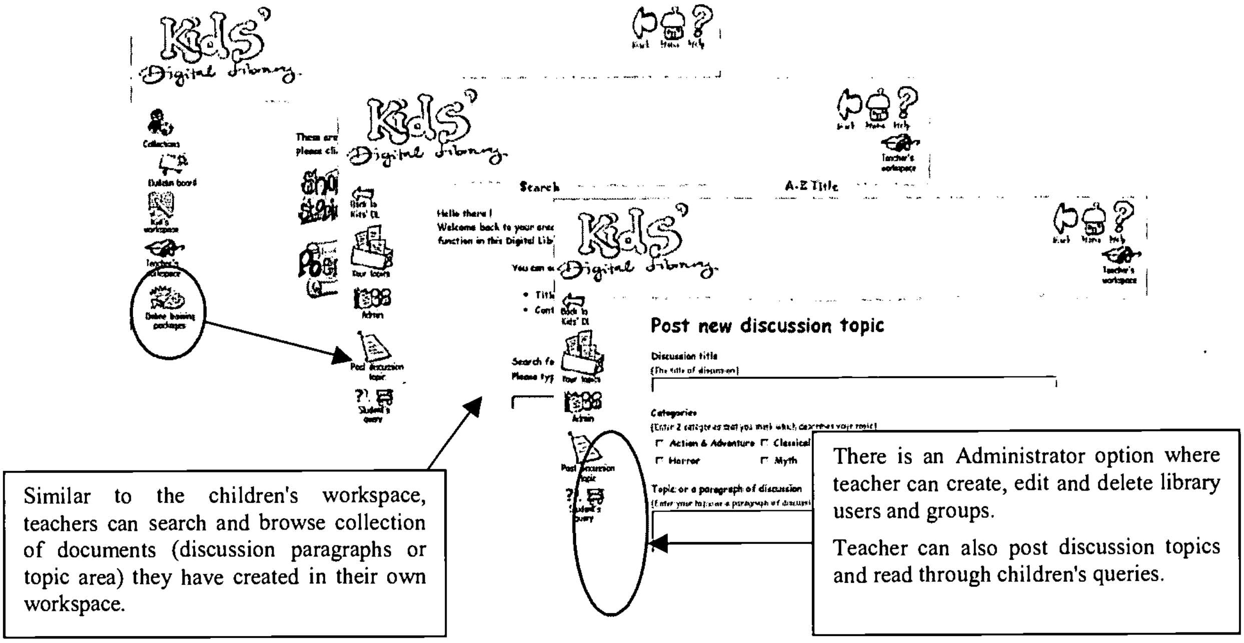
Table 2 maps initial requirements and results of observational study to affirmation/suggestion for improved design.

Column A indicates what Class 99/00 boys wanted in the DL; qualities desired in the DL and dynamic features to allow for user-initiated activities (e.g., submitting to the DL, etc.) and social environment to promote collaboration and feedback (e.g., chatting with friends, etc.). Column B bullet lists the writing behaviors of Class 00/01 boys observed in our observational study. The actions are put into four different stages in collaborative writing: starting and pre-writing; writing; reviewing and reporting. In each of these stages, we observed certain behaviors. Column C draws up a list of proposed new features suggested by Class 99/00 boys and reinforced in our observational study on Class 00/01 boys.

The classroom collaborative writing task (Column B) shows differences with the original expectations of the Class 99/00 boys (Column A). During the writing class, exploitation of some qualities Class 99/00 boys desired, for example, games and reading of other stories, was minimal in the classroom, paper-based environment. On the other hand, dynamic activities such as discussions were engaged more often as observed involving Class 00/01 boys, validating the suggested dynamic features in the children's DL. This profile of use, thus, leads to a significant emphasis of the dynamic features (Items 5-7, Column A) in the re-design of the children's DL (Column C).

Table 2. From initial requirements (see section 1.2) to results of observational study (see section 2.2) to refinement of initial design

Column A: What boys wanted	Column B: Observational study	Column C: Proposed new features (suggested in Column A, reinforced in Column B)	Column D: Comments
▪ **Qualities** 1. Be like "traditional" library 2. Be more game-like 3. Be efficient in searching for relevant materials 4. Offer children with fun features to search for relevant books, etc. ▪ **Dynamic features** 5. Offer children with opportunities to submit to the DL. 6. Give recognition of good stories submitted by listing the top 10 books/authors. 7. Provide opportunities to chat with and to get feedback from other readers.	▪ **Starting and pre-writing** a) Some kind of external stimulus (e.g., teacher) was required before students started working on their assignments. b) Reference was made to the book recommended by the teacher. c) Some boys tried to attract attention by using hands, fingers and voice. d) Some boys were unable to join in the discussion (possibly due to shyness). e) A couple of the boys looked confused. f) Lots of activities on Conversing/discussing/arguing with focusing/listening. ▪ **Writing** g) A couple of boys needed other boys to help them with spelling. h) Cancellation of the whole or bits of document by some boys. ▪ **Reviewing** i) Drafts were passed around. j) Some boys were talking to boys in other groups. ▪ **Reporting** k) Boys were eager to read drafts in front of class.	i. Provide spell checking *(3, g)*. ii. Have editing and deleting facilities *(5, h)*. iii. Features where the users can contribute. Ranking feature where users can contribute by reading and providing some kind of feedback to the document *(5, k)*. iv. Have personal writing space *(5, j)*. v. Flags to show that there are new reviews in the bulletin board or new essays in the DL *(6, i)*. vi. Post drafts in bulletin board *(6, k)*. vii. Read related essays written by other students not belonged to assigned groups *(7, f)*. viii. Links to related topic of discussion *(7, b)*. ix. Facility to post questions to teacher *(7, a/c/d/e)*. x. Ability for group members to contribute to the review session *(7, j)*. xi. Teacher area and teacher-guided discussion *(7, a)*.	Not yet implemented. Yes, already implemented. Not yet implemented. Yes, already implemented. No, not in initial design (see Figure 7 for new feature added in the improved version).



Similar to the children's workspace, teachers can search and browse collection of documents (discussion paragraphs or topic area) they have created in their own workspace.

There is an Administrator option where teacher can create, edit and delete library users and groups.

Teacher can also post discussion topics and read through children's queries.

Figure 7. New feature - teacher workspace

What Class 99/00 boys wanted are generally reinforced by the behaviors of Class 00/01 boys. The following are features we have already implemented in our initial DL prototype:

- Links to related topic of discussion;
- Flags to show new reviews posted in the bulletin board;
- Features to contribute essays, rank essays, give reviews and give feedback;
- Bulletin board in which drafts, questions and answers can be posted;
- Features to edit and delete; and
- Personal work space for writing;

From the observational study, we identified three new features when observing the natural collaborative writing behaviors of Class 00/01 boys:

- *Teacher area and teacher-guided discussion.* The teacher was the "stimulus" to get the boys to start thinking and writing. The discussion at the end was helpful in rounding up the activity. Throughout the lesson (when the teacher was in class), individuals and groups would be asking the teacher questions, clarifying task, etc. We have now implemented a teacher area where the students can pose questions and engage in discussions initiated by the teacher (see Figure 7).
- *Ability to select own friends.* Some of the children preferred to work with their own group of friends and not with those assigned by the teacher. A couple of boys walked across the room to look for friends to read their writing and give them feedback. At the moment, the children's DL has not included the facility to allow the children to add their circle of friends whom they want to get feedback from. We are in the process of incorporating this feature.
- *Spell checker.* This feature was also identified when we asked Class 99/00 boys, and this was reinforced by a couple of boys needing help with spelling. We are also in the process of implementing this feature.

DISCUSSION

The goal of integrity is that a DL represents the collection that "should" be there. In dynamic library, especially one with children as authors, one wants to encourage creativity, diversity and the unexpected: but this should be aimed at the content rather than the integrity of the system itself!

Our project highlights these issues summarized as follows:

- Conventional reliability, integrity and security issues. The software the DL is built on must be reliable, it must provide an adequate security system that the teacher can handle, and which the students cannot easily circumvent. All this is possible with NT, Macintosh and Unix based systems, and these can be set up to guarantee that students do not mix up their submissions to the DL. Doing better than proprietary systems would be a research project in its own right!
- If students can submit materials that contain active components, these components may be accidentally or maliciously damaging. Even Microsoft Word documents can contain executable code, and the effect of other users reading it can initiate virus infections. It is therefore desirable to restrict dynamic content to basic HTML or to other easily-restricted formats.
- Systems with very little security can be surprisingly successful. In a school environment, mutual trust might easily be built up *within* a classroom, and this could be supported by the DL having restricted physical/firewall access or permitting access from designated ports or IP

numbers (i.e., particular machines rather than particular users). However in our project we found that the school's firewall (which is designed to stop external security breaches) made it impossible for us to maintain the DL server, which was located physically inside the school premises!

- Examples of successful open systems include Wiki (`http://wiki.org/wiki.cgi?WikiWay`) and early versions of Unix. Wiki [10] can be understood as an open dynamic DL with absolutely no restrictions on what users can do to content. (It provides automatic indexing and cross-referencing of content, though typically delayed by a few days.) Users can change or create any content, whether or not they authored it, and Wiki distinctively makes it extraordinarily easy (and, indeed, tempting) to do this. The result is that constructive social conventions emerge, and users rewrite and edit content to make it better. Wiki systems are typically strongly subject-based; a generic Wiki system (e.g., the equivalent of a public rather than a research library) would, in our opinion, be unlikely to succeed. Wiki systems achieve their success by having a distinctive *brand*: just like substantial real libraries which, without trying, instill a sense of awe or peace in their users! The Wiki brand is enforced by a distinctive markup language (equivalent to a very small subset of HTML) which might be said to restrict user's freedom of expression, and hence encourages conformity.
- In the early (1970s) days of Unix, Queen Mary College (QMC, London) and Melbourne (Australia) had different philosophies of student access. QMC had open access to source code: the result was that students found system bugs and helped staff fix them. In Melbourne, a stricter system was in force, with source code off-line, and students were implicitly seen as a threat to security. As a result of the lack of cooperation, when student problems arose, they were quite serious. In comparison, QMC had no student problems, and in fact found them helpful.
- The moral is that if it is technically possible to do so, students should be actively involved in all aspects of the DL, and encouraged to take responsibility for it. Obviously, our experimental approach used the students as co-workers with us in developing the user interface and other features. This gave the students a sense of self-worth and, crucially, of investment in the system itself, and hence made the DL a positive experience for all involved. It is possible that future dynamic children's DLs which "big bang" with a working system will be susceptible to different, possibly destructive, attitudes from students.

Some constraints of the implementation technologies available affected the facilities that could be provided.

The movement of a document between collections means that the source and destination collections both need to be re-indexed for searching. For very small collections, this takes a matter of seconds. However, even modestly sized collections can start to take more substantial amounts of time with some algorithms. As with almost all algorithms, there is a trade-off between pre-processing and run-time costs. The MG search engine behind Greenstone provides a high-quality level of compression and fast recall times, but at the cost of processing time when indexes are rebuilt. Similarly, MG does not provide incremental indexing which would substantially reduce the time cost of rebuilding the index.

Moving a single document between two small indexes would probably not itself cause a problem. However, the pattern of use within a school environment is for a high density of use within a small period (less than one hour), so many indexes can be rebuilt at once, alongside high densities of saving and editing activity.

Most of the user actions which would lead to re-indexing are focussed during class time, concurrent with editing and other creative tasks. As seen from our observational study, much of the review and other user-to-user interactions that occur within the classroom happen outside the core DL system. In order to provide good response times to the users during this peak time of use, we have scheduled the rebuilding of indexes to follow each class in a predictable manner. Texts then appear in a timely manner to support later review outside of the class.

The DL has provided an active environment where a user (in this case, a reader) can participate in the environment by giving feedback to the author. In our previous design, we have included authors' email addresses for readers to contact. From our evaluations, the children testers were not in favour of having their email addresses listed to avoid them from getting "junk" mails. For security reasons, we were advised by ethic experts not to display personal information mainly full names and addresses together with their photos. However, in order to provide an active environment and feedback to the authors, we allow the reader to send feedback using a rating system. The rating system was created by adopting our own DocMan [4] tool with a minimal degree of alteration (to be able to access certain metadata in the Greenstone system).

3. CONCLUSIONS & ON-GOING WORK

Ethnographically-based design projects are still few in number and primarily exploratory in nature. They are just beginning to provide concrete examples of the value of bringing knowledge about specific work practices to the designed artifacts, and of the requirements for creating an environment wherein the worlds of design and work analysis can come together [15].

We structured the results of our observational study to bridge the gap between the study of children's natural collaborative writing behaviors and the design of a dynamic collaborative environment within the children's DL, as if it were to be done over distance.

This is on-going work for us. The initial work has created a useful DL for children, which has novel features *with a rationale for those features*. Certainly, more can be done: careful analysis of data; refinement of taxonomy of embodied actions and greater understanding of how actions can be interpreted to support design. If we had had a mixed school, and the gender issues were really relevant, we would have to have done more complex structured experiments, with controls and what not. Instead, we have been clear we used males, and therefore raised sharp questions that others working with females or mixed groups might like to explore as specifically gender issues, rather than DL issues. One might make similar comments about the age range, the income group, and so forth. These are all large and relevant issues for the success of DLs in the world. What matters — in our view — is that we can create a useful DL for a well-defined part of the real population of users.

The pilot work suggests many exciting avenues to research in greater depth. It will be interesting to repeat work with other age groups and control for other factors such as web skills and gender. We will be carrying out longer-term observational studies to study the impact DLs have on collaborative writing.

4. ACKNOWLEDGEMENTS

We would like to thank the teachers and children at St. Albans School for their participation and ideas. David Bainbridge, University of Waikato, assisted in the customisation of the Greenstone software. Colleagues at Middlesex University, Ann Blandford and Hanna Stelmaszewska, helped us to video-tape the children at work. This work was carried out as part of EPSRC grant GR/M72098.

5. REFERENCES

1. Ackerman, M.: Providing Social Interaction in the Digital Library, in Proceedings of Digital Libraries'94: First Annual Conference on the Theory and Practice of Digital Libraries, pp198–200, ACM Press (1994).
2. Ackerman, M. and Fielding, R.: Collection Maintenance in the Digital Library, in Proceedings of Digital Libraries '95, ACM Press (1995).
3. Borgman, C., Gilliland-Swetland, A., Leazer, G., Mayer, R., Gwynn, D., Gazan, R. and Mautone, P.: Evaluating Digital Libraries for Teaching and Learning in Undergraduate Eduaction: Case Study of the Alexandria Digital Earth Prototype (ADEPT), http://is.gesis.ucla.edu/adept/pubs/lt822.html
4. Buchanan, G., Marsden, G. and Thimbleby, H.: Dynamic Metadata for Monitoring Digital Library Management, in Proceedings of Digital Libraries'99, pp219–220, ACM (1999).
5. Crabtree, A., Twidale, M., O'Brien, J. and Nichols, D.: Talking in the Library: Implications for the Design of Digital Libraries, in Proceedings of Proceedings of Digital Libraries' 97, pp221–228, ACM Press (1997).
6. Druin, A.: DLs for children: Computational tools to support Children as researchers. http://www.cs.umd.edu/hcil/kiddiglib/ (2000).
7. Druin, A., Bederson, B., Boltman, A., Miura, A., Knotts-Callahan, D. and Platt, M.: Children as technology design partners, in Druin, A. (ed.), The Design of Children's Technology. Morgan Kaufman. pp51–72. (1999).
8. Fox, E.A., Akscyn, R.M, Furuta, R.K and Leggett, J.J, Digital Libraries, in Comm. Of the ACM, Apr '95, Vol.38 No.4, pp23-28.
9. Kafai, Y.: Children as designers, testers and evaluators of educational software. in Druin, A. (ed.), The Design of Children's Technology. Morgan Kaufman. pp123-145. (1999).
10. Leuf, B. and Cunningham, W.: The Wiki Way. Addison Wesley Longman (2001, in press).
11. Koenemann, J., Carroll, J., Shaffer, C., Rosson, M. and Abrams, M.: Designing collaborative applications for classroom use: The LiNC Project, in Druin, A. (ed.), The Design of Children's Technology. Morgan Kaufman. pp99-122. (1999).
12. Levy, D. and Marshall, C.: What Color was George Washington's White Horse? A Look at the Assumptions Underlying Digital Libraries, in Proceedings of Digital Libraries'94, pp163–169, ACM Press (1994).
13. McMillan, G. Digital Libraries support Distributed Education or put the library in Digital Library, 9th National Conference of Association of College and Research Libraries, April, 1999
14. Miksa, F. and Doty, P.: Intellectual Realities and the Digital Library, in Proceedings of Digital Libraries'94, pp1–5, ACM Press (1994).
15. Monk, A. and Gilbert, N.: Perspectives on HCI: Diverse Approaches. Computer and People Series. Academic Press (1995).
16. Nichols, D., Pemberton, D., Dalhoumi, S., Larouk, O., Belisle, C. and Twidale, M.: DEBORA: Developing an Interface to Support Collaboration in a Digital Library, in proceedings of Fourth European Conference ECDL'00, Borbinha, B. and Baker, T. (Eds.), pp239–248, Springer (2000).
17. Robertson, T.: Building bridges: negotiating the gap between work practice and technology design. International Journal of Human-Computer Studies, 53, pp121–146 (2000).
18. Rowland, F. The Librarian's Role in the Electronic Information Environment, ICSU Press Workshop'98, http://www.bodley.ox.ac.uk/icsu/rowlandppr.htm.
19. Shapiro, D.: The limits of ethnography: Combining social sciences for CSCW, in Proceedings of CSCW'94, ACM (1994).
20. Suchman, L.: Plans and situated actions: The problem of human-machine communication. Cambridge University Press (1987).
21. Suchman, L. and Trigg, F.: Understanding practice: Using video as a medium for reflection and design, in Design at work: Cooperative design of computer systems, Greenbaum, J. and Kyng, M. (Eds.), pp65–89, Lawrence Erlbaum Associates (1991).
22. Suleman, H., Fox, E. and Abrams, M.: Building Quality into a Digital Library, in Proceedings of Digital Libraries'00, pp228–229, ACM Press (2000).
23. Theng, Y.L., Mohd-Nasir, N., Buchanan, G., Fields, B. and Thimbleby, H.: Children as Design Partners and Testers for a Children's Digital Library, in proceedings of Fourth European Conference ECDL'00, Borbinha, B. and Baker, T. (Eds.), pp249–258, Springer (2000).
24. Theng, Y.L., Mohd-Nasir, N., Thimbleby, H., Buchanan, G. and Jones, M.: Designing a children's digital library with and for children, in Proceedings of Digital Libraries'00, pp266-267, ACM Press (2000).
25. Twidale, M. and Nichols, D.: A Survey of Applications of CSCW for Digital Libraries. Technical Report CSEG/4/98, Computing Department, Lancaster University (1998).
26. Witten, I., McNab, R., Boddie, S. and Bainbridge, D.: Greenstone: A Comprehensive Open-Source Digital Software System, in Proceedings of Digital Libraries'00, pp113–121, ACM Press (2000).

Looking at Digital Library Usability from a Reuse Perspective

Tamara Sumner and Melissa Dawe
Center for LifeLong Learning and Design
Dept. of Computer Science and the Institute of Cognitive Science
University of Colorado at Boulder
+1 303 492 2233
[sumner, meliss]@colorado.edu

ABSTRACT

The need for information systems to support the dissemination and reuse of educational resources has sparked a number of large-scale digital library efforts. This article describes usability findings from one such project – the Digital Library for Earth System Education (DLESE) – focusing on its role in the process of educational resource reuse. Drawing upon a reuse model developed in the domain of software engineering, the reuse cycle is broken down into five stages: formulation of a reuse intention, location, comprehension, modification, and sharing. Using this model to analyze user studies in the DLESE project, several implications for library system design and library outreach activities are highlighted. One finding is that resource reuse occurs at different stages in the educational design process, and each stage imposes different and possibly conflicting requirements on digital library design. Another finding is that reuse is a distributed process across several artifacts, both within and outside of the library itself. In order for reuse to be successful, a usability line cannot be drawn at the library boundary, but instead must encompass both the library system and the educational resources themselves.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: *Systems issues*, *User issues*; J.2 [**Physical Sciences and Engineering**]: *Earth and atmospheric sciences*.

General Terms

Design, Human Factors

Keywords

Location, Comprehension, Modification, Sharing, Reuse, Educational Resources, Digital Libraries, Learning Impact

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

1. INTRODUCTION

Science educators have repeatedly called for information systems that can effectively deliver quality educational materials in formats that are readily accessible, with a high degree of confidence that the materials will be useful, interesting, and effective [27]. This has largely been interpreted as a call for digital library systems or information systems with library-like services [16]. Towards this end, there are many digital library efforts underway aimed at improving the quality of undergraduate or K-12 science education [7, 23, 24]. Typically, these systems are similar to web portals, providing access to a managed collection of community-constructed educational resources with services for resource discovery, and possibly peer-review and resource creation as well. One prominent example of such a system is the NSDL (National Science, Mathematics, Engineering, and Technology Education Digital Library), a program initiated by the National Science Foundation to integrate multiple digital libraries in these areas and provide access to a broad variety of educational materials [36]. These efforts are based on the assumptions that providing digital libraries of educational resources can [21]:

- Improve the quality of education by promoting the reuse of educational resources that are proven to be effective;
- Improve the productivity of faculty through resource reuse and sharing;
- Help foster an active community of learning and innovation, where best practices and resources are developed and shared.

The underlying belief is that the *reuse* of existing resources or knowledge to create new educational resources will lead to improvements in both product (better educational resources) and process (teaching and learning). These assumptions and beliefs mirror similar goals for shared resource collections, henceforth referred to as 'repositories,' in both the business world and the software engineering community. In the business world, much recent activity has focused on using shared repositories to support 'knowledge management' and 'best practice' sharing to improve organizational efficiency and foster innovation [1, 17, 35]. The software engineering community has been vigorously promoting code sharing and software reuse for years since reuse is linked to measurable improvements in both programmer productivity and software quality [8, 10, 30, 31].

In the field of software engineering, reuse has been defined as the utilization of pre-existing components to ultimately create something new [10]. We feel that this is an appropriate definition for reuse of educational resources as well. For software developers, the product created might be an application; for educators it is a class or a curriculum. Although the products differ, we posit that the process of reuse in these two domains is essentially the same. Thus we feel that important lessons can be drawn from examining software engineering models of reuse: in particular, what promotes and what inhibits successful reuse.

Currently there is little research on the effects of reuse on learning. In literature from the library community [25], it has been observed that digital libraries can affect learning in different ways: through direct student interaction with the digital library system, and through the utilization of digital library resources by educators in their classroom. Studies show a positive correlation between learning and student interaction with a digital library [25]. Less is known about the relationship between educator reuse and learning. However, it is important to note that reuse is not an activity new to educators. On the contrary, the reuse process is deeply embedded in the way educators work today, as they locate, utilize and share physical as well as digital resources. In theory, digital libraries should support these current practices and extend them further, by making available a richer variety of educational resources, at all levels of granularity: ranging from single images, lesson plans and applets, to lab modules, to entire courses. Ideally, such facilities could enable educators to have more creativity and control of the materials they use in the classroom.

Past experiences in software engineering suggest that while a library is essential for supporting systematic reuse at the institutional level, simply providing a library of resources is insufficient to guarantee that effective reuse will occur at the individual level:

> *..., although the library metaphor has guided early work in classification, storage systems, and other areas of reuse technology, it does not provide the best focus for setting up and running a reuse program. [...], it has simply not yielded a major change in the way most people develop software.* Poulin, page 1, [30].

One survey on the state-of-the-practice in reuse found that (1) taking the library approach to software reuse and (2) the effectiveness or efficiency of the software library were two of the most significant *non*-predictors of reuse capability in an organization [31]. These results are quite disturbing given the current emphasis on 'digital libraries' as the primary means for facilitating the reuse of educational resources, and ultimately for changing the way that educators develop teaching resources.

In this article, we examine existing studies and theories on resource reuse from the software engineering literature. We believe that this prior research has much to offer digital library projects that are deeply concerned with fostering reuse within their communities of users. In particular, we develop and discuss a cognitive model of resource reuse derived from the software engineering literature – the location, comprehension, modification, and sharing cycle – and compare this model with existing digital library information lifecycle models. We focus mainly on comprehension and modification processes, as they are critical yet relatively understudied aspects of digital library use. We use this model to critically examine our own experiences with the Digital Library for Earth System Education (DLESE) project. This reformulation of the problem from 'providing a library' to 'supporting reuse' has significant implications for system design, community outreach and training, and future research directions, which we will discuss. We will first briefly present this model of reuse before proceeding to the analysis.

2. A MODEL OF REUSE

Reuse is cognitive activity that is embedded in an overall task-directed design process; as described above, the person's primary goal is to create something new. For instance, a faculty member may want to design a new lecture or even a new course, and may consider combining and reusing existing resources (e.g., lesson plans, exercises, maps, data sets) as part of this overall process. Reuse involves both composing new resources from existing resources and developing resources that can be reused in the future [9]. Previous studies of software programmers [8] suggest that when composing new resources from existing resources, reuse involves three closely intertwined cognitive activities: location, comprehension and modification (Figure 1).

As shown in Figure 1, the first step in the reuse process is forming a 'reuse intention'; i.e., deciding to use the resource repository in the first place. Ye et. al. outline numerous ways the reuse process can fail at this first step [38]. Software developers can be reluctant to disrupt their workflow and endure the cognitive effort it takes to switch back and forth between their development environment and the repository. Developers also chronically underestimate the amount of time it takes to develop from scratch, which leads to a cognitive bias against deciding to reuse in the first place. Ye argues that these behaviors stem from loss aversion tendencies in human decision-making processes; i.e., people are more sensitive to avoiding potential losses (like wasting time looking for resources that don't exist) than to realizing potential gains (finding a useful resource).

Location refers to the process of finding potentially useful resources in the resource repository. The user must be able to translate his or her reuse intentions (their situation model) into appropriate system queries (the system model) [12]. Numerous studies indicate the many people, not just software developers, have difficulties with this; Norman referred to this challenge as the gulf of execution and evaluation [26].

In reuse, location is tightly intertwined with comprehension. Comprehension involves not only making relevance judgments, but also understanding the function, structure, and context of use of the resource in order to decide if it needs modification and how to go about making the necessary modifications. Unsurprisingly, research has found that people are often unwilling to take the time to thoroughly comprehend a resource, and instead employ "comprehension avoidance" strategies [9]. For example, a developer might execute a piece of software to see what it does, rather than read all its documentation. Often times, developers just abandon the reuse process at this point.

Typically, resources cannot be used 'as is' but instead must be modified before they can be used. Modification can take many forms. Studies of the evolution of resources in software repositories show that software components evolve through four main processes: refinement, composition, abstraction, and factorization [9]. These processes will be defined and discussed in the following section. However, it is sufficient to note that many people drop out of the reuse cycle here because they are either unable to make the necessary changes (because they lack the technical skills or an adequate understanding of the resource), or are unable or unwilling to spend the necessary time.

In this context, 'sharing' refers to the resource producer making a conscious decision to make a new resource available to others to reuse and undergoing the work necessary to do so. Sharing resources is a crucial part of an effective and sustainable reuse cycle, but it is often viewed as being an extra step that is outside of the overall design process. Bannon and Bodker discuss the extra work it takes to place items into a shared resource repository [3]. They argue that both the resource producer and the resource consumer must make a conscious effort to understand each other's context of use and that reusability is enhanced when the producer can anticipate the consumer's context of use to some degree and makes the effort to 'package' aspects of this context with the resource. In software engineering, studies show that this packaging for reuse takes a significant effort and can double the production costs of more complex resources [10]. Many organizations have found it necessary to implement reward programs, often involving direct monetary compensation, to encourage developers to do the extra work required to submit resources to the repository [30].

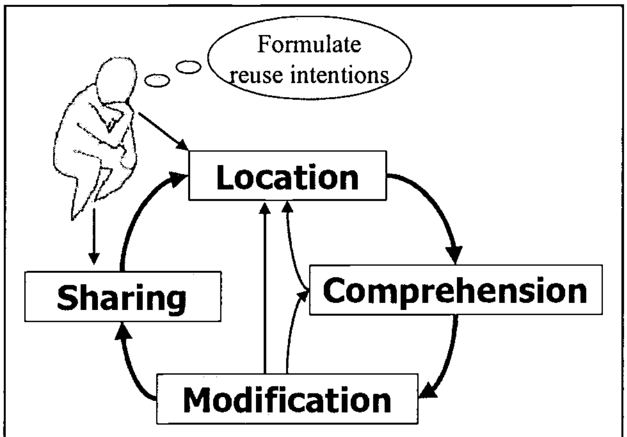


Figure 1: Location-Comprehension-Modification-Sharing Cycle (L-C-M-S)

In the following section, we will use this model to examine some of our experiences in the DLESE project, particularly noting the similarities and differences in the observed reuse behaviors of undergraduate teaching faculty and software developers.

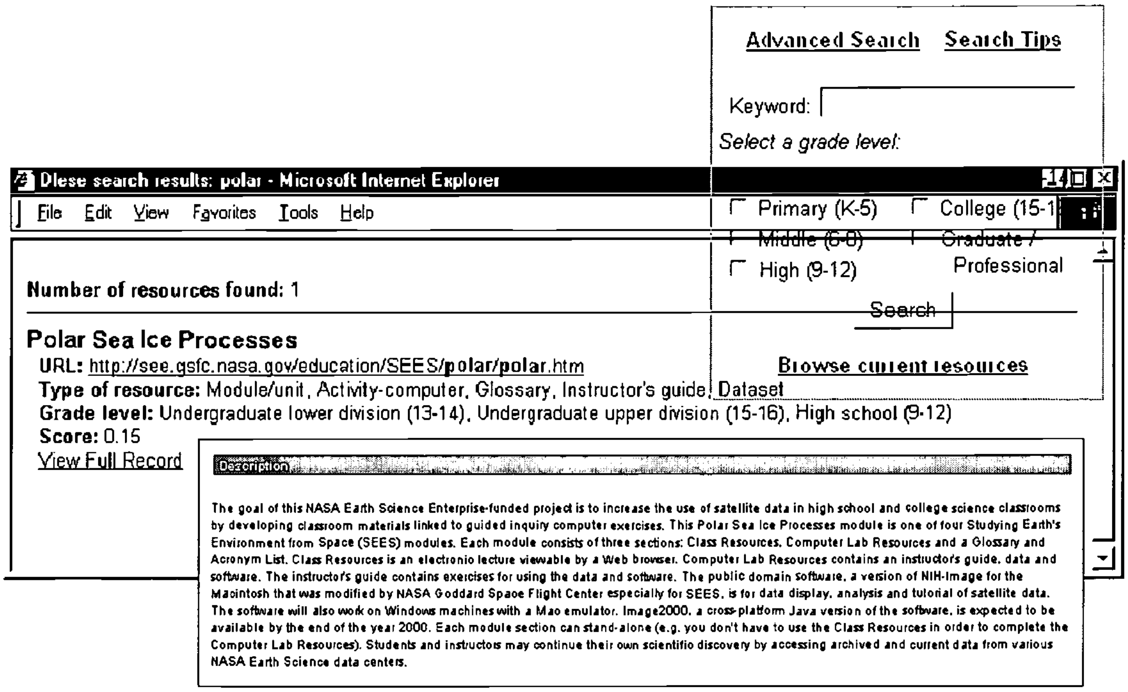
3. CASE: DIGITAL LIBRARY FOR EARTH SYSTEM EDUCATION

DLESE is a grassroots, community-led project to provide searchable access to high-quality, online educational resources for K-12 and undergraduate earth system science education. These educational resources include objects such as maps, simulations, lesson plans, lab exercises, data sets, virtual field trips, etc. These resources are created by either individual faculty members or by institutions (e.g., NASA, USGS) and are held (stored) on their local servers. When resources are contributed to the library, either the resource creators or DLESE catalogers create metadata describing the resource to support resource discovery. The metadata scheme, based on the IMS LOM standard [14], is quite detailed and is fully described elsewhere [11]. This metadata is centrally stored and managed and is, in effect, the chief holding of the library.

As detailed in the DLESE Community Plan [19], the success of DLESE will ultimately be measured in two ways. Firstly, it will be measured by the library's impact on earth system science teaching and learning. Is student learning improved? Do faculty teach more effectively or efficiently? Both of these outcomes depend on faculty effectively reusing the resources available in DLESE in their own teaching. Secondly, success will be measured by the sustainability and viability of the community contribution process. Do faculty create reusable resources to share with others? Can they create the necessary metadata to help their colleagues discover and use their resources? Thus, success in DLESE heavily depends on faculty participating deeply in a culture of reuse in terms of both resource design and teaching practices.

For the past 18 months, the project has focused on building and evaluating a library prototype, designing a resource review process, and setting up the community governance structure [23]. Part of the current library interface is shown in Figure 2. Faculty can search by keyword and grade level to locate resources. The search results are presented as a list of information summarizing the potentially relevant resources such as resource title, brief description, grade level, etc. Clicking on either the resource title or URL will bring up the educational resource directly in a new browser window.

During this period, a number of formative studies were conducted to help us understand community needs and to evaluate the evolving library prototype. As part of the requirements analysis process, workplace interviews were conducted with seven earth science faculty and two students. Participants were selected that were already using digital resources in their teaching and learning. Each interview lasted between 60 to 90 minutes and was taped, transcribed, and analyzed. The focus of the interviews was to understand in detail how participants located and selected digital educational resources. We found that there is an important distinction in the way instructors prepare for a course, and for a class, which suggest that these two processes must be treated differently (Table1).



Number of resources found: 1

Polar Sea Ice Processes

URL: http://see.gsfc.nasa.gov/education/SEES/polar/polar.htm
Type of resource: Module/unit, Activity-computer, Glossary, Instructor's guide, Dataset
Grade level: Undergraduate lower division (13-14), Undergraduate upper division (15-16), High school (9-12)
Score: 0.15
View Full Record

Description

The goal of this NASA Earth Science Enterprise-funded project is to increase the use of satellite data in high school and college science classrooms by developing classroom materials linked to guided inquiry computer exercises. This Polar Sea Ice Processes module is one of four Studying Earth's Environment from Space (SEES) modules. Each module consists of three sections: Class Resources, Computer Lab Resources and a Glossary and Acronym List. Class Resources is an electronio lecture viewable by a Web browser. Computer Lab Resources contains an instructor's guide, data and software. The instructor's guide contains exercises for using the data and software. The public domain software, a version of NIH-Image for the Macintosh that was modified by NASA Goddard Spaoe Flight Center especially for SEES, is for data display, analysis and tutorial of satellite data. The software will also work on Windows machines with a Mao emulator. Image2000, a cross-platform Java version of the software, is expected to be available by the end of the year 2000. Each module section can stand-alone (e.g. you don't have to use the Class Resources in order to complete the Computer Lab Resources). Students and instructors may continue their own scientifio discovery by accessing archived and current data from various NASA Earth Science data centers.

Figure 2. Discovery system interface in DLESE. The search interface is in the upper left corner. The search results page is shown in the middle. Clicking on 'View Full Record' brings up the full metadata record, which includes a short description of the resource.

When preparing for a class, educators are searching for items that will immediately plug in to the existing framework of their curriculum. In this task, reuse occurs *downstream* in the design process, where most of the overall design has been done, and so the search is highly targeted and specific according to the existing design. At this point educators are usually working under a severe time constraint, and don't have time to digest large amounts of information or adapt items for use. Thus they are looking for resources that are as *context independent* as possible, so that they require little adaptation effort. Such items include images, diagrams, vocabulary lists, and other resources that have minimal environmental or other dependencies.

Table 1. Two example DLESE tasks.

Prepare for a Class	Prepare for a Course
Jeff has an hour and a half before his class on environmental issues of greenhouse gases and ozone depletion. He wants to spice up his lecture. He browses online repositories he knows for pictures, charts, animations, or interactive tools to use during the class. He wants to conduct a pre-screening of sites for suitability and note the locations of these materials for his students, so they can study it further. The materials he uses must be from a source he trusts and be useable without a lot of alteration, since there is little time.	Kim is an introductory earth science teacher. She wants to teach her students about deep time, and how climate has varied over time in one location. She knows this is a difficult concept. She is looking for resources that can educate her in the area, as well as some pedagogical tools. In particular, she needs some help teaching students how to understand and get past their problems with "deep time." She would like to locate tutorials in this area, and experts in the field who she might be able to contact for teaching ideas.

This type of preparation is different from the work an educator goes through to prepare for a course. In this task, reuse plays a role *upstream* in the design process, in that much of the structure of the course itself is in a formative stage. Preparing for a course is generally on the time scale of weeks or months, rather than minutes or hours. In this process, instructors appear to be more willing to explore new resources and test them for their classroom. They may have a vague idea of what they are looking for, and are interested in browsing collections to gain ideas. They can afford to take more time to synthesize larger resources and adapt them to their classroom, or design curriculum around them.

One of the challenges of designing a digital library is to support these two processes simultaneously, which can create tensions and conflicts in the system requirements. To analyze how well the evolving DLESE prototype supports these two key user tasks, we conducted two rounds of usability testing with earth science faculty (one with five participants and one with ten). Each participant was asked to think aloud as he or she performed a series of class and course preparation tasks using the current DLESE prototype. Detailed observational notes were independently taken by two observers and compared after each session. Additionally, we analyzed the structure and content of some of the resources with an eye towards developing heuristics for creating metadata that best facilitates location and resource comprehension. We now present the results of these studies and analysis in terms of the L-C-M-S reuse cycle. Table 2 summarizes the results for each of the two types of tasks.

3.1 Formulation of reuse intentions

Our research has shown that faculty formulate reuse intentions, as exemplified in the above reuse tasks taken directly from user interviews. We discovered that their primary source for identifying materials for reuse is through personal interaction of some kind: they ask colleagues for suggestions, they exchange ideas and resources at workshops and conferences, they even go to the library and ask the librarian for known sources. A crucial aspect of this process is the reliance on a trusted source for resource location and evaluation. Faculty appear to use the source (either the individual creator, or an institution) of the material as the primary way to determine its quality and effectiveness.

The Internet plays a role for many faculty in the reuse process, though this is highly dependent on the conditions of access and other environmental factors. However even in the best cases of Internet availability, we found that faculty do not commonly rely on search engines to locate resources, but rather go to sites that they already know and trust. Examples of trusted sites include colleague's personal home pages, or known organizations' websites, like NASA, USGS, or Discovery.com. In one case, a faculty member actually walked to the library and asked the librarian to search the web for a particular item, and then scribbled down the URL to later type in to his browser!

Faculty resist "cold" searching the web for a number of reasons: the thousands of hits returned by search engine queries and the time required to evaluate them; the frequently unrelated or low quality websites returned from a search query; and the many links that are broken or no longer point to relevant material. In short, web searching is perceived as inefficient and frustrating because of the enormous amount of information to sort through and high variability of quality of results.

The observation that instructors are only willing to visit and search trusted sites suggests that quality filtering for resources occurs mainly at the location stage, not at the comprehension stage as one might guess, in that faculty are resistant to even search sites with which they are not familiar. This finding echoes results from software engineering which found that while the perceived quality of the repository as a whole is an important factor, the certification of individual resources to some quality level appears to have no measurable impact on reuse [31].

Research has shown that an important factor in successful web searching is knowledge not only in the domain of the topic being searched, but also in techniques of web searching itself [13, 20]. In general, we observed that most faculty have a fairly low level of expertise in web searching, most likely due to lack of training. There is evidence that effective searching and evaluation are skills that need to be taught, regardless of the searcher's domain knowledge [15]. The most common search technique we observed for faculty was to type one or two keywords into a text box (although one faculty member claimed to always compose queries of one word or less). Many faculty did not recognize a difference between different search engines, and expressed frustration when different search engines returned divergent results to the same query. The frustration felt is equally due to the limitations and often minimal query support offered by current web search engines, which in many cases only provide a single search box with no instructions for query formulation.

When considering the World Wide Web as a repository, we have found that there is a gap between the formulation of reuse intentions, and the location of effective material. We have evidence from the interviews that faculty are forming reuse intentions, but these intentions are not being translated into an effective search strategy. The implication of this finding for the design of DLESE is two-fold: the system must provide better search capabilities than traditional search engines, and it must encourage and educate faculty to utilize this functionality. In other words, part of the role of DLESE is to help educators translate their intentions into an effective search query that will lead to the location of desired resources.

3.2 Location

The first step to successful location of resources through DLESE is to ensure that faculty know and accept DLESE as a trusted site. To make DLESE known in the community of earth science

Table 2. Examining two DLESE tasks from a reuse perspective.

	Forming reuse intentions	Location	Comprehension	Modification	Sharing
Preparing for a class	Downstream: Specific, well-defined	Highly targeted	Minimal time and effort available Comprehension avoidance	Little or none: Bookmark lists; Factorization of complex resources	Relatively easy
Preparing for a course	Upstream: Broad, loosely defined	May be explorative	More time and effort available Metadata records important Supplementary aids in resources (tables of contents, indexes, etc.) necessary	Modification likely: Composition of resources; Refinements	More effort required for 'packaging'

educators, the project has actively engaged the community in many aspects of DLESE development. It has also undergone extensive outreach efforts, which are described more fully in the DLESE Community Plan [19].

In order to best support our diverse audience, the DLESE resource discovery system has been designed to be both simple and powerful. One of the important advantages of DLESE over traditional search engines is that the resources are well described with a detailed metadata scheme, and therefore can be searched for in more powerful ways. The discovery system supports both searching and categorical browsing. These two search methods have been fully described and compared elsewhere [4, 37], and it has been shown that users often use a combination of both processes in the discovery process and it is important to support both. Browsing, in particular, supports search where the query is vague or explorative in nature, which may be the case in the task of preparing for a course.

In addition to entering keywords, the discovery system also allows the user to limit the search by specifying various search parameters (type of resource, grade level, computer requirements, etc.). In order to maintain a simple and easily usable discovery system, DLESE provides two different interfaces: a simple and advanced search. The simple search interface, shown in Figure 2, is similar to a common search engine and thus familiar to many users. It provides a text box for keywords and allows the user to specify the grade level of the resource. The advanced search interface supports other search parameters, such as educational resource type, and can be used by more advanced users.

After a search is executed, a results page is shown with a listing of the resources matching the search query. Along with the URL, some key information is shown for each resource: the title, brief description, grade level, and resource type. The quality and quantity of information presented at this stage is crucial both for resource location and comprehension, as described below. The user is also provided a link to the full metadata record for each resource. In preliminary studies, we have found that users frequently utilize the metadata record when evaluating the relevancy before visiting the resource itself, demonstrating that the two processes of location and comprehension are tightly intertwined.

3.3 Comprehension

As mentioned above, the comprehension of a resource begins with the search results page, so it is important that the results page provide the right type and right amount of information. In preliminary testing, we observed comprehension avoidance strategies at this stage, in the form of keyword scanning or skipping over text entirely, when the amount of information presented was considered too great. Interestingly, we found that there was not a graceful degradation of comprehension at this point, but rather complete failure because the text was ignored entirely if it was too long!

The design of the results page is one case where there exists a tension between supporting the two tasks described earlier: preparing for a course and preparing for a class. In the former, the instructor may want more detailed resource descriptions and secondary information, like suggested variations of use, related material, etc. In the latter case, where support for extremely rapid evaluation of a resource is essential, keywords and brief descriptions are ideal. These conflicting goals must be considered not only in the design of the search results page, but everywhere that comprehension takes place. This includes the presentation of the metadata record, and even the resource itself, as described below.

Possibly the most important finding is that comprehension is a distributed process across several artifacts, both within and outside of DLESE. Indeed, we observed instructors go back and forth between the search results page, the metadata description page for a particular resource, and the resource itself in the process of comprehension. This implies that in order to support reuse, all of these components must be considered.

A frequent and unexpected cause of failure at this stage was a granularity mismatch between what was promised by the metadata and what was actually displayed at the resource URL. For example, in one case an instructor restricted his search query to search only for lesson plans. The results page returned links to large educational sites that had numerous holdings, or entire courses that had lesson plans embedded in them. This was a frustrating and ultimately unsuccessful search experience for the instructor, as he expected to be taken to a lesson plan and instead was taken to a site where he had to search further. In many cases it was not at all obvious that the site was even relevant to the query the instructor had constructed.

An implication of the granularity mismatch problem is that the system needs to be designed to better support different levels of granularity of large websites. This problem manifests in both the comprehension and modification stages. The problem of defining what comprises an item vs. a collection in a digital library is an entire research topic on its own, but it is worth stating here that accurate correspondence between the metadata and the resource it points to is essential. To achieve this it may be necessary for DLESE to separately classify components of a resource, e.g. key images or lessons plans, with their own metadata and thus treat them as separate resources.

As mentioned above, we observed that failure in comprehension can occur at the resource itself, regardless of the level of detail or accuracy of the metadata. This implies that to be serious about reuse, we cannot draw a usability line at the library boundary. Usability of the resources held in the library is equally if not more important in the success of the reuse cycle as the rest of the library system. Thus, an essential function of the digital library is to encourage resource creators to create resources that are more usable. This can be viewed as a 'packaging' problem, which will be described in detail in the sharing section.

3.4 Modification

At first blush, modifying educational resources might appear to be outside of the purview of a digital library system like DLESE. Our studies, and experiences of other educational component libraries such as the Educational Object Economy project, suggest that many educational resources, even simple textual resources, need to be modified in some way before they can be used [34]. As mentioned earlier, studies of software repositories have observed resources to evolve through four main processes: composition, factorization, refinement, and abstraction. We have observed the first three processes in our studies, which we will discuss in turn.

Composition refers to creating a new resource by combining existing resources. We observed composition activities in both the preparing for a class and preparing for a course tasks. When preparing for a class, the instructor is typically looking for a resource with fairly specific requirements to plug into their existing lecture framework and is unwilling to spend much (if any) time in modification activities. However, instructors often also desire to create a compilation of pointers to interesting 'further reading' resources for the students to use after class. This type of bookmarking or list-creating activity is also observed in the planning stages when preparing for course. Interestingly, composition activities have received the most attention in the research community. There are many digital libraries, such as the ACM's, that provide simple bookmarking and bookmark sharing facilities. There are also more ambitious projects aimed at creating digital library tools and services to support more advanced forms of composition such as Walden Paths (a tool enabling teachers to construct linear paths through web resources) [33], Iscapes (a tool enabling faculty to create a collection of mixed media resources to share with their students) [28], and ESCOT (a set of tools and services which help faculty and software developers to combine interactive educational components without programming) [32].

Factorization refers to creating new resources by partitioning more complex resources into simpler parts that can be more easily shared. We observed the need for factoring in the granularity mismatches observed during 'preparing for a class' type tasks; i.e., the faculty member wants to find a lesson plan or a certain image but it is buried in a complex 'whole course' web site. There are several different approaches that could be taken to support factoring. One approach already mentioned would be to support it at cataloging time; i.e., create separate metadata records for each object in a complex resource. For example, a cataloger might break up a course by separately indexing particular units or even key images from the course. This is extremely effortful and time-intensive, and introduces the challenge of not returning many overlapping resources from the same complex object in response to a single query. Another approach would be to work with resource creators to encourage them to create these sorts of supplementary aids as part of the library's outreach activities. Alternatively, tools could be created that factor complex resources at comprehension time. For instance, when the user chooses to view an online course or other complex resource from the DLESE search results page, a computational tool could analyze the structure of the course to construct an active table of contents to all the main sections and major media elements. This could help both comprehension and modification processes by making the structure more readily apparent and by quickly identifying simpler parts that could be used separately. Our experiences suggest that support for factoring could be an important digital library service, but to the best of our knowledge there are no library projects currently looking at this area.

Refinement refers to creating a new resource by modifying or adding to an existing resource without significantly modifying its structure. For simple textual resources, refinements require a willingness and ability to engage in basic HTML authoring. For interactive resources such as simulations or visualizations, refinements could involve programming. Our resource analysis suggests that many resources, even simple textual ones, will require refinements before they can be used. As a mundane, yet pervasive example, we have observed that many educational resources cannot be reused 'as is' simply because they are overly specific; i.e., the contact details and classroom times and locations from the last time the resource producer taught are embedded directly in the resource. Research into end-user programming and authoring tools, such as that begin done in the ESCOT project, are looking at how to help faculty refine interactive resources [32]. Existing HTML authoring tools such as FrontPage^{TM}, Dreamweaver^{TM}, or even some word processors such as MS Word^{TM}, could be used to refine textual resources. However, many faculty may lack access to these tools, may not know how to use these tools, or may simply be unwilling to spend the time to switch to another application to make even simple changes. It would be fruitful to calculate more precisely the nature of many textual refinements; it may be the case that libraries such as DLESE could provide a tool with extremely simple editing functions that could deal with many of the necessary refinements.

Abstraction refers to identifying common features across several existing resources and creating a new resource that captures these commonalities. We did not observe abstraction processes in any of our studies or analyses. Our feeling is that abstraction processes may occur in more mature reuse communities, where there are multiple similar resources to generalize across and there are users willing and able to do this type of analysis.

3.5 Sharing

As discussed under comprehension and modification, many educational resources available on the web haven't been designed with reuse by other faculty in mind, and they are mainly focused on supporting students. Several educational publishers have also observed this, even for some commercial products:

> *"If the needs of the adopting instructor or the adopting community are considered along with the needs of the student, it will be much easier to transition the innovative work that is being done into the classroom at large."* [5]

Essentially, many faculty are placing resources on the web so that they can be used by students (notably theirs), but they are not taking the extra step to do the 'packaging'; i.e., to prepare the useful supplementary materials such as tables of contents, indexes, summaries, instructors guides, etc that make complex resources, such as courses, reusable by other faculty. Even for simpler resources like interactive applets, it is rare to find syllabi or lesson plans demonstrating how an applet might be used in a course [22]. Traditionally, these ancillary materials are developed and integrated into the resource as part of the publishing process [5, 22]. One challenge for educational libraries like DLESE will be to come up with alternative processes that encourage the development of such ancillaries.

At the moment, the metadata created by DLESE staff members for library resources is serving as the reuse 'packaging' to some degree. This metadata includes basic information on the content of the resource, technical requirements, and intended users. Thus, the reuse packing is in fact distributed across the library interface and the resource itself. In the usability testing, some participants were observed to make extensive use of the descriptive metadata records as comprehension aids, but then stumble when moving to the resource itself because the resource was not as clearly structured as the metadata. Other participants did not make use of the descriptive metadata, and some became frustrated trying to comprehend complex resources. One possible remedy would be for DLESE to provide services to resource creators to feed back some of the descriptive metadata into the resource itself. For instance, once the technical requirements are summarized for the metadata it would be fairly straightforward to embed this information in a summary section in the resource.

4. RELATED WORK

The L-C-M-S model shares some important similarities and differences to more general information lifecycle models proposed by others [6, 29]. Figure 3 shows the information lifecycle model developed by Borgman et. al [6]. One important difference between the models is that the Borgman model clumps location and comprehension into a single activity: searching. We believe that it is necessary to distinguish between these two steps, because it is important that both are supported. Being able to locate good material does not necessarily mean it will be comprehended, and the gap between these two stages is frequently where reuse fails. In effect, the L-C-M-S cycle is a much more specific model that depicts the specific demands of resource reuse from an individual's perspective, whereas the Borgman model depicts a general digital library process from the resource's perspective. Interestingly, Paepcke presents a rough model, based on the field studies of information needs in an engineering firm, consisting of five processes: discover, retrieve, interpret, manage, and share [29]. While this model highlights similar processes as being important, it does not develop the implications of these processes for digital library design and outreach, nor does it situate these processes into particular types of user tasks specifically, or a reuse process more generally.

As mentioned in the Introduction, in both software engineering and knowledge management, other approaches for supporting reuse have proven to be more effective than large, centrally managed libraries. Two promising alternatives are the product line approach [31] and the small, locally-managed library approach [2]. In the product line approach, the organization makes a concerted effort to analyze existing products and determine what underlying components are shared across products. 'Packaging' efforts are focused on making these identified components reusable, and subsequently, the product line is more easily maintained and expanded using these shared components. Perhaps an analogous approach could be applied to digital library outreach: facilitators could work with an educational program or department to identify shared components or opportunities for reuse across a suite of courses. In the small library approach, repositories are created with a narrow focus on either a specific domain (e.g., avionics software) or a suite of related tasks (e.g., financial analysis). Typically, these small libraries are locally run and managed, which often means that the resource producers and the resource consumers are part of the same workgroup. Reuse is easier in these cases because producers and consumers share a similar use context, which reduces the need for and reliance upon 'packaging'.

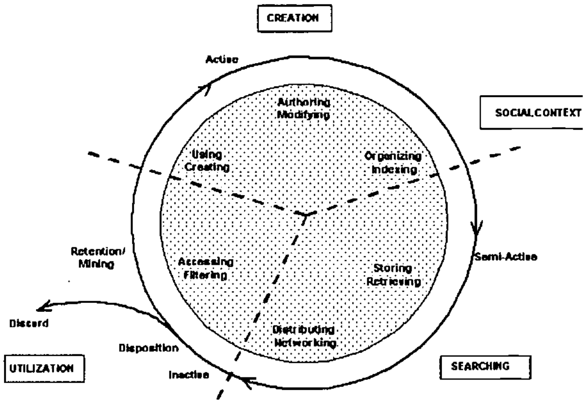


Figure 3. Information Life Cycle in Digital Libraries. Reprinted from Borgman, et.al. [6]

5. CONCLUSIONS

The preceding analysis falls into the broad category of task-centered design and analysis, which is a well-known approach to understanding software usability [18]. In our case, we examined two main tasks – preparing for a class and preparing for a course – from the perspective of a reuse model derived from the software engineering literature. We have tried to discuss how the experiences of faculty using a digital library in their teaching share some important similarities and differences to software engineers using a shared code repository. One of the key points highlighted by this analysis is that for these types of reuse tasks, we cannot draw a sharp boundary around the digital library system and only concern ourselves with the design and evaluation of the library itself. We must take into account the usability of the resources as well since the information necessary to complete the task, from the user's perspective, is distributed across the library and the resource. Analyzing our experiences from a reuse perspective highlighted several possible implications for library system design and library outreach activities.

In terms of system design, the design of the search results page is critical for supporting resource comprehension. There must be enough information to effectively summarize the resource, but not so much text that people skip reading it. Our evaluations show this to be a very fine line! Also, the library's metadata plays a central role in documenting the resource enough to support comprehension and modification processes. In effect, the descriptive metadata is 'filling in' for the lack of summarization and overview materials in the resources themselves. This observation has implications for the use of automatic cataloging tools or full-text retrieval methods. For these techniques to replace human catalogers creating metadata, they must be considered in terms of how well their output supports resource comprehension, in addition to supporting resource location.

The analysis of modification activities suggests that libraries could consider providing tools that support factorization and some simple refinement activities. While the DLESE community has always been interested in providing content creation tools as part of the library's services, modification tools have different needs in terms of functionality and in terms of when such a tool might be used. Our analysis suggests that these tools should be available in the context of the L-C-M-S cycle; thus they should be easily accessible from the search results page and designed to reduce context switching between the discovery system, the resource, and the tool itself.

In terms of library outreach, clearly one of the top priorities of the outreach effort should be helping potential library users associate resource reuse intentions with using DLESE. This is consistent with the focus of DLESE's current outreach activities. Perhaps more surprisingly is the strong need suggested by our analysis for outreach activities to work equally closely with resource creators. We must take steps to raise the level of the playing field overall in terms of designing and structuring educational resources for reuse by other faculty, not just students, and this may be part of the library's overall remit to 'manage' the collection. Currently, DLESE offers workshops on cataloging and metadata creation to interested resource creators. One possibility would be to extend these workshops to include a component on designing resources for effective reuse, which emphasizes the importance of clear labeling, summarizing, and indexing. Additionally, the 'product line' approach to promoting reuse suggests that outreach efforts might consider working with programs or departments, in addition to individual faculty members.

Finally, we tried to demonstrate that certain types of libraries, such as DLESE, serve as 'reuse repositories' and can benefit from prior reuse research in other disciplines. In the software engineering discipline, empirical studies have established that reuse improves software quality and programmer productivity [10]. This discipline has tried a number of approaches for fostering reuse in programming organizations, of which libraries are just one approach. To the best of our knowledge, no research has established the link between educational resource reuse and improved student learning or improved faculty productivity. Yet, this is the assumption underlying most educational digital library projects. Clearly research is needed that looks at whether resource reuse does improve teaching and learning, and under what conditions.

6. ACKNOWLEDGEMENTS

We wish to particularly thank our colleague Lynne Davis for her hard work in numerous areas, including the interviews and protocol studies presented here. Thanks to Mary Marlino for great discussions on reuse and helpful comments on earlier drafts of this article. This article has benefited from discussions with Gerhard Fischer and other colleagues at the Center for LifeLong Learning and Design over the years. The research described here was partially funded under Cooperative Agreement #ATM-9732665 between the National Science Foundation (NSF) and the University Corporation for Atmospheric Research (UCAR), and Grant #9978338. The views expressed herein are those of the author(s) and do not necessarily reflect the views of NSF, its sub-agencies, or UCAR.

Product Credit and trademark notifications for the products referred to are given here: FrontPage and MS Word are registered trademarks of the Microsoft Corporation. Dreamweaver is a registered trademark of the Macromedia Corporation.

7. REFERENCES

[1] Ackerman, M., Definitional and Contextual Issues in Organizational and Group Memories. in *Hawaiian International Conference on System Sciences (HICSS 27)*, (Hawaii (Jan), 1993), IEEE Computer Press.

[2] Ackerman, M. and Mandel, E. Memory in the Small: An Application to Provide Task-Based Organizational Memory for a Scientific Community, UC Irvine, 1997.

[3] Bannon, L. and Bødker, S. Constructing Common Information Spaces. in Hughes, J. ed. *Fifth European Conference on Computer Supported Cooperative Work, (Sep 7-11, Lancaster, UK)*, Kluwer Academic Publishers, The Netherlands, 1997, 81-96.

[4] Bilal, D. Children's Use of the Yahooligans! Web Search Engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of the American Society for Information Science (JASIS)*, *51* (7). 646-665.

[5] Bondaryk, L. Publishing New Media in Higher Education: Overcoming the Adoption Hurdle. *Journal of Interactive Media in Education*, *98* (3). [www-jime.open.ac.uk/98/93].

[6] Borgman, C.L. Social Aspects of Digital Libraries, UCLA-NSF Social Aspects of Digital Libraries Workshop (February 15-17), 1996.

[7] Borgman, C.L., Gilliland-Swetland, A.J., Leazer, G.H., Mayer, R., Gwynn, D., Gazan, R. and Mautone, P. Evaluating Digital Libraries for Teaching and Learning in Undergraduate Education: A Case Study of the Alexandria Digital Earth Prototype (ADEPT). *Library Trends*, Fall 2000, Vol. 49, No. 2, 228 - 250.

[8] Fischer, G., Henninger, S.R. and Redmiles, D.F. Cognitive Tools for Locating and Comprehending Software Objects for Reuse *Thirteenth International Conference on Software Engineering (Austin, TX)*, IEEE Computer Society Press, Los Alamitos, CA, 1991, 318-328.

[9] Fischer, G., Redmiles, D., Williams, L., Puhr, G., Aoki, A. and Nakakoji, K. Beyond Object-Oriented Development: Where Current Object-Oriented Approaches Fall Short. *Human-Computer Interaction*, *10* (1). 79-119.

[10] Frakes, W. and Terry, C. Software Reuse: Metrics and Models. *ACM Computing Surveys*, *28* (2 (June)). 416-435.

[11] Ginger, K., Devaul, H., Kelly, K., Sumner, T.R. and Dawe, M. DLESE Metadata Working Group Homepage, Ginger, Katy, 2000.

[12] Henninger, S. An Evolutionary Approach to Constructing Effective Software Reuse Repositories. *ACM Transactions on Software Engineering and Methodology*, *6* (2). 111-140.

[13] Hoelscher, C. and Strube, G., Searching on the Web: Two Types of Expertise. in *22nd annual international ACM SIGIR conference on Research and development in information retrieval*, (Berkeley, CA, 1999).

[14] IMS. The Instructional Management System (IMS) Project, 1994.

[15] Jacobson, F.F. and Ignacio, E.N. Teaching Reflection: Information Seeking and Evaluation in a Digital Library Environment. *Library Trends*, *45* (4). 771-802.

[16] Khoo, M., Designing a Collection Review Policy: A Case Study. *Submitted to: Joint Conference on Digital Libraries*, (Roanoke, VA (June 24-28), 2001).

[17] Kuehn, O. and Abecker, A. Corporate Memories for Knowledge Management in Industrial Practice: Prospects and Challenges. *Journal of Universal Computer Science*, *3* (8 (Special Issue on Information Technology for Knowledge Management)). Springer Science Online.

[18] Lewis, C.H. and Rieman, J. *Task-centered User Interface Design: A Practical Guide*. http://home.att.net/~jrieman/jrtcdbk.html, 1993.

[19] Manduca, C. and Mogk, D. Digital Library for Earth System Education: A Community Plan, University of Oklahoma, 2000, 44, Aailable at: www.dlese.org.

[20] Marchionini, G. Information-Seeking Strategies of Novices Using a Full-Text Electronic Encyclopedia. *Journal of the American Society for Information Science (JASIS)*, *40* (1). 54-66.

[21] Marchionini, G. and Maurer, H. The Roles of Digital Libraries in Teaching and Learning. *Communications of the ACM*, *38* (4). 67-75.

[22] Marion, A. and Hacking, E. Educational Publishing and the World Wide Web. *Journal of Interactive Media in Education*, *98* (2). [www-jime.open.ac.uk/98/92].

[23] Marlino, M., Sumner, T.R., Fulker, D., Manduca, C. and Mogk, D. The Digital Library for Earth System Education: Building Community, Building the Library. *To appear in: Communications of the ACM, Special Issue on Digital Libraries (May)*.

[24] Muramatsu, B. and Agogino, A. NEEDS — The National Engineering Education Delivery System: A Digital Library for Engineering Education *D-Lib Magazine*, 1999, Available at:http://www.dlib.org/dlib/april99/muramatsu/04muramatsu.html.

[25] Neuman, D. Learning and the Digital Library. *Library Trends*, *45* (4). 687-707.

[26] Norman, D.A. *The Psychology of Everyday Things*. Basic Books, New York, 1988.

[27] NSF. Shaping the Future: New Expectations for Undergraduate Education in Science, Mathematics, Engineering, and Technology, National Science Foundation, Arlington, VA, 1996.

[28] Olsen, F. 'Iscapes' Combine Data Sets, Computer Models, and More to Solve Problems *The Chronicle of Higher Education*, 1999, http://chronicle.com/free/99/08/9908250 1t.htm.

[29] Paepcke, A. Digital Libraries: Searching Is Not Enough; What We Learned On-Site *D-Lib Magazine*, 1996, http://www.dlib.org/dlib/may96/stanford/05paepcke.html.

[30] Poulin, J.S., Balancing the need for large corporate and small domain-specific reuse libraries. in *ACM Symposium on Applied Computing*, (Phoenix, AZ (Mar 6-8), 1994), ACM Press, 88-93.

[31] Rine, D.C., Success factors for software reuse that are applicable across domains and businesse. in *ACM Symposium on Applied Computing*, (San Jose, CA (Feb 28 - Mar 1), 1997), 182-196.

[32] Roschelle, J., C. , DiGiano, C., Koutlis, M., Repenning, A., Phillips, J., Jackiw, N. and Suthers, D. Developing Educational Software Components. *IEEE Computer*, *32*. 50-58.

[33] Shipman, F., Furuta, R., Brenner, D., Chung, C. and H., H. Guided Paths through Web-Based Collections: Design, Experiences, and Adaptations. *Journal of the American Society for Information Science*, *5* (3). 260-272.

[34] Spohrer, J., Sumner, T.R. and Buckingham Shum, S. Educational Authoring Tools and the Educational Object Economy: Introduction to this Special Issue from the East/West Group. *Journal of Interactive Media in Education*, *98* (10). Available at: http://www-jime.open.ac.uk/98/10/.

[35] Sumner, T., Domingue, J., Zdrahal, Z., Hatala, M., Millican, A., Murray, J., Hinkelmann, K., Bernardi, A., Wess, S. and Traphöner, R., Enriching Representations of Work to Support Organisational Learning. in *1st Interdisciplinary Workshop on Building, Maintaining, and Using Organizational Memories (OM-98). In conjunction with the 13th biennial European Conference on Artificial Intelligence (ECAI -98)*, (Brighton, UK (August 23-24), 1998), 09-127.

[36] Wattenberg, F. A National Digital Library for Science,Mathematics, Engineering, and Technology Education *D-Lib Magazine*, Arlington, VA, 1998, Available at:http://www.dlib.org/dlib/october98/wattenberg/10wattenberg.html#contents.

[37] Xie, H.I. Shifts of Interactive Intentions and Information-Seeking Strategies in Interactive Information Retrieval. *Journal of the American Society for Information Science (JASIS)*, *51* (9). 841-857.

[38] Ye, Y., Fischer, G. and Reeves, B., Integrating active information delivery and reuse repository systems. in *Eighth international symposium on Foundations of software engineering for twenty-first century applications*, (San Diego, CA (Nov 6 - 10), 2000), ACM Press, 60-68.

Vannevar Bush Award Winner

Building a Hypertextual Digital Library in the Humanities: A Case Study on London

Gregory Crane, David A. Smith, Clifford E. Wulfman
Perseus Project
Eaton Hall
Tufts University
Medford MA 02155
E-mail: {gcrane, dasmith, cwulfman}@perseus.tufts.edu

ABSTRACT

This paper describes the creation of a new humanities digital library collection: 11,000,000 words and 10,000 images representing books, images and maps on pre-twentieth century London and its environs. The London collection contained far more dense and precise information than the materials from the Greco-Roman world on which we had previously concentrated. The London collection thus allowed us to explore new problems of data structure, manipulation, and visualization. This paper contrasts our model for how humanities digital libraries are best used with the assumptions that underlie many academic digital libraries on the one hand and more literary hypertexts on the other. Since encoding guidelines such as those from the TEI provide collection designers with far more options than any one project can realize, this paper describes what structures we used to organize the collection and why. We particularly emphasize the importance of mining historical "authority lists" (encyclopedias, gazetteers, etc.) and then generating automatic "span-to-span" links within the collection.

KEYWORDS: automatic linking, collection development, document design, reading, browsing.

INTRODUCTION

Two years ago, we set out to create a new, densely hyperlinked digital library of materials pertaining to pre-twentieth century London and its environs [1]. The first results of this work are now available in the Perseus Digital Library (http://www.perseus.tufts.edu). This paper describes some of the results from our initial work on this collection.

Before the London work, we had spent more than a decade developing a collection of Greco-Roman cultural materials [2]. Although we learned a great deal about the tasks building such a resource entailed and the benefits such a resource could provide, we knew that these were in some ways unique to classical studies and did not necessarily pertain to humanities digital libraries as a whole. Consequently, we began to explore other domains of humanistic interest, like early modern English and the history of science [3], [4], in order to more concretely distinguish general from domain-specific issues. A collection housed at Tufts University intrigued us particularly. In 1922, the university had acquired a major set of books, maps, and pictures of London and its environs [5]. The collection is an important one, because its materials shed light on London when it was arguably the greatest city in the world; unfortunately, sequestered in the archives, it was accessible only to specialists who made their way to Tufts' special collections.

One difference from our Greco-Roman collection particularly intrigued us. The classical record is sparse: scholars spend a great deal of time determining who people were, where places were located, and what things may have looked like. By contrast, our data for the past few centuries of European and North American history are vast, and their organization and presentation raise different challenges and opportunities from those presented by the remains of the ancient world. We wanted to see how effectively we could use data available in printed form to create a digital collection that would have properties that built on, but were distinct from, its print sources. In particular, we wanted to see how time and space could be used as axes along which to organize the materials.

We also wished to discover how some of the technologies we had developed for classical study could be adapted to more modern texts. For Greek and Latin, we had surmounted a major technical hurdle that has bedeviled novices and experts since non-native speakers began to study these languages [6]. The morphology of Greek and Latin is far more complex than that of Western European languages such as English, French, Spanish, German or Italian: a single Greek verb can in theory appear in millions of different forms. We developed a system that could map inflected forms to their dictionary entries and were thus able to create links from inflected words to dictionary entries, a feature which has proven enormously popular among students of the languages. The same system allowed us to create much better retrieval tools

JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

to aid those conducting philological research. We wanted to see whether similar dense links might be useful in a collection in English, most of whose users were not desperate for all the linguistic help they could get.

This paper is aimed at two audiences. First, we hope to present one strategy of collection building for those who are themselves contemplating similar projects. We expect that many who use the London collection will be working with literary texts; nevertheless, rather than starting with major literary works (many of which were in any event already on-line), we chose to emphasize histories and descriptions of London and its environs, sources that might not occupy such a prominent position in the curriculum or public eye but that would add value to canonical literary texts. Since these reference works tend to be larger, more complex in format, and thus more expensive to digitize than literary works, such an approach was not easy, but our experiences with Greco-Roman Perseus, and now with the London collection, suggest that reference works are, in fact, a logical starting point for collection building.

We also hope to present an audience of information technologists and interface designers with a reasonably well structured test bed that is distinct in form and content from those built for the fields of science, technology, and medicine. People have been making books and reading them for thousands of years: long before the digital age, technology was shaping the organization and display of information (see, for example, a recent essay entitled "The Early Modern Search Engine: Indices, Title Pages, Marginalia and Contents," [7] part of a book called *The Renaissance Computer: Knowledge as Technology in the First Age of Print* [8]). The strategies we pursue today as we develop digital libraries build on traditions of information organization that have evolved since antiquity. The London collection provides a new historical text, primarily in English, with which to test various strategies for organizing collections. We have been particularly interested in seeing how effectively the organizational elements in these pre-twentieth century books support visualization strategies and the automatic generation of links.

Building a digital library of materials on any major city — especially on one so vast and important as London — is an open-ended task that can easily absorb decades of labor and millions of dollars. The results that we offer here constitute baseline observations after two years of work. The Greco-Roman collection in the Perseus Digital Library contains c. 35,000 images and 22 million words, of which 5 million are in classical Greek, 2.6 million are Latin and the rest primarily English. It has been evolving on the Web since 1995 and now has a substantial user base: in 2000, we served 67,000,000 pages to 6,800,000 sessions. The London collection has c. 10,000 images and 11 million words — substantially smaller but large enough to begin exhibiting problems and advantages of scale. We have only just begun to make the initial London materials available.

SUPPORTING SCHOLARLY READING

Collection design (whether the collection is digital or print) often presupposes a model for collection use. Our model requires some explanation, because it differs from those assumed by other digital library resources.

Most libraries of journal articles and monographs assume a rather utilitarian model of reading, in which the best document is the one that yields the most useful information in the shortest time with the least effort. In this model, reading is driven by explicit goals: the need to prepare a briefing on security concerns in a Latin American country, or to develop a new procedure for treating a form of hepatitis, or to find the most appropriate methodology for clustering related documents. The documents themselves are means to an end, to be absorbed and discarded. The digital support of journal-reading practice has been the object of study in its own right [9], while some worry (with good reason) about the superficiality of such "hyperextensive" reading [10].

Literary reading (insofar as there is any single practice by that name) defies (and to some extent is defined in opposition to) such utilitarian models of reading; it abjures the extraction of discrete, well-defined messages from closed works for open texts with meanings that are problematic at best. It is the most theorized and hotly contested of reading practices, and its digital formations have been written about extensively (see, for example, Landow [32], Joyce [33], Murray [34], Aarseth [11] and Douglas [12])

Those who "historicize" documents — struggling to experience them as parts of past cultures —often occupy a position that partakes of each extreme, occupying less a stable mid-point than the third point of a triangle, midway between the other two extremes but as far from each as they are from each other. On the one hand, they must immerse themselves in information: countless factoids are the raw material for larger narratives and often allow us to breathe life into the past (this is underlying argument of [35], for example). They must be, like any good researcher in any field, cold and passionate at once, able to react with delight to the dry where possible and to drag themselves through the frankly dull where necessary, ploughing through large stretches of material and retaining as much as they can.

On the other hand, many of the objects historicists study cannot be reduced to containers of information, but are objects whose meaning and interest deepen with study. Thucydides' *History of the Peloponnesian War* is not simply a historical source but an object of intensive analysis and indeed pleasure in its own right. The rise of cultural studies has brought some of the practices of literary reading to documents and objects far outside the canons of high culture. In this view, every historical text requires intense and thoughtful study if it is to be properly evaluated: we cannot even accept "objective" data sets (such as census records) unless we understand the process whereby the structuring categories are designed and the data collected.

A different view of documents and libraries emerges from this third, "historicist" perspective. On the one hand, the documents that we produce are not disposable tubes of information that can be squeezed dry and cast aside. We can expect readers to go through documents from beginning to end (often more than once), and from end to beginning; to jump from one point to another; to race through some passages and linger over others. At the same time, we can expect them to search for those materials that can give context to the words or images before them — to find "information" that will cast the object of immediate interest in a different light. Such information can range from preliminary background information (e.g., who is a particular person? where is a given place located? what is the traditional custom being mentioned?) to more complex issues touching the culture as a whole (e.g., the relationship of mass and elite, or ways of describing space or broad ritual practices). To develop their own mental models, readers need information and lots of it.

The above outline has a number of implications for digital library system design:

- Digital libraries are not designed to generate short-term remuneration, be it massive traffic (indicating that the content is heavily used) or financial gain. We want to help individuals systematically expand not only their knowledge of a particular subject but also their ability to approach problems in general.
- Collection builders want to maximize their audiences, but interpreting cultural artifacts is inherently complex, and we defeat our own purposes if we artificially simplify our materials. Good design is crucial for the broad acceptance and sophisticated use of digital collections, and attractive and engaging presentation is as important for libraries as it is for commercial sites.
- Size matters. Digital libraries need substantial bodies of material if they are to become useful, and these bodies of material may need to contain heterogeneous categories of data (e.g., animations, statistical datasets, and geospatial data as well as texts).
- If the documents in our digital library are not simply containers for information but objects of study in their own right, we need to be able to work with them at a fine level of granularity. Document to document links are not enough: we need "span to span" links connecting arbitrary subsections of documents.
- Above all we need as many links as possible between the objects of study and related materials. While most web designers aim for a small number of highly pertinent links, we need more, rather than fewer, links. We want to support free browsing and are willing to tolerate a limited number of false leads, since false leads are inherent in all serious inquiry. In the language of information retrieval, where conventional web design stresses precision (a small number of focused links), we seek to emphasize recall (getting as many links as possible).

Human-generated links are useful and critical editions traditionally provide rich connections to supporting materials. The *New Variorum Shakespeare* series (NVS), for example, produces editions of individual plays that collate every significant edition ever published, provide line by line commentary summarizing the important findings in scholarship, and include major source materials and essays on stage history, character studies, actors' interpretations, criticism, and other topics. A single such edition can require ten years of labor. One recent edition contained more than 10,000 bibliographic citations and 5,000 links to parts of the play, each of which was the product of substantial thought. However, the NVS cannot keep pace with ongoing Shakespeare scholarship, and its print volumes begin drifting out of date the minute the author hands the manuscript over to the copy-editor. The Shakespearean canon comprises 36 or 40 plays (depending on who decides the marginal cases). Even if the NVS could produce a new edition each year, the series would, on the average, reflect the state of Shakespearean scholarship eighteen years in the past. Thus, hand-generated links do not fully satisfy our needs, because they are too labor intensive to keep pace with rapid changes in scholarship. Furthermore, even if we had the scholarly labor to produce the equivalent of a variorum for every important document, we would still not be entirely satisfied because we would inevitably have questions that went beyond the editor's interests.

Thus if we are to support scholarly reading, we need to connect each document to a hypertextual digital library — a digital library that is not only large enough to support serendipitous discovery but is broken up into logical chunks that can (when appropriate) be rapidly digested. Where many Web designers strive for a few well-chosen links, our goal is to provide information about as many words and phrases as possible. We strive to create an environment that encourages the widest possible browsing and searching strategies. Rather than creating a few choice links to augment a single editorial voice, we challenge readers to refine from a superfluity of data their own paths and distinctive interpretive voice.

GENERATING LINKS FROM TEXT TO TEXT

A great deal of previous work has gone into the automatic generation of semantic links — content-based links which connect different documents that are related to one another by subject ([13]; [14, 15]), and other approaches that use automated semantic analysis to link documents (e.g., [16]). We have drawn upon this research, particularly on those aspects most relevant to our collections (e.g., cross-language document comparison between Greek and Latin: [17], [18]). Much of our present work on the London collection has centered on leveraging the information encoded in print to generate hypertextual links within our digital library.

In viewing 19th century English texts from the perspective of a twenty-first century American reader, two things about their original readership stand out. First, they knew more Latin and Greek than does the average reader today. Latin and Greek remained widespread in the British curriculum through the nineteenth century. The current London collection includes more than 1,500 passages in Latin — some of them fairly substantial, almost none translated into English. Second, they were familiar with many people, places, and topics that are no longer part of an average reader's general knowledge. These observations suggest two useful services a digital library could provide:

- By tagging classical languages (and not simply as italics), we could, as we do in classical Perseus, link inflected words to grammatical analyses, dictionaries and other linguistic support tools, making the embedded Latin and Greek quotes accessible to a wider audience.
- By tagging the names of people, places, and topics, we could link them to reference works that provide glosses and further information for readers unfamiliar with the period.

In the Greco-Roman Perseus, we have long added links from English words to what we optimistically termed the Perseus encyclopedia: several thousand small glossary entries and several hundred essays. The approach was simplistic: we added links from every instance of Homer to information about the poet and made no attempt to separate out references to, for example, "Winslow Homer". Our knowledge base also emphasized the fifth century; thus users would find eight Cleopatras, but not the famous one who allied herself with Marc Antony. Furthermore, precision and recall measures were hard to establish because they differed from text to text. Nevertheless, users made considerable use of the automatic links and they seemed to provide an important service.

We therefore set out to create a similar service for the new London collection, where we faced two basic problems. First, we were starting from scratch and had access to very little preexisting digital data that we could ourselves make freely available; we needed to build up a useful knowledge base in a relatively short period of time and with limited resources. Second, the form of proper names was more complicated: where most of our classical names were single words, the London collection had much more complex phrases: e.g. "St. Martin in the Fields" (with various combinations of spaces and hyphens).

In an ideal world, we would create a unified authority list for every proper name in the collection, but in practice, of course, this was not feasible. We needed to collect as many authority lists as we could and automatically create a unified resource that could create links from text spans to supplementary information. We wanted the user to be able to click on *Fleet Street* and then see pictures of Fleet Street, articles about Fleet Street, maps that included Fleet Street and any other relevant information.

We drew upon a variety of different resources. First, we collected conventional reference works and mined these for links. For information about famous people, we entered the 1903 one-volume index and summary volume of the *British Dictionary of National Biography:* this included names, dates and brief biographies for roughly 34,000 famous individuals. For places, we included Henry Wheatley's 1891 three volume *London Past and Present,* an encyclopedia with 9,800 entries. For geographically referenced street names, we entered a gazetteer to G. F. Cruchley's 1843 *New Plan of London.* This compact lists mentions 4,800 streets and locates them within 1/2 mile quadrants, providing a coarse but often effective georeference. We also acquired (from Bartholomew Mapping Solutions) a modern dataset for London, with vector data for more than twenty thousand contemporary streets. The two datasets complemented each other, because while the Bartholomew dataset provided vector data and much better information, many streets have disappeared since the mid nineteenth century either because of development or simple name changes. Of the 90,000 phrases that we have tagged as possible street names, only 53% (48,000) were in the contemporary Bartholomew dataset.

Besides these conventional authority lists, we also identified other sources of proper names for linking. Chapter and section headers, for example, tend to be discursive (e.g., "A Description of the Westminster Abbey"). Also, because many of the books we selected describe the history and topography of London, their tables of contents are often hierarchical and the section headers rich in proper names. We extracted 4,500 explicit headers. Less structured books provided other typographic clues as to the relevance of a page or paragraph. Augustus Hare's *Walks in London,* for example, uses italics to mark (among other things) significant place names, and several hours of labor allowed us to tag 1,500 italicized phrases as place names suitable for generating automatic links. Thomas Pennant's early nineteenth century *Popular London* italicizes every proper name, thereby reducing the heuristic value of italics for automatic place-name extraction, but it includes marginalia that we quickly mined for another 1,500 link phrases.

We also turned to image captions as an additional source of links. While we were able to collect a modest number (at present less than 1,000) of new and newly captioned color images, we drew heavily upon engravings from books in the collection. We now have more than 10,000 captions linking the user to images. If the user clicks on a link from Fleet Street, for example, she will discover that there are 98 images whose captions contain "Fleet Street."

The aggregate authority lists described above allow us to generate 284,000 automatic links for the 11,000,000 words in the collection — roughly one word in forty has an automatically generated link. We generate most of these links at runtime, using a fairly efficient algorithm to compare each text against a large list (> 200,000) of multiword matches. To minimize false hits, we prevent common words from initiating matches.

The following reproduces the output from a paragraph on Fleet Street:

> There were certainly rough doings in Fleet Street in the Middle Ages, for the City chronicles tell us of much blood spilt there and of many deeds of violence. In 1228 (Henry III.) we find, for instance, one Henry de Buke slaying a man named Le Ireis, or Le Tylor, of Fleet Bridge, then fleeing to the church of St. Mary, Southwark, and there claiming sanctuary. In 1311 (Edward II.) five of the king's not very respectable or law-fearing household were arrested in Fleet Street for a burglary; and though the weak king demanded them (they were perhaps servants of his Gascon favourite, Piers Gaveston, whom the barons afterwards killed), the City refused to give them up, and they probably had short shrive. In the same reign, when the Strand was full of bushes and thickets, Fleet Street could hardly have been continuous. Still, some shops in Fleet Street were, no doubt, even in Edward II.'s reign, of importance, for we find, in 1321, a Fleet Street bootmaker supplying the luxurious king with "six pairs of boots, with tassels of silk and drops of silver-gilt, the price of each pair being 5s." In Richard II.'s reign it is especially mentioned that Wat Tyler's fierce Kentish men sacked the Savoy church, and part of the Temple, and destroyed two forges which had been originally erected on each side of St. Dunstan's Church by the Knight Templars. The Priory of St. John of Jerusalem had paid a rent of 15s. for these forges, which same rent was given for more than a century after their destruction.

Proper names such as *Piers Gaveston*, *Fleet Street*, and *church of St. Mary* are automatically recognized. The authority list does not include *Priory of St. John of Jerusalem* but it contains *Priory of St. John* and *Jerusalem.* Although (or because) Wat Tyler is a famous historical figure, the header with his name in the DNB is unusually complex and the phrase *Wat Tyler* was not generated. Nevertheless, a reader clicking on *Tyler* would find the appropriate Tyler among the six Tyler entries in the DNB.

Clicking on a link (e.g., *Fleet Street)* calls up a list of resources whose headers, marginalia, and entry keywords are related.

Fleet Street" is in descriptions of...
1 Hare Chapter
By Fleet Street to St. Paul's., Fleet Street
95 Images
2 London sites
1.Fleet Street
2.Fleet Street Hill
10 Thornbury chapters
1.Fleet Street (Northern Tributaries--Shire Lane and Bell Yard).
2.Fleet Street (continued).
3.Fleet Street (continued).
4.Fleet Street (Northern Tributaries--Chancery Lane).
5.Fleet Street (Northern Tributaries--continued).
6.Fleet Street--General Introduction.
7.Fleet Street Tributaries--Shoe lane.
8.Fleet Street Tributaries--South.
9.Fleet Street Tributaries.
10.Fleet Street (Tributaries--Crane Court, Johnson's Court, Bolt Court).
Wheatley, London Past and Present (1891)
1.Fleet Street,

In practice this method of secondary link generation works much better than we had hoped. Developing useful precision and recall measures is problematic because the applicability of this strategy varies from document to document; furthermore, judgments of relevance vary depending upon the reader's purposes. Nevertheless, in practice it is not difficult to recognize dubious links. The casual user accustomed to carefully edited links may find the system off-putting, but the active reader who is eager to find out more about James Barry, for example, will welcome the ability to find a picture of the artist and will be willing to determine which of three James Barrys in the Dictionary of National Biography is the appropriate one. Nevertheless, a feature that lets users switch among different kinds of display depending on their preferences and information-seeking needs is clearly desirable.

Documents that discuss many disparate places and historical personages that were famous in the nineteenth century obviously benefit most from this environment, but these links also help contextualize works that occupy a largely fictive London. We automatically identify, for example, more than one hundred and fifty London locations in Dickens' *Our Mutual Friend.* Likewise, the reader confronting the phrase "the Lord Chancellor sitting in Lincoln's Inn Hall" in the opening chapter of *Bleak House* will find links to a picture and a description of that building.

Links to Visualize Time and Space

Space and time are fundamental axes for most historical collections. We decided to extract as much temporal spatial information as possible, with the goal of generating useful maps and timelines automatically.

Given its chronological focus, it is not surprising that the London collection contains many dates. Early dates usually have labels such as *A. D.* and *B. C.* to disambiguate them from other small numbers. (The consistency of this practice varies, of course, from book to book.) Furthermore, in the samples that we have examined, more than 98% of the unlabelled numbers between 1000 and 2000 in running text are dates. Most of the falsely recognized dates in this corpus come from tables — a class of data structure on which we have not yet begun serious analytical work. Overall, we have automatically identified more than 69,000 dates in the London materials; by contrast, classical source texts contain few precise dates.

Electronic timelines are hardly new (e.g. [19], [20]). In our case, we generate them from automatically extracted data as a visualization tool for documents and collections of documents (see figures 1 and 2).

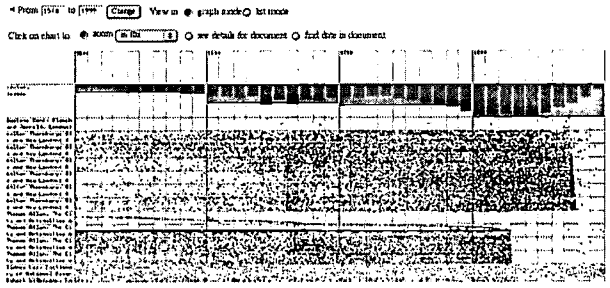


Figure 1: Part of the timelines generated for the London collection. The x-axis tracks dates and the y-axis lists the titles of books within the collection. The top bar exhibits aggregate date counts by decade and century and shows that the collection as a whole increases its coverage over time, with richest coverage focused on the 19th century. The bottom section plots dates in separate books, including six-volume and four-volume descriptions of London and, at the bottom, the summary volume of the Dictionary of National Biography. Note that the slight rightward creep of the timelines above lets us see that the two multi-volume descriptions of London were produced in installments over time.

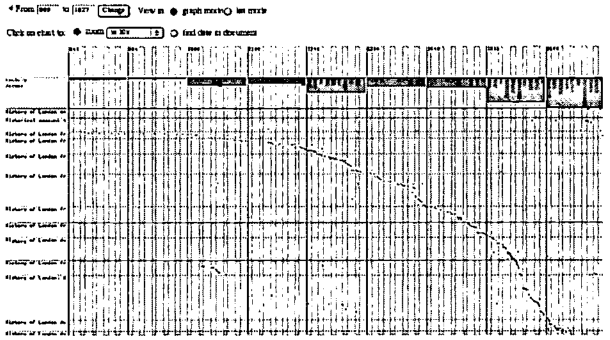


Figure 2: A Timeline for an Individual Document (in this case a narrative history of London). The y-axis lists chapters while the x-axis plots dates. A user can zoom into the timeline and/or use this as a front end to the text: clicking on a dot for 1666, will retrieve the particular page and will highlight the date. The stretch of red dots curving downwards in the middle is the temporal signature of a narrative history moving through time: the dates move steadily forward in time (i.e., they move right on the timeline on top) as we move through the text (measured by the y-axis, with chapter breaks as blue horizontal lines and marked by labels in the left hand margin).

Integrating maps with each other and with texts

The London collection at Tufts contains approximately 50 historical maps ranging in date from 1790 through the end of the nineteenth century. The integration of geographic information systems (GIS) with a larger digital library has been a long-term interest for us [21] and the extensive and precise spatial data available for London opened up possibilities not feasible with our much sketchier knowledge of ancient Rome or Athens. We georeferenced each map, aligning the historical maps to a modern GIS. Each map varies somewhat from the others, but the overall alignment works well and we can now locate the same subset of London in any map within the collection, comparing historical maps to one another or to the modern GIS. At present, we have georeferenced two dozen maps. The time required for georeferencing is less than one hour per map.

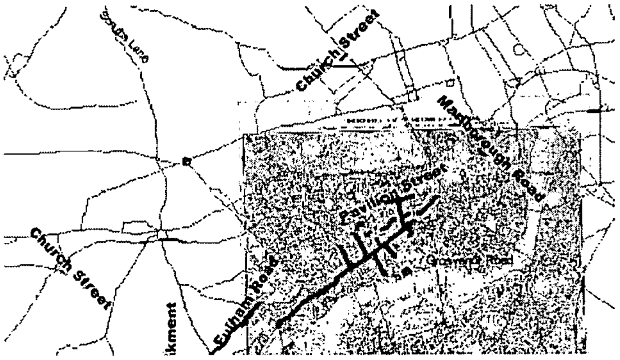


Figure 3: The above map plots vectors from a modern GIS for all the streets mentioned in a table from the 1902 edition of Charles Booth, *Life and Labour in London*. We have overlaid the modern GIS data on a georeferenced map from the period (in fact from Booth). Although some street names (such as Church Street) are ambiguous (thus limiting precision) and the modern GIS picks up no more 53% of the possible street names (thus limiting recall), the automatically generated map clearly reveals the geographic context. The user can now zoom into the modern GIS, historical map, or both.

Finally, we used the *Getty Thesaurus of Geographic Names* (TGN) to search for major geographic features. The TGN has proven to be the most difficult source to leverage. Not only is the TGN huge (more than 1,000,000 names for 886,000 locations) and ambiguous (92% of the place names that we actually encounter can refer to more than one place), but American practices of naming render semantic classification particularly challenging: Hot Coffee is the name of a town in Mississippi, for example, and there is a Monday in both Ohio and Missouri (as well as a Paraguay).

FUTURE WORK

We are clearly at an early stage of development and a great deal more could be done at every level. The scattered authority lists should be unified. We need to develop better tools to disambiguate and to filter what the users see when they pursue an automatic link. We need more content to create a richer environment for browsing and exploration. We need to develop evaluation measures that take into consideration the disparate materials within, and audiences for, this collection. Short term issues include the following:

- **Other sources of link data: Arguments and Conventional Indices:** There are other sources of information that we can use to generate useful links. Many 19th century books include brief, itemized "arguments" that summarize the content of a chapter. These break down into lists with items separated by dashes and can easily yield discrete phrases similar to the headers that we have already mined. Conventional indices likewise provide a wealth of information, including brief descriptions of people and places that do not appear in the larger reference works. Even brief hand-generated indices can disambiguate referents (e.g., the Smith on page 12 is "John" and that on page 32 is "Mary"; or the "All Saints' Church" on page 212 is in Blackheath while that on page 461 is on Margaret Street). Older books often have separate indices for people and places, thus helping bootstrap the problem of semantic classification (e.g., is Wellington a person or a place?). Some indices are as long and informative as entire books: the index to the six volume *Old and New London* contains half a megabyte of raw text and 15,000 page references to 5,600 disambiguated people and places. We need to develop strategies to mine such resources.
- **Quotes and Citation Linking**: Designers of digital libraries now routinely scan their source documents for citations and where possible convert these into active links ([22]; [23]; [24]). Classicists have been careful to establish and then maintain standard reference schemes so that the citations in nineteenth century commentaries, grammars and lexica normally work with contemporary editions. We have thus been able to mine our on-line classical reference works for more than 900,000 links. Of these, 380,000 are "commentary" notes that cite not only "Vergil Aen. 1.1" but one or more words within that reference (e.g., *arma virumque*): since each commentary note is part of a defined chunk of text, the 380,000 commentary notes are "span-to-span" links. If we follow the Dexter Hypertext reference model [25], we can generate 900,000 "LinkTo" and 380,000 "LinkToAnchor" objects, thus converting each citation into a bi-directional link. The average page of Greek and Latin text in Perseus has nine links pointing into it. For highly canonical texts such as the *Iliad*, the number of links already exceeds 100 per page. For us such density is a feature as it allows us to study problems of filtering and visualizing dense, relatively stable collections of links.

  The London collection is highly intertextual. Many of the works cite earlier authorities extensively — in some cases, more than half of a text consists of quotes from earlier authorities. In fact, over twenty percent of the collection as a whole consists of quoted material. Many of these earlier authorities are, or will in the foreseeable future become, parts of the collection. We should be able to generate a rich web of links, allowing us to see links to and from individual passages and to visualize the relationships between documents (e.g., who cites which parts of which documents). For anyone studying the development of discourse about London such links are essential.

Unfortunately, the London books rarely use conventional citations. They will often refer to "Stow" without providing any typographic or formatting clue that Stow is an author. Even when an author cites another by page number, the edition (and pagination) cited may be different from the one that we have online. And, indeed, the citing work often contains no bibliography and fails to specify which edition it happens to be citing. We have carefully used the distinction marked by the <Q> and <QUOTE> tags in the TEI DTD [26] to distinguish between literary inventions (e.g., the dialogue of characters within a novel), and true quotations drawn from sources external to the text. A digital library system should be able to search its own and federated holdings to locate the source for any text enclosed in <QUOTE> tags. If the query string is extensive enough and the source text is on-line, the chances of retrieval are good (if one can choose ahead of time, five words are usually enough to define a document: [27]). The average <QUOTE> element contains more than fifty words and this should be enough data to retrieve the source document if it is available and on-line.

- **Tabular Information**: The London collection contains at present 1,600 tables with 154,000 elements. These need to be mined for data. Several of the works that we include (Mayhew's *London Labour* and Booth's *Life and Labour*) contain important statistical information that would benefit from visualization within a GIS. Many of the books contain scattered tables with prices and wages illustrating social and economic history.
- **Monetary sums**: Monetary sums are another class of easily extracted and historically significant data — relative prices for commodities and labor are both important for scholarship and useful for giving students a sense of what people purchased and how expensive things were at a given time. The precision of monetary sum extraction is good because the texts contain various labels to indicate when number defines a currency. Where tables primarily affect the precision of our date tagging, they conversely reduce our recall of monetary sums. Our collection contains many historical lists of products and their prices: e.g. a table of prices for fowl in 1274 ("the best hen," for example, cost 3s. 2d.). As we do not yet interpret the forms of tables, we currently lose these values. Even parsing simple tables will be useful because such a process will not only yield more monetary sums but will firmly bind these monetary sums to their referents. Nevertheless, we have extracted more than 10,000 monetary sums. Simply allowing

users to search for similar sums of money would be useful. Our goal is to associate those sums with their probable referents as well (e.g., "3s. 2d." refers to the cost of the "best hen").

- **Temporal Spatial Querying**: Given automatically generated timelines and maps, the next logical step will be to query the collection by time and space: e.g., search for documents relevant to the area around St. Paul's in the 1630s.
- **Providing Link Services to External Datasets**: However much work we do on London (or any other subject), no one collection will contain everything of value. We have worked to create an initial critical mass of information on London both because we felt that this would be useful in itself and because we hoped to build an extensible environment. We will continue to expand our internal collection, but we also plan to provide linking services for third party resources (e.g., "value added surrogates" [28]). Others (e.g. [29]) could filter their documents through our linking and visualization tools. We would thus offer linking services similar to those contemplated as part of the Open Citation Project ([30]) but covering other categories including people and place names, as well as specialized language tools (e.g., links from inflected foreign language terms to their dictionary entries). The rise of XML will immensely simplify such services, since well-formed XML fragments can readily contain detailed formatting information that could enhance the precision and recall of any third party linking service.

CONCLUSIONS

Generating metadata from diverse and opportunistically acquired sources has proven extremely useful. While a great deal of effort could profitably be spent merging and resolving inconsistencies between the various authority lists that we have collected, the quickly assembled materials at hand have proven surprisingly effective. While the approach that we are pursuing may not scale to collections that contain thousands of distinct authority lists, we tentatively believe that this relatively simple approach will work well with hundreds of documents and hundreds of millions of words. We suspect that scalability will not prove a major problem for the foreseeable future because crucial reference works are much scarcer than general documents. Thus if the collection increased by two orders of magnitude, the number of key reference works would increase much more slowly and the aggregate pool of potential automatic links slower still.

- We would urge anyone bootstrapping a digital collection on a coherent subject to begin, if at all possible, with creating well-structured on-line key reference works. Such reference works are often very expensive and difficult to manage, but they lay a foundation that may make more conventional materials easier to add and then make these materials more useful when integrated with the online reference environment. We found this to be the case when we started a Digital Library on Roman culture by entering a dictionary and only then adding texts [31]. The same principle seems to be holding true with London.
- Well-organized XML documents are enormously useful for any finely grained, hypertextual digital library, but the value of XML resides in its ability not only to describe overall document structure but to precisely associate unambiguous identifiers with references to people and places. While readily available XML editors are a desideratum, we also need connectivity between these editors and external databases. We can generate useful automatic links, but these automatic links are only a starting point. Subject experts should be able to go through and refine these links, adding some, removing others and disambiguating still others. A great deal of work needs to be done on user systems (e.g., click on a map to indicate which Springfield is meant in a particular text) and on back end data processing (e.g., systems that can intelligently compare local indices or particular reference works against more global resources like the authority lists from the US Library of Congress or the Getty *Thesaurus of Geographic Names*).

ACKNOWLEDGMENTS

Support for digitizing the London Collection has come from Tufts University and especially the *Berger Family Fund for Technology Transfer.* A grant from the *Digital Library Initiative Phase 2* (NSF IIS-9817484), with particular backing in our case from the National Endowment for the Humanities, supported the system development. We would particularly like to thank our colleagues in the Perseus Project, our collaborator, Gregory Colati, Tufts University Archivist and his colleagues for the contributions in every facet of this work.

REFERENCES

1. Colati, G., *Bolles Collection Overview.* 2000, Perseus Digital Library. http://www.perseus.tufts.edu/cgi-bin/ptext?doc=2000.01.0043.
2. Smith, D., J.A. Rydberg-Cox, and G. Crane, *The Perseus Project: A Digital Library for the Humanities.* Literary and Linguistic Computing, 2000. **15**(1): p. 15-25.
3. Rydberg-Cox, J., et al., *Knowledge Management in the Perseus Digital Library.* Ariadne, 2000. **25**.
4. Crane, G., et al., *The Symbiosis between Content and Technology in the Perseus Digital Library.* Cultivate Interactive, 2000. **1**(2): http://www.cultivate-int.org/issue2/perseus/.
5. Crane, G., *Designing Documents to Enhance the Performance of Digital Libraries: Time, Space, People and a Digital Library on London.* D-Lib Magazine, 2000. **6**(7/8).
6. Crane, G., *Generating and Parsing Classical Greek.* Literary and Linguistic Computing, 1991. **6**: p. 243-245.

7. Corns, T.N., *The Early Modern Search Engine: Indices, Title Pages, Marginalia and Contents*, in *The Renaissance Computer: Knowledge Technology in the First Age of Print*, N. Rhodes and J. Sawday, Editors. 2000, Routledge: New York. p. 95-105.

8. Rhodes, N. and J. Sawday, eds. *The Renaissance Computer: Knowledge Technology in the First Age of Print*. 2000, Routledge: New York. 212.

9. Blustein, W.J., *Hypertext Versions of Journal Articles: Computer-aided Linking and Realistic Human-based Evaluation*, in *Computer Science*. 1999, University of Western Ontario: London, Ontario, CA. p. 180.

10. Levy, D.M. *I Read the News Today, Oh Boy: Reading and Attention in Digital Libraries.* In *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space-structure in hypermedia systems*. 1997. Pittsburgh, PA USA: ACM Press.

11. Aarseth, E., *Cybertext: Perspectives on Ergodic Literature*. 1997: Johns Hopkins University Press. 216.

12. Douglas, Y. and A. Hardagon. *The Pleasure Principle: Immersion, Engagement, Flow.* In *Proceedings of the eleventh ACM conference on Hypertext and hypermedia.* 2000. San Antonio, TX USA: ACM Press.

13. Allan, J. *Automatic Hypertext Link Typing.* In *Proceedings of the seventh ACM conference on Hypertext*. 1996. Bethesda, MD USA: ACM Press.

14. Green, S.J. *Building Hypertext Links in Newspaper Articles Using Semantic Similarity.* In *Third Workshop on Applications of Natural Language to Information Systems (NLDB '97)*. 1997. Vancouver, CA.

15. Green, S.J. *Automated Link Generation: Can We Do Better Than Term Repetition?* In *Proceedings of the 7th International World-Wide Web Conference.* 1998. Brisbane, Australia.

16. Shin, D., S. Nam, and M. Kim. *Hypertext Construction Using Statistical And Semantic Similarity.* In *Proceedings of the 2nd ACM international conference on Digital Libraries*. 1997. Philadelphia PA USA: ACM Press.

17. Rydberg-Cox, J., *Announcing a Greek and Latin Synonym Tool.* 1999, Tufts Universiity: Medford, MA. http://www.perseus.tufts.edu/PR/syn.ann.html.

18. Rydberg-Cox, J., *Word Co-Occurrence and Lexical Acquisition in Ancient Greek Texts.* Literary and Linguistic Computing, 2000. **15**(2): p. 121-129.

19. Kumar, V., R. Furuta, and R.B. Allen. *Interactive Timeline Editing and Review.* In *Digital Libraries '98.* 1998. Pittsburg PA USA: ACM.

20. Kumar, V. and R. Furuta. *Visualization of Relationships.* In *Proceedings of hypertext '99 on Hypertext and hypermedia.* 1999. Darmstadt, Germany: ACM Press.

21. Chavez, R.F. *Using GIS in an Integrated Digital Library.* In *Proceedings of the fifth annual ACM Digital Library Conference.* 2000. San Antonio, TX USA: ACM Press.

22. Hitchcock, S., et al. *Citation Linking: Improving Access to Online Journals.* In *Proceedings of the 2nd ACM International conference on Digital Libraries.* 1997. Philadelphia PA USA: ACM Press.

23. Baldonado, M.Q.W. and T. Winograd. *Hi-cites: Dynamically Created Citations with Active Highlighting.* In *Conference proceedings on Human factors in computing systems*. 1998: ACM Press.

24. Lawrence, S., C.L. Giles, and K. Bollacker, *Digital Libraries and Autonomous Citation Indexing.* IEEE Computer, 1999. **32**(6): p. 67-71.

25. Halasz, F. and M. Schwartz, *The Dexter Hypertext Reference Model.* Communications of the ACM, 1994. **37**(2): p. 30-39.

26. Sperberg-McQueen, C.M. and L. Burnard, eds. *Guidelines for Electronic Text Encoding and Interchange.* 1994, Text Encoding Initiative: Chicago and Oxford.

27. Phelps, T.A. and R. Wilensky. *Robust Intra-document Locations*. In *9th World Wide Web Conference.* 2000.

28. Payette, S. and C. Lagoze, *Value-Added Surrogates for Distributed Content.* D-Lib Magazine, 2000. **6**(6).

29. Levenson, M., D. Trotter, and A. Wohl, *Monuments and Dust: The Culture of Victorian London.* Institute for Advanced Technology in the Humanities. http://www.iath.virginia.edu/london/.

30. Harnad, S. and L. Carr, *Eprint Archives Through Open Citation Linking (the OpCit Project).* Current Science, 2000. **79**(5): p. 629-638.

31. Crane, G., *Extending a Digital Library: Beginning a Roman Perseus.* New England Classical Journal, 2000. **27**(3): p. 140-160.

32. Landow, G.P., *Hypertext 2.0.* 1997, Baltimore: Johns Hopkins University Press.

33. Joyce, M., *Of Two Minds: Hypertext Pedagogy And Poetics.* Studies in literature and science. 1995, Ann Arbor: University of Michigan Press. viii, 277.

34. Murray, J.H., *Hamlet on the Holodeck: the Future of Narrative in Cyberspace.* 1997, New York: Free Press. xii, 324.

35. Baker, N., *Double fold : Libraries and the Assault on Paper.* 1st ed. 2001, New York: Random House. xii, 370, [4] of plates.

Document Quality Indicators and Corpus Editions

Jeffrey A. Rydberg-Cox
University of Missouri at Kansas City
Department of English
Kansas City, MO 64110

rydbergcoxj@umkc.edu

Anne Mahoney, Gregory R. Crane
Tufts University
Perseus Project
Medford, MA 02155

{amahoney, gcrane}@perseus.tufts.edu

ABSTRACT

Corpus editions can only be useful to scholars when users know what to expect of the texts. We argue for text quality indicators, both general and domain-specific.

Categories and Subject Descriptors

Collaboration, Design and methodology, Communities of Use

General Terms

Documentation, Design, Standardization, Languages, Theory

Keywords

Editing, hypertext, corpus linguistics

INTRODUCTION

One of the challenges faced by scholars in the humanities and digital librarians alike is the need to digitize large bodies of material relatively quickly. Humanists need their source materials in digital form if they are to study them with computational methods, while digital librarians face demands to provide electronic access to large portions of their collections such as back issues of journals and special collections. One efficient and quick mechanism for this sort of digital conversion involves scanning these documents, creating minimal meta-data (such as tables of contents) and providing access to the digital images. This method, however, leverages few of the advantages of an electronic environment: texts cannot be searched, documents cannot be analyzed and mined for useful information, etc. All of these methods require that texts not simply be presented as images but that they be converted to text, whether by typists or by OCR software. This conversion introduces a new set of considerations: should the texts be tagged, what DTD should be used, what kinds of information should be tagged, and so on. But this process inevitably conflicts with the initial ideal of the rapid conversion of a large body of texts into digital form.

Previously, we have suggested the ideas of a corpus editor and a corpus edition as one possible solution to the need for rapid digitization [4]. A corpus edition is a thematically coherent collection of documents whose structure and content are tagged, according to the needs of scholars, by mostly computational techniques. In our experience creating corpus editions for the Perseus Digital Library (http://www.perseus.tufts.edu), we have found that corpus-style editing allows us to produce significant collections of useful materials in relatively short periods of time. As we will discuss below, because a corpus edition relies on automatic tagging methods, some elements of these texts will not be tagged perfectly. While we believe that the level of error introduced by automatic tagging methods is acceptable, and that a large group of texts with some errors may even be preferable to a smaller collection of carefully tagged texts, making texts in this way requires the addition of an additional piece of meta-data to the digital library, a document quality indicator that allows users to see the methodology employed and tells them what level of detail and accuracy to expect from the documents in the edition.

What Is A Corpus Edition?

A corpus edition stands in contrast to a 'clean' collection of documents with either no tagging or minimal tags preserving basic information such as page numbers or how the text was laid out on a page (i.e. *Project Guttenberg* http://promo.net/pg/ or the *Thesaurus Linguae Graecae* http://www.tlg.uci.edu/). A corpus edition also stands in contrast to carefully crafted electronic editions with extremely detailed tagging of a text's content, features, and context (i.e. the *Analytical Onomasticon to Ovid* http://www.kcl.ac.uk/humanities/cch/wlm/Onomasticon/ or the *Electronic Text Corpus of Sumerian Literature* http://www-etcsl.orient.ox.ac.uk/). The corpus editor working with a collection of texts carefully considers a minimal number of elements that should be tagged in order to make the text useful to the scholarly community (much as the designer of a hypertext system must consider how users will work with the texts [1]). A scholar working on Renaissance scientific texts must, for example, decide whether it is worthwhile to mark such formal elements of the texts as propositions, theorems, or proofs. Likewise, a person preparing an electronic edition of Shakespeare's works must decide whether to tag the original 'long s' contained in printed editions or simply to represent it as an ordinary 's'. Issues such as these exist for almost every collection of documents and the answer is not immediately obvious even to those with specialist knowledge of a field.

Decisions about what elements of a text should be tagged must always be balanced against considerations of time and scale. Corpus editions may contain dozens or hundreds of documents, representing thousands of printed pages. The corpus editor must consider not only what scholars might like to know about a text, but also which elements can practically be tagged in the large collection of documents.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

Once the corpus editor has made decisions about what elements in a text should be tagged, it is then necessary to develop scalable procedures to tag these elements in every text within the corpus. The requirements of scalability and relatively rapid production require that much of this tagging be done with computational techniques, using information extraction algorithms to identify proper names, dates, geographic locations, street names, speakers in dramatic texts, headwords in dictionaries and encyclopedias, and whatever other features are required.

Because this process relies on computational techniques, we do not assume that the corpus editor will (or even ought to) proofread every tagged element in a text. Rather, the corpus editor need only proofread enough tags to ensure that the information extraction routines are working as expected.

The Need for Document Quality Indicators

One of the fundamental precepts of the corpus edition is that purely automatic markup does not introduce so much error as to obviate the advantages of the rapid conversion of a corpus to electronic form. A corpus tagged with minimal human intervention can serve as the basis for valuable tools for patrons of digital libraries and scholars in the humanities.

This precept, however, runs counter to traditional notions of scholarship in the humanities. Scholars traditionally cultivate their editions, only publishing them to a wider audience when they have approached a certain level of perfection. Scholars generally have similarly high expectations for the works that they consume. While a digital library researcher might be well pleased to produce a system that correctly identifies 95% of the geographic locations, proper names, and dates within a text, scholars trained in the tradition of detailed and careful study of texts often find the missing 5% unacceptable.

When corpus- based editing is explained to users, however, many complaints disappear. Users need to know what they can expect from a text, and are often willing to accept errors if they know why they are there. The use of computational techniques to tag a text [i.e. 3] and the development of computational editing environments for the creation of traditional editions are, of course, well known. [i.e. 2, 5] The essential difference between these projects and corpus editions lies in the belief that texts have value before they become handcrafted editions.

Some indicators of document quality are relatively simple and can apply to any electronic corpus. Was the text entered by hand or acquired by OCR? How thoroughly the text has been proofread? Other indicators are might be relevant for only one discipline or corpus. A reader of Shakespearean texts, for example, will want to know if and how the spelling was modernized; the reader of a scientific text will want to know whether the tagged proofs were identified by hand or automatically.

Corpus editors must also document the meaning of their indicators. Does "thoroughly proofread" mean that a graduate student in the field has read the text, or that a relatively unskilled worker has checked it against a copy text? We expect that each discipline will ultimately reach a consensus about what indicators are the most important and what should be considered a high-quality text. Until this happens, corpus editors must ensure that users can find out what "good" means in a particular collection.

The document quality indicators are a form of meta-data, which must be easily available to users just as are more usual meta-data fields like the title, creator, or date. Further, this meta-data must be made available along with all the rest of the meta-data, to catalogs or to 'harvesters' (in the sense of the Open Archives Initiative http://www.openarchives.org).

Laying the Groundwork for New Editions

Careful documentation of the elements and standards used in the creation of a corpus edition has another additional benefit. Corpus editions can serve as the basis of handcrafted editions at some point in the future. It will be easier for subsequent editors to begin with the automatically tagged text than to restart the process from scratch. This possibility, however, also counters traditional ideas of scholarship in the humanities. Building a new edition or commentary based on a previously marked-up text appears at first like cheating or cutting corners. Using and enhancing a corpus edition, however, is really a form of collaboration, especially when the enhanced text is returned to the digital library. Humanists will not be able to exploit the potential of corpus editions until we develop a culture that values this kind of collaboration.

Conclusions

A corpus edition can be a useful tool for scholarship, even though its texts may contain errors. Users of these texts need to know what kinds of errors are likely and why. Each discipline will establish its own guidelines for which elements in a text should be marked, and what level of quality is acceptable. An essential part of the meta-data for each document is an indication of how it was created and how well it meets the discipline's standards for a good text. As corpus editions become more widely available, we expect further that humanists will develop new forms of collaboration based on shared electronic texts.

REFERENCES

[1] Blunstein, J. "Methods for Evaluating the Quality of Hypertext Links" *Information Processing and Management* 33.2 (1997), 255-271.

[2] Bunker, G, Zick G. "Collaboration as a Key to Digital Library Development: High Performance Image Management at the University of Washington." D-Lib 1999. 5:3. http://www.dlib.org/dlib/march99/bunker/03bunker.html

[3] Chestnutt, David R. "The Model Editions Partnership: 'Smart Text' and Beyond" D-Lib July/August 1997 http://www.dlib.org/dlib/july97/07chesnutt.html.

[4] Crane, G. and Rydberg-Cox J. "New technologies and new roles: the need for corpus editors". in *Proceedings of the 5th ACM Conference on Digital Libraries,* 2000, ACM Press, 252-253.

[5] Lecolinet, E. Likforman-Sulem, L Robert, L. Role F. and Lebrave, J-L.; "An Integrated Reading and Editing Environment for Scholarly Research on Literary Works and their Handwritten Sources" *Proceedings of the Third ACM Conference on Digital Libraries*, 1998, ACM Press, 144-151.

The Digital Atheneum: New Approaches for Preserving, Restoring and Analyzing Damaged Manuscripts *

Michael S. Brown and W. Brent Seales
Department of Computer Science
University of Kentucky
Hardymon Building, 2nd Floor
301 Rose St.
Lexington, KY 40506, USA
{mbrown,seales}@dcs.uky.edu

ABSTRACT

This paper presents research focused on developing new techniques and algorithms for the digital acquisition, restoration, and study of damaged manuscripts. We present results from an acquisition effort in partnership with the British Library, funded through the NSF DLI-2 program, designed to capture 3-D models of old and damaged manuscripts. We show how these 3-D facsimiles can be analyzed and manipulated in ways that are tedious or even impossible if confined to the physical manuscript. In particular, we present results from a restoration framework we have developed for "flattening" the 3-D representation of badly warped manuscripts. We expect these research directions to give scholars more sophisticated methods to preserve, restore, and better understand the physical objects they study.

Keywords

Digital Preservation, Humanities Computing, Image Restoration, Document Analysis, Digital Libraries

1. INTRODUCTION

There are now major efforts being undertaken throughout the world to digitize and preserve significant materials [13, 8]. Digital acquisition, which is the conversion of physical materials into a digital format, allows the possibility of efficient dissemination, and serves as a means of preservation. In addition, the digital facsimile can be manipulated in ways that are not possible for a fragile, physical artifact. Such manipulation can be used to digitally restore or enhance damaged materials. This is particularly true for digitized handwritten documents, where image processing algorithms can enhance illegible materials and provide improved data for the interested scholarly community [10, 3].

Traditionally, digitization and subsequent digital enhancement has been limited to 2-D images. This limitation is now changing, with several recent digitization efforts [2, 12, 5] focused on capturing highly detailed facsimiles using 3-D acquisition techniques. As the media stored in the digital library evolves into new and more expressive forms, we must develop new approaches and algorithms for manipulating, processing, and enhancing it.

In this paper we present research results from aspects of the *Digital Atheneum* [1], a National Science Foundation Digital Library Initiative Phase Two project. The *Digital Atheneum* encompasses research into new techniques to restore and analyze digitized collections. In particular, we are interested in new methods for acquiring and manipulating realistic facsimiles of damaged manuscripts for the purpose of enabling scholars to use these facsimiles in new ways to gain a better understanding of the physical items. Because many damaged manuscripts are no longer flat, our work involves capturing both the *images* of the manuscript, and the *three-dimensional structure* of the manuscript in the form of a high resolution shape model. Such 3-D models offer an array of uses beyond the 2-D images. For example, an accurate 3-D representation allows metric measurements to be made on the surface of the model. As described in Section 3, such measurements are valuable in a number of contexts. Furthermore, in the case of warped and crinkled documents, our recent research shows how to use the 3-D model for "virtual" flattening.

The remainder of this paper details three aspects of our research. Section 2 presents results from a 3-D acquisition effort in conjunction with collaborators at the British Library. Section 3 gives examples of how the 3-D data can be analyzed via user-specified measurements, and Section 4 presents a technical framework for restoring warped documents by flattening their 3-D facsimile.

2. 3-D ACQUISITION

2.1 Creating Digital Facsimiles

Digital photography is the most common means of creating digital content from non-traditional library materials, such as items found in special collections. While the 2-D image provides a representation that is familiar and widely accepted, it has fundamental limitations. A solitary image cannot unambiguously represent metric scale for all points within the photograph. The usual solution to this problem is to insert meta-data that describes dimensions, or to visually place a ruler next to the object during imaging. This

* We gratefully acknowledge support for this work by the NSF DLI-2 award #9817483

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

[1] www.digitalatheneum.org

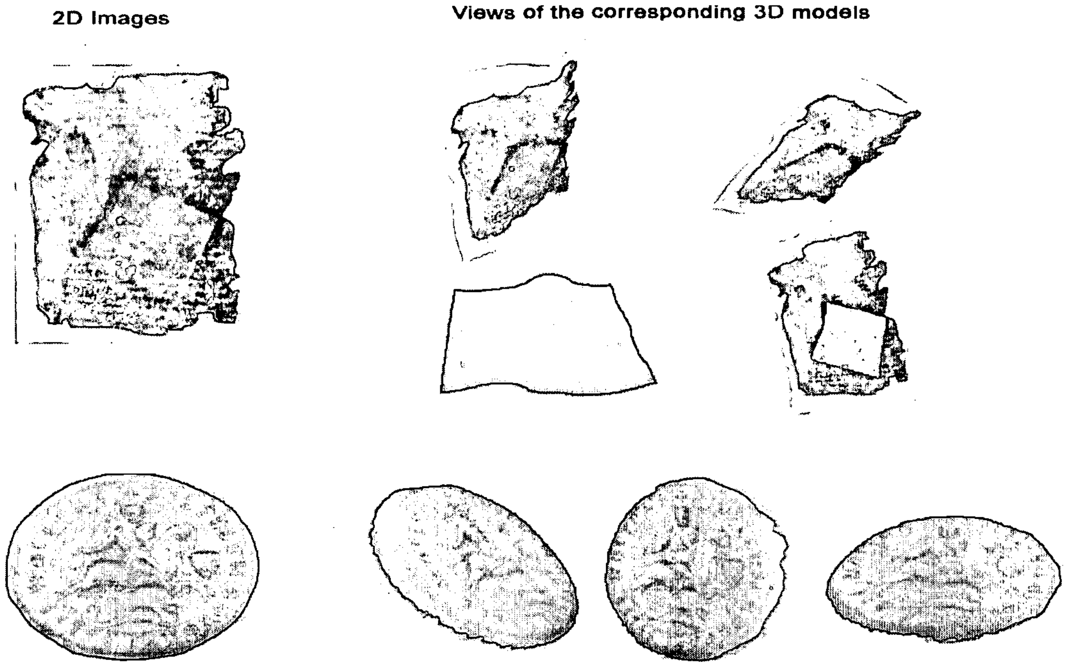


Figure 1: Top Row: A 2-D image juxtaposed with renderings of an acquired 3-D model of a manuscript shows the amount of relief the manuscript contains. This manuscript was imaged under white-light and UV light. The two images can be composited together to form a new texture for the 3-D model. Bottom Row: This acquired 3-D model of a wax seal captures detailed metric shape information.

approach to digitization makes the assumption that the object is flat, which is reasonable when considering many printed materials. There are many older, damaged texts, however, which have become warped and crinkled from age and deterioration. In addition, there are hosts of other items that have inherent 3-D shape, such as wax seals, coins, tablets, leather book bindings, etc. For such items, the image alone is insufficient to capture true 3-D shape.

We are addressing this acquisition problem as part of the Digital Atheneum [5], and have developed a structured-light computer vision technique which uses a light projector and camera to capture 3-D models. In this technique, the projector projects vertical or horizontal stripes of light onto the object. The camera observes these projected stripes and can determine the 3-D shape of the illuminated object by measuring the warp in the stripes. In the following section, we discuss issues in using this technique, and present results from manuscripts scanned at the British Library.

2.2 Acquiring 3-D Materials

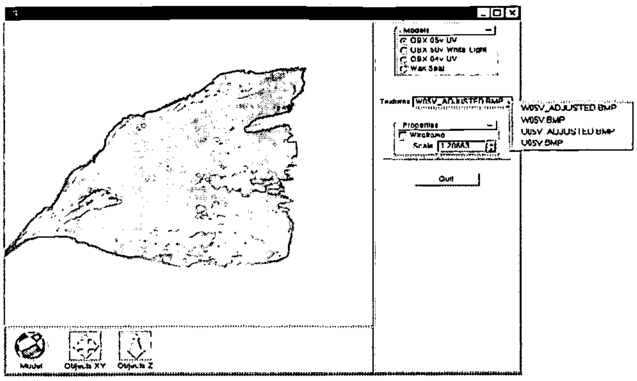
The British Library has imaged a number of collections with its ultra high resolution digital color camera from Kontron Elektronik GmbH [11]. This camera is capable of capturing images at a pixel resolution of roughly $4K \times 3K$. Unfortunately, the interface for acquiring an image is proprietary, and a software development kit (SDK) is not available. In addition, capturing an image at the highest resolution takes several minutes and must be performed through an Adobe Photoshop plugin. As a result, it was impractical to use this camera for capturing a large number of images for the purpose of recovering a 3-D representation.

The Kontron camera has a continuous PAL signal that can be used for external monitoring of the camera field of view. This feature allows continuous feedback when positioning and aligning the materials beneath the camera. We captured this PAL signal (768 × 576 pixel resolution), which is generated from the same optical path used to scan high-resolution data, at 24 frames per second. Using the PAL signal we were able to recover the 3-D shape of manuscripts using structured light [5]. Because the PAL signal and the high-resolution images are created by the same optical pathway (i.e. same lens and same sensor), registration between the imagery is straightforward. In this way we acquired 3-D data using the PAL video signal, and acquired higher resolution imagery for textures.

2.3 Results

Figure 1 shows views of the some of the acquired 3-D models. Many of the manuscripts were photographed using both white light and ultra-violet (UV) light. UV light has been successfully used to enhance certain texts that are badly damaged and difficult to see with the unaided eye [15]. One advantage of our 3-D acquisition technique over commercial laser scanners is the ability to register multiple textures easily and accurately to a single 3-D model, thus allowing for accurate compositing of textures. Figure 2 shows a visualization application for these models. This tool allows the user to select a particular model, and choose from any number of corresponding textures. The user can interactively rotate, translate, and zoom the 3-D model.

In addition to manuscript pages, we tested the acquisition system on other items, such as a wax seal (Figure 1). Overall, we acquired 3-D shape and accompanying texture models for twenty three items.



Textures W05V_ADJUSTED.BMP

Scale 1.20683

Figure 2: This screen-shot shows one of our applications for viewing an acquired 3-D model. The tool allows the user to manipulate the view of selected models and to visualized their structure in 3-D with any number of corresponding textures.

2.4 Improving the 3-D Scanner

Although we obtained good 3-D results using the PAL signal from the Kontron digital camera, a preferable solution that is likely to be more reliable and accurate is to use a high resolution camera that is supported by an available application programming interface. For example, the Kodak Professional DCS series of cameras use the high-speed IEEE 1394 interface (commonly called*firewire* or *iLink*). These cameras are available at megapixel resolutions, and Kodak provides an SDK. We are currently designing a new scanner using the Kodak DCS 330 camera, which is capable of capturing a $2K \times 1.5K$ color image and transferring the image to a host machine in roughly 10 seconds. The IEEE 1394 interface allows the camera to be driven by a notebook computer, such as a Sony VAIO. With a compact light source such as the 5 lb. Epson Powerlite projector, which can easily be mounted on a tripod, the entire system becomes even more portable. Our goal is to develop a compact, fully portable 3-D acquisition setup, which is affordable and can produce very accurate digital facsimiles.

3. ANALYSIS

The 3-D model that we acquire captures the metric scale of the original[2]. This model can be converted into a depth image. A depth image is an extended image where each pixel is given an associated "depth" value. Thus each image value $I(u, v)$ is represented by a tuple (r, b, g, x, y, z), where (u, v) is the depth image coordinate, r, g, b represent the pixel's Red, Green, and Blue color values, and x, y, z is the 3-D point recovered for the pixel position. Although this representation is larger than a standard intensity image, it directly incorporates a recovered 3-D depth representation and is easy to manipulate.

The user can perform a number of interesting operations using the depth image, which tightly couples 3-D points to pixels in the image. For instance, if the user selects two image points, $I_1(u, v)$ and $I_2(s, t)$, the metric distance, d, between these two points can be calculated directly as

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \qquad (1)$$

where x_k, y_k, z_k corresponds to the respective metric 3-D coordinates for pixel I_k stored in the depth image. The ability to make such direct metric measurements provides users with a powerful means to analyze digital facsimiles.

3.1 Examples of Metric Measurements

Figure 3 shows some examples of measurements made using 3-D facsimiles. These measurements can be computed as the direct distance between two points, or calculated as the distance along the surface of the object. In addition, irregular regions, such as the holes in the manuscript in Figure 3(c) and (d), can be selected by the user and measured. This is done by specifying a region with several connected line segments. The overall distance is simply the sum of the individual segments. Making these same measurements on the real object would be tedious if not impossible, when performed with standard tools such as a caliper or ruler.

3.2 Uses of Measurements

We envision that metric measurements may be useful in the following instances:

Monitoring Damage: From the measurements made on the surface of an object, it may be possible to monitor damaged areas over time. For example, the hole measured in Figure 3(c) could be measured periodically to see if has become enlarged. Such measurements could be made before and after an item is loaned to another institution to monitor damage from shipping and handling.

Surface Area and Volume: In addition to measuring surface distances, the 3-D representation makes it possible to determine the surface area and volume of objects. This data, combined with weight measurements, can be used to determine an object's density and thereby possible composition.

Handwriting Analysis: Brush stroke metrics and metric letter form analysis can be performed. These measurements may by useful as another tool for making arguments about authorship. Moreover, accurate measurements may help determine how to re-assemble or re-associate fragments that are physically separate but may be part of the same collection.

We have provided a technical framework that will allow scholars to perform metric measurements on collections. Our framework is independent of the importance and semantics of a particular collection. We believe that by placing this new capability into the hands of scholars who are keenly interested in the content and meaning of various objects, we will enable them to conduct a substantially more sophisticated study.

4. RESTORATION: VIRTUAL FLATTENING

Although we are able to create a 3-D model that encodes the shape of a manuscript, it is quite desirable to produce a flat facsimile even when the physical manuscript is no longer flat. A flat facsimile would make a warped document easier to read. In addition, subsequent image processing operations that derive features from a digital image, such as Optical Character Recognition (OCR) [14] and Hand Writing Recognition [7] algorithms, rely on the assumption that the input images are of flat documents.

We have developed a framework to help restore an image of a warped document by virtually flattening its 3-D model. This is achieved using a physically-based mass-spring system. Physically-based systems are typically used in computer graphics algorithms to simulate the dynamic deformation of 3-D models over time. One

[2] We acquire models at correct metric scale, within an error tolerance. We have estimated the mean value of this error to be 0.3mm for the 3-D models acquired at the British Library. See [5] for further details regarding how these error estimates are made.

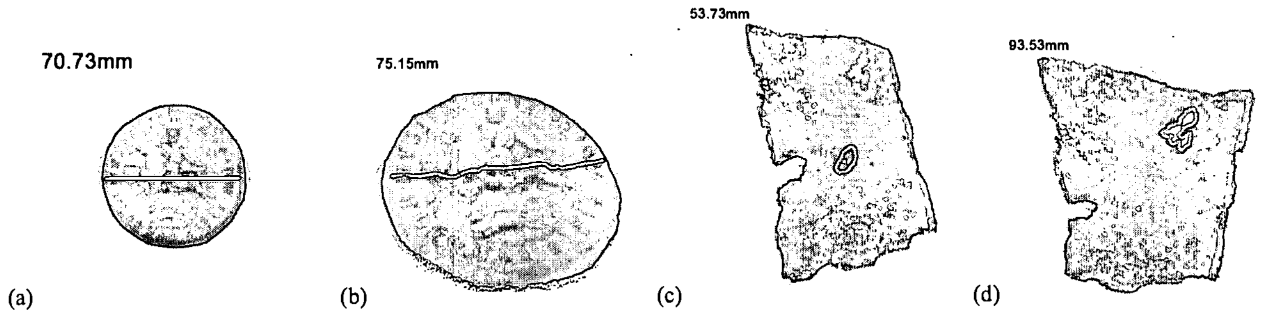


Figure 3: (a) The distance measurement, shown in the upper left corner of the image, can be specified by the user. (b) Using the same user-selected points, the measurement can be made along the 3-D *surface* of the seal, giving a slightly larger distance. This measurement would be extremely difficult to make on the physical object. (c) and (d) show circumference measurements of irregularly shaped holes on the manuscript.

notable application is in *cloth modeling*, where a flat sheet, representing a piece of cloth, is "dropped" and deforms as it hits obstacles in the simulated environment [17, 1]. The shape of the cloth changes according to the geometry of the colliding obstacles and properties of the simulation, such as gravity, the elasticity of the cloth, and so on. The manuscript flattening process can be cast as the inverse problem: given a sheet (a manuscript) in which deformations have already been applied, how can the simulation undo them to obtain the original, flat shape? The starting point for the simulation is the exact, warped 3-D shape, which we can obtain with our acquisition system. We initialize a mass-spring "sheet" with this warped 3-D shape, and force it to collide with a flat plane, which unwarps the manuscript.

The next section gives an overview of the mass-spring system and shows how we apply it to obtain results using this approach. Further details and experiments can be found in [4, 6].

4.1 Mass-Spring System

Recovered 3-D points on the surface of a manuscript form what can be considered as a system of *particles* that are able to move in 3-space. A particle system is governed by the classic second order Newtonian equation, $f = ma$, where f is a force, m is the mass of a particle, and a is an acceleration. A particle modeled by this equation can be described by its *phase state* with six variables $[x_1, x_2, x_3, v_1, v_2, v_3]$, where x_i represents the particle's 3-space position, and v_i represents its velocity. The phase state derivative with respect to time, and the subsequent motion equation, is $[v_1, v_2, v_3, f_1/m, f_2/m, f_3/m]$. This system describes a particle's mass, position and velocity at a given instance in time. During simulation, dynamic external forces such as gravity and collision forces are exerted on these particles over time. New particle positions are calculated based on these forces applied according to the equations as the time variable advances.

In a basic particle system, individual particles respond only to *external* forces, and have no influence on other particles. However, this basic system can be extended to incorporate forces between particles. One common extension, referred to as a mass-spring particle system, is formulated by logically connecting particles together via *springs*. The resulting forces in such a system can be classified into two types: *internal*, or forces between particles; and *external* forces. The slightly modified equation expressing this is

$$F_{int} + F_{ext} = ma \qquad (2)$$

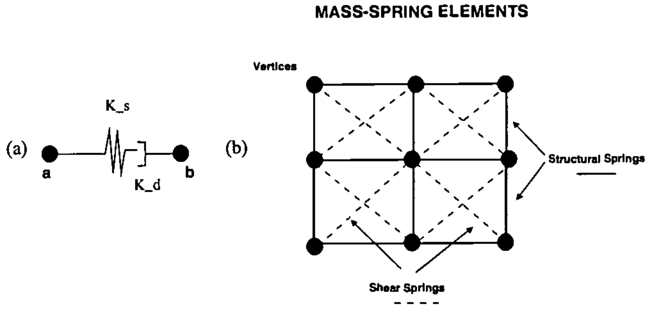


Figure 4: (a) The ideal Hookian spring, with damper, acts on two particles. K_s is the stiffness coefficient of the spring, and K_d is the dampening coefficient. (b) The finite element structure of the particles consists of structural and shear springs.

Figure 4 shows the finite elements of the mass-spring model. Particles form the vertices of quadrangles in which springs are attached. Using Provot's [16] naming convention, each element is composed of *structural* springs, which form the quadrangles' hull, and two *shear* springs, which connect diagonally. This structure is robust for modeling flexible sheet materials, such as cloth. More springs may be used to create additional rigidity if required [16].

The springs exert forces on connected particles when the two particles are moved from their resting length. These forces, governed by the ideal Hookian spring (shown in Figure 4(a)) act to keep the particles together. The Hook spring coefficients can be adjusted to control spring stiffness.

4.2 Flattening

The finite element structure described above is initialized using the acquired 3-D shape model for a manuscript. The manuscript shape is sub-sampled producing a "sheet" at a particular resolution (for example, we used 45 × 45 particles). This sheet is textured with the acquired 2-D image. As described in Section 3, texturing is straightforward using the depth image. Figure 5(Row II) shows examples of models viewed as non-planer sheets.

A flat collision plane is placed directly below the manuscript. A downward force (gravity) is exerted on the sheet. As the particles move downward, they collide with the plane. While this collision force tends to move particles away from one another, the internal spring forces tend to keep connected particles together. Eventually the surface of the sheet will come to rest on the collision plane when all of the internal and external forces have been minimized. At this point, the manuscript's 3-D structure has been unwarped and is flat. The flattened sheet can be textured with the original image, and the result is an unwarped 2-D image.

4.3 Experiments and Results

4.3.1 Controlled Trials

The first experiment is intended to quantify the ability of the mass-spring system to restore a deformed document to its original planar shape. Figure 5 shows images of two documents: one document is a checkerboard pattern, and the other is a set of printed letters. The documents are imaged while they are flat, serving as the experimental control. The documents are then crumpled by hand and imaged. The 3-D shape models of the documents are acquired as described in Section 2 (shown in Row II). These models provide the starting point for the mass-spring system. These initialized mass-spring meshes are subsequently flattened using the technique previously described.

The resulting *restored* images are compared to their respective control images. For the checkerboard image, we compare how closely the corners of the checkerboard align. We found that the mass-spring system provides a mean alignment error between corners in the restored image and corners in the original (control) image of $0.25mm$.

For the documents with printed letters, we compare the results under a commercial optical character recognition (OCR) package, Readiris Pro [9]. OCR is performed on the control image, the *unrestored* image, and the *restored* image. We compare the number of misses made by the OCR algorithm for these three documents. A miss is defined as any letter that is misclassified and any "noise" letters that are inserted by the character recognition algorithm. There are 176 letters present in the document. The control image was recognized with 100% accuracy, i.e., 0 misses. The *unrestored* image had 39 misses. The *restored* image was recognized with 100% accuracy (0 misses). These experiments were performed a number of times, with repeatable results [4].

4.3.2 Experiments With Manuscripts

The second experiment flattens a manuscript. Since there is no ground-truth for such an experiment, it is not possible to compare the simulation results to what the manuscript looked like before it became warped. However, this experiment shows the flexibility of the mass-spring framework for restoring such data. Two different materials are present in the scanned item. The original velum[3] document is embedded in a paper sleeve to preserve it and allow it to be bound without directly binding the vellum. These two materials, the vellum and the paper sleeve, have very different properties. Their interaction is often a cause for the overall page deformation. In cases such as this, where mixed materials must be modeled, the user can experiment with the flattening process by setting different internal force coefficients at portions of the mesh corresponding to each material. To demonstrate this, we first model the velum material with stiffness values making it *stiffer* than the surrounding paper. We compare this to the inverse setting, where the paper sleeve is made to be stiffer than the velum material. Figure 6 shows the results. Notice that the difference image between these two settings shows large variations between the *restored* images from the two experiments.

4.4 Restoration Summary

Our restoration framework performs well with objective measures on controlled experiments when documents have undergone rigid deformations, such as paper being crumpled by hand. For a decaying manuscript, however, it may be impossible to model all of the physical phenomena contributing to the deformed state. For such items, we are interested in manipulating the model in a reasonable and flexible way to help restore the *perceptual* quality of the digital representation. Our hope is to extend the current framework to allow user-specified constraints, which can be supplied by scholars who have specific knowledge about the *content* of the imagery. Experts who understand the intricacies of letter forms and page layout, for example, may be able to use this framework to direct the "flattening" simulation for better restoration.

5. CONCLUSION

This paper has presented several aspects of the research being conducted by the DLI-2 *Digital Atheneum* project. We have presented results from a novel 3-D acquisition effort, deployed and tested at the British Library, where several high quality 3-D models of manuscripts and similar artifacts were acquired. In addition, we presented *(1)* how metric measurements, corresponding to the real metric distances on an object's surface, can be calculated using the 3-D facsimile, and *(2)* how the 3-D representation of a warped document can be "virtually" flattened. This research is part of a broader effort to establish sound principles and practices for the creation, restoration, and manipulation of quality archives, thereby aiding those communities that increasingly rely on digital content in their scholarly activities.

6. REFERENCES

[1] D. Baraff and A. Witkin. Large steps in cloth simulation. In *Computer Graphics (Proc. SIGGRAPH)*, pages 43–52, August 1998.

[2] F. Bernardini, J. Mittleman, and H. Rushmeier. Case Study: Scanning Michelangelo's Florentine Pietà. In *SIGGRAPH 99 Course 8*, Los Angeles, August 1999.

[3] G. Braudaway. Restoration of faded photographic transparencies by digital image processing. In *Proceedings of IS and T's 46 Annual Conference*, pages 287–289, 1993.

[4] M. S. Brown and W. B Seales. Document restoration using 3d shape. Technical report no. 312-01, University of Kentucky, Lexington, Kentucky, Jan 2001.

[5] M. S. Brown and W. B. Seales. Beyond 2D images: effective 3D imaging for library materials. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pages 27–36, June 2000.

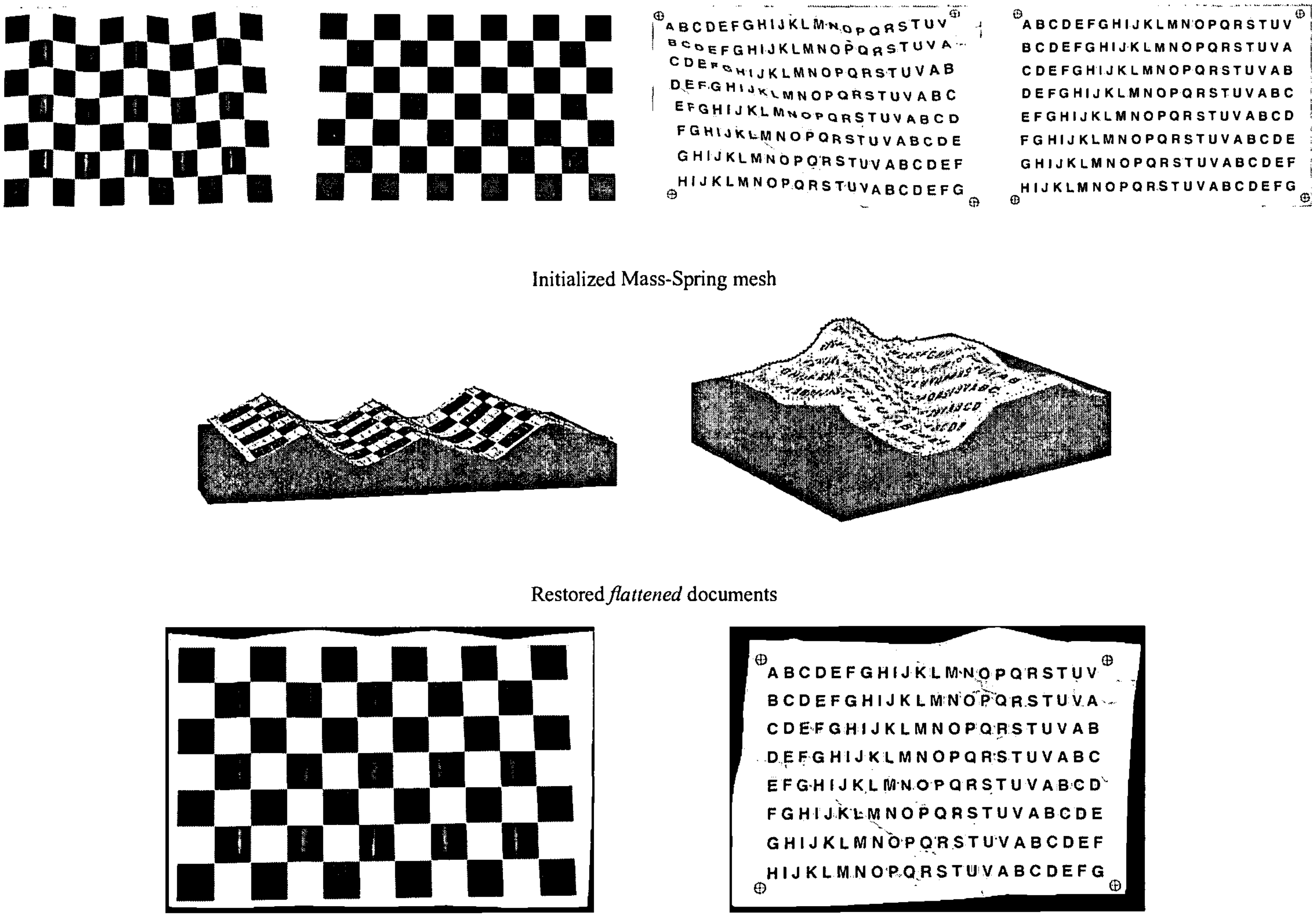
[6] M. S. Brown and W. B Seales. Document restoration using 3d shape. In *International Conference on Computer Vision (ICCV)*, June 2001. to appear.

[7] K. W. Cheung, Yeung D. Y., and Chin R. T. A Bayesian framework for deformable pattern recognition with application to handwritten character recognition. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pages 12(20):1382–1388, Dec 1998.

[8] H. M. Gladney, F. Mintzer, Schiattarella F., Bescos J., and Treu M. Digital Access to Antiquities. In *Communications of the ACM*, pages 41(4):49–57, April 1998.

[9] Image Recognition Integrated Systems (I.R.I.S). Readiris Pro. *Rue Du Bosquet 10, 1348 Louvain-la-Nueve - Belgium*, http://www.irislink.com.

[3] parchment made from animal skin



2D Images of original document (flat) and warped documents

Initialized Mass-Spring mesh

Restored *flattened* documents

Figure 5: Experiment I Row I: 2-D images of the original flat documents serve as control before crumpling them by hand. Row II: the Mass-Spring finite-element mesh is initially structured from the corresponding 3-D facsimile. Row III: the original 2-D image is correctly texture-mapped onto the restored (*flattened*) shape model.

[10] K.S. Kiernan. Digital Image Processing and the Beowulf Manuscript. *Literary and Linguistic Computing: Special Issue on Computers and Medieval Studies*, 1991.

[11] Kontron Elektronic GmbH. Kontron embedded computers ag. *Oskar-von-Miller-Strabe 1, 85386 Eching - Germany*, http://www.kontron.com.

[12] M. Levoy. The Digital Michelangelo Project. In *Proceedings of the Second International Conference on 3D Digital Imaging and Modeling*, October 5-8, 1999.

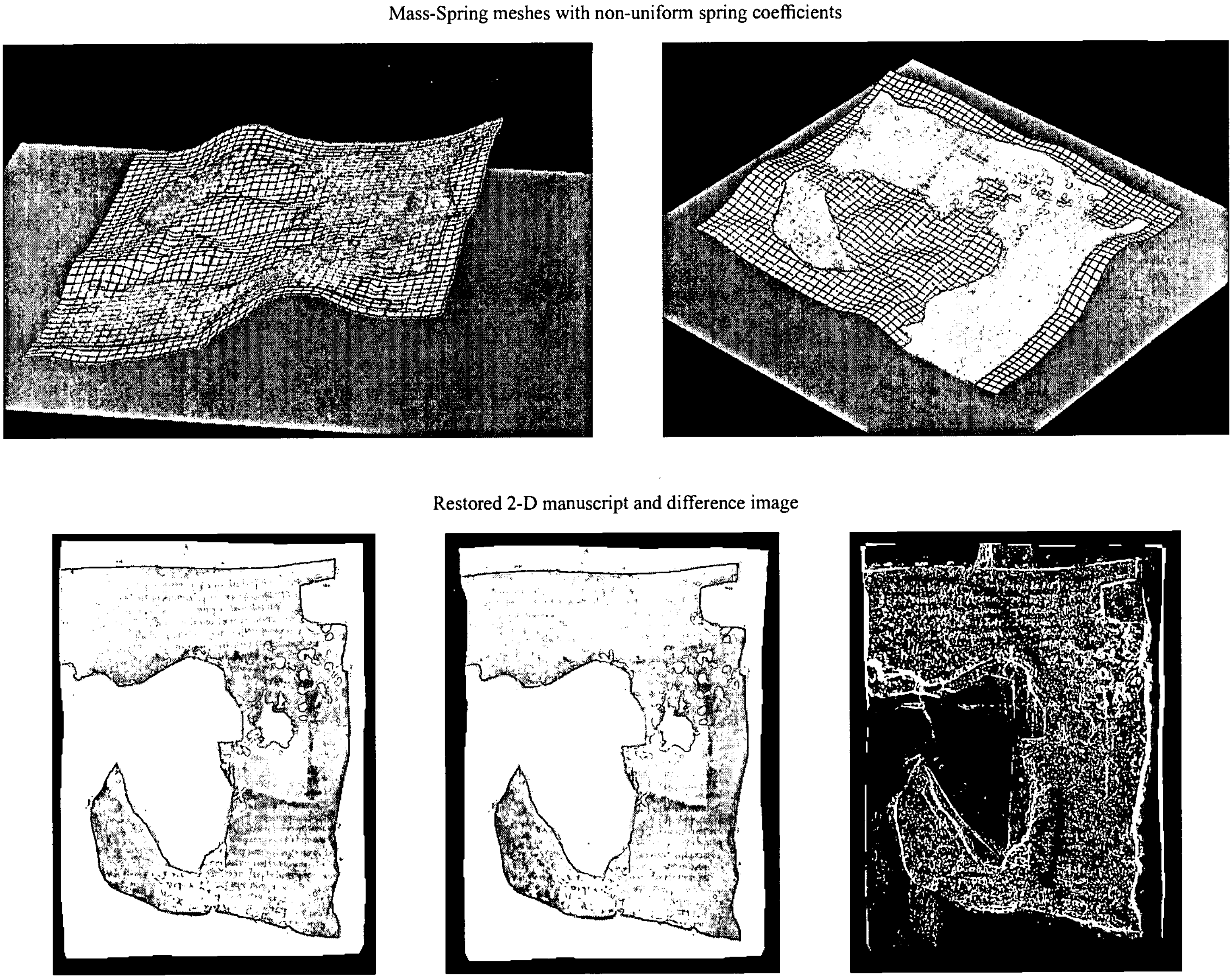
[13] B. Mannoni. Bringing Museums Online. In *Communications of the ACM*, pages 39(6):100–105, June 1996.

[14] Pavlidis, T. and Mori, S. (eds.). Optical character recognition. In *Special Issue of Proceedings of the IEEE*, pages 7(80), 1027–1209, July 1992.

[15] A. Prescott. The electronic Beowulf and digital restoration. *Literary and Linguistic Computing*, pages 12(197),185–95, 1997.

[16] Xavier Provot. Deformation constraints in a mass-spring model to describe rigid cloth behavior. In *Graphics Interface*, pages 174–155, 1995.

[17] D. Terzopoulos, J.C. Platt, and A.H. Barr. Elastically deformable models. In *Computer Graphics (Proc. SIGGRAPH)*, pages 21:205–214, 1987.



Mass-Spring meshes with non-uniform spring coefficients

Restored 2-D manuscript and difference image

Figure 6: Experiment II Top Row: The spring stiffness coefficients are non-uniform across the Mass-Spring finite-element meshes for a manuscript. The first mesh has stiffer spring parameters for the velum portion, and the second mesh is stiffer in the paper portion. Bottom Row: restored images and difference images between the two simulations.

Towards an Electronic Variorum Edition of *Don Quixote*

Richard Furuta, Shueh-Cheng Hu, Siddarth Kalasapur, Rajiv Kochumman, Eduardo Urbina, Ricardo Vivancos-Pérez[1]

Center for the Study of Digital Libraries
Texas A&M University
College Station, TX 77843-3112, USA
{furuta, shuehu, ssk9770, rajiv, e-urbina, rv} @csdl.tamu.edu

ABSTRACT

The Cervantes Project is creating an Electronic Variorum Edition of Cervantes' well-known *Don Quixote*. This paper gives an overview of the computer-based tools that we are using in this endeavor, and summarizes the current status of the project. The Electronic Variorum Edition will join the other content elements maintained by the project, which focuses on electronic resources in support of the study of Cervantes, his works, and his times.

Keywords

Humanities digital libraries, Hispanic culture, Cervantes Project, Cervantes Digital Library (CDL)

1. INTRODUCTION

The Cervantes Project, housed under the auspices of the Center for the Study of Digital Libraries at Texas A&M University, seeks to provide a comprehensive on-line reference and research site on the life and contributions of Miguel de Cervantes (1547-1616), the author of the classic *Don Quixote de la Mancha*. The project, initiated in 1995, contains a number of components: the Cervantes Digital Library (CDL), with copies of electronic editions of Cervantes' novels, plays, and other related writings; the Cervantes Digital Archive of Images (CDAI), a developing archive of photographic images on Cervantes' times and places suitable for teaching and research purposes; and the Cervantes International Bibliography Online (CIBO), a comprehensive bibliography of studies, editions, and translations of Cervantes' works.

Currently, we are creating an Electronic Variorum Edition of *Don Quixote* (EVE) for inclusion in the CDL. The EVE will contain all of the significant early editions of the text in interlinked textual and image form. In addition, it will support scholarly analysis of the differences between the editions along with interpretative commentary. The reader of the EVE will be able to customize their view of the text, examining and selecting among (or possibly combining) the choices made by independent editors. Original source material will be available to the reader, as well.

Don Quixote was published in Madrid in two volumes—1605 and 1615 (these editions are called the *princeps*). The original manuscripts have not survived and indeed only 18 copies are known to exist of the 1605 *princeps*. Consequently, our work centers around the significant editions printed during Cervantes' life. Including the two *princeps*, there are five such editions for volume 1, two for volume 2, and two containing both in a single book. Each volume is about 700 pages in length. To date we have obtained and scanned five microfilmed copies each of the two *princeps* and one or two copies of the remaining editions, for a total currently on-hand of 21 copies.

Our work-plan for each volume is first to create an "ideal:" version (a *base text*) of the *princeps* by consulting with multiple printings, identifying differences in the individual copies, and then making reasoned judgments as to a preferred rendition; this process is already underway for volume 1. To the extent permissible by our use agreements with the libraries that own the copies, we also plan to find the best available image for each page of those available to us. Subsequently, we will collate the base text against the other editions to produce the *documentary variorum edition*.

Editors of an EVE need to collate editions to find differences, resolve the differences among variants, and provide additional annotation; this is supported by the MVED (the multi-variant document editor). Readers of an EVE need the abilities to examine and customize editions; this is supported by a separate Reader's Interface. Earlier versions of these interfaces have been described elsewhere [2]; here we will sketch some of the issues we have encountered.

2. COMPONENTS OF THE EVE

2.1 Microfilms and Images

We are producing the EVE from scanned microfilm images, both for the practical reasons of access and expense, but also to investigate what could be achieved with available-quality images (our collection currently greatly exceeds what is generally available to the Cervantes scholar). Figure 1 shows a representative image.

The straightforward processing will be to "trim" the images to remove background and to correct imaging skew. This process already is underway. We also are evaluating transformations that might be applied to the text area. Here, we must take a cautious approach; see for example Donaldson's description of semantic differences introduced by interpretation of an ambiguously-printed character as "f" or "s" [1]. The clear implication here is

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

[1]The authors are listed alphabetically. Richard Furuta, Siddarth S. Kalasapur and Rajiv Kochumman also are affiliated with the Department of Computer Science. Eduardo Urbina and Ricardo Vivancos-Pérez also are affiliated with the Department of Modern and Classical Languages. The project's Web pages are at http://www.csdl.tamu.edu/cervantes/

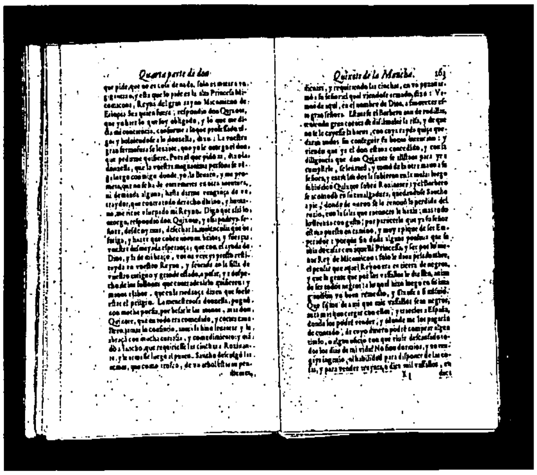


Figure 1: A representative page from microfilm.

that for major "corrections", both cleaned and original images must remain available for inspection by editors and readers.

2.2 MVED

The MVED is our software tool that enables scholars to identify, analyze, and edit variances in collated texts, given a chosen base text and different editions of the same. The MVED also enables scholars to select and annotate sections of the text.

The MVED's collator module automatically identifies variances between the base and one or more other texts. The scholar can classify acceptable variances, and can also identify variances not brought out by the collator module. Figure 2 shows a collation in progress using the MVED.

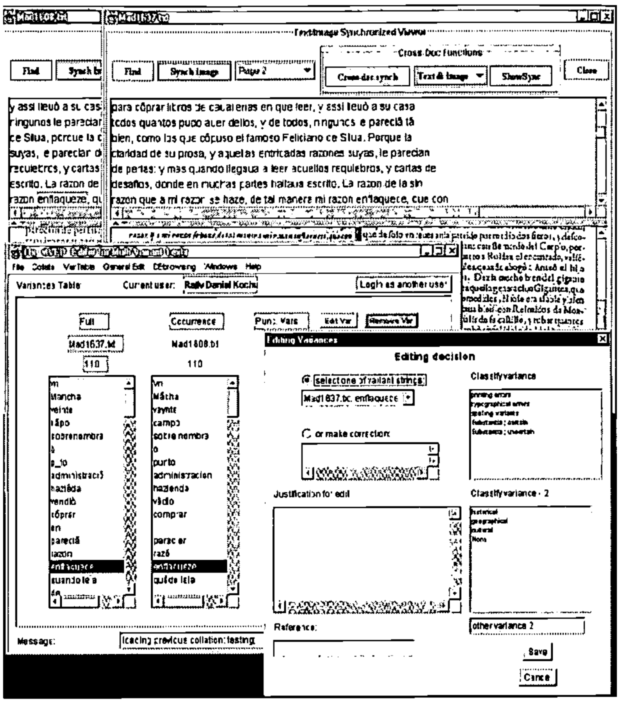
In order to aid the scholar in working with multiple texts, the MVED has a dual-form document viewer, which allows the editor to view the synchronized image of the actual document, along with the corresponding textual transcription.

Evaluation of the MVED has helped us better understand the role of annotations and the distinctions between the uses of annotation-like features (see also [3]). While variances require justification, a separate mechanism is needed to allow commentary. Additionally, the classifications appropriate for justifications are different from those appropriate for classifications. We currently support variants and annotations as separate mechanisms, each associated with free-text commentary. Additionally, as an accelerator, annotations can be attached directly to variants in addition to selected portions of the base text.

2.3 Readers' Interface

The Readers' Interface enables users to view the result of collations via the World-Wide Web. The MVED potentially makes a very large amount of data available to users, from the base texts and the comparison texts, to the collation results and statistics, along with various editors' comments on the same.

Our initial design adopted what might be called a "category-centric" model of desirable customizations (e.g., selection by category of variance, for example). Use of the interface by Humanities scholars pointed out the need instead to adopt an editor-centric model for customizations, as comprehensibility is closely tied to issues of authority.



para cõprar libros de caualerias en que leer, y assi lleuò a su casa
todos quantos pudo auer dellos, y de todos, ningunos le parecia tā
bien, como los que cõpuso el famoso Feliciano de Silua. Porque la
claridad de su prosa, y aquellas entricadas razones suyas, le parecian
de perlas: y mas quando llegaua a leer aquellos requiebros, y cartas de
desafios, donde en muchas partes hallaua escrito. La razon de la sin
razon que a mi razon se haze, de tal manera mi razon enflaquece, que con

Variantes Table   Current user: Rajiv Daniel Kochi   Login as another user

Mad1637.txt	Mad1608.txt
110	110
vn	vn
Mancha	Mãcha
veinte	vaynte
cãpo	campo
sobrenombra	sobre nombra
à	o
p_to	punto
administraciõ	administracion
haziēda	hazienda
vendiò	vēdio
cõprar	comprar
en	
pareciã	parac er
razon	razō
enflaquece	enflaqueze
quando le a	quãdo leia

Message: loading previous collation: testing

● select one of variant strings: Mad1637.txt: enflaquece
○ or make correction:

Justification for edit

Reference:

Classify variance: printing errors; typographical errors; spelling variants; substantive; certain; substantive; uncertain

Classify variance - 2: historical; geographical; cultural; None

other variance 2

Figure 2: A collation in progress using the MVED.

3. FUTURE WORK

As our work continues, and we continue to gain experience with our potential users, additional requirements are identified that lead to rethinkings of our underlying models. A strong characteristic of our use environment is the need to support both English-language and also Spanish-language speakers—both editors and also readers. Consequently, the language, or languages, understood and preferred lead to further generalizations of the interrelationships among components in our system's architecture. Additionally, the broad popularity of Cervantes, and of *Don Quixote*, raise questions of appropriately supporting quite different categories of readers—ranging from University-level researchers to grade-school children.

4. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0081420. Support for this work was provided (in part) by the Interdisciplinary Research Initiatives Program, administered by the Office of the Vice President for Research, Texas A&M University.

5. REFERENCES

[1] Peter S. Donaldson, *Digital Archive as Expanded Text: Shakespeare and Electronic Textuality.* In Katheryn Sutherland, editor, *Electronic Text: Investigations in Method and Theory.* Oxford Clarendon Press, 1997, pp. 173–197.

[2] Shueh-Cheng Hu, Richard Furuta, and Eduardo Urbina, "An Electronic Edition of Don Quixote for Humanities Scholars". *Document Numérique,* 3(1–2), June 1999, pp. 75–91.

[3] William Proctor Williams and Craig S. Abbott, *An Introduction to Bibliographical and Textual Studies*, Third Edition. Modern Language Association of America, 1999, pages 69–125.

Using Markov Models and Innovation-Diffusion as a Tool for Predicting Digital Library Access and Distribution Rates

Bruce R. Barkstrom
Atmospheric Sciences, NASA Langley Research Center
Hampton, VA 23681-2199
1-757-864-5676
b.r.barkstrom@larc.nasa.gov

ABSTRACT

This paper, discusses a general approach to predicting data access rates and user access patterns for planning distribution capacities and for monitoring data usage. The approach uses a steady-state Markov model to describe user activities and innovation-diffusion to describe the rate at which a naïve population adopts accessing data from a digital library.

Keywords

user modeling, Markov models, innovation-diffusion, user access rates, user access patterns, EOSDIS.

1. INTRODUCTION

Predicting the rate at which digital library users will access a digital library to search the holdings and the rate at which the library needs to distribute them is a problem with important economic ramifications. In the case of NASA's Earth Observing System (EOS) Data and Information System (EOSDIS), data producers will add 1 – 2 Terabytes per day to its data store over the next decade. The number of files in the system grows at a corresponding rate of 10,000 – 30,000 per day. It makes a difference to this system's planners whether there are 10,000 users accessing the system a few times a year or 10,000,000 accessing it every day. If the users need data on 8 mm tapes and require 1 TB/day distribution, media costs alone may be several $M per year.

2. MARKOV MODELS FOR USERS

While we generally recognize the folly of trying to predict the continuity of individual human behavior with 'mathematical exactitude,' we still expect such behavior to exhibit quantifiable statistical regularity. In many cases, a Markov model can provide a reasonable description of these regularities.

In a Markov model, the activities of users are described by a finite state machine with additional information related to the rate at which transitions occur. Once the user states, the transition branching ratios, and the mean times users spend in the states are quantified, it is easy to solve the steady-state equations [1]. The access rate at which a user makes a transition from being out of contact with the library to one in which he is in contact is readily derivable.

3. INNOVATION-DIFFUSION

Sociologists have studied how a community adopts an innovation, such as the spread of hybrid corn or birth control information, characterizing the spread by 'word-of-mouth' as 'innovation-diffusion.' Marketing researchers have adopted a quantitative model of such diffusion to predict the way potential customers become real customers [2].

A simple form of this kind of innovation-diffusion model is provided by assuming that there is a subpopulation of innovators and a larger subpopulation of imitators. If *P* is the population that might adopt the innovation and *C*(*t*) is the population that have adopted the innovation at time *t*, then *C* changes over time according to the equation

$$C(t + 1) - C(t) = Ini^* [P - C(t)] + Imi^* [C(t)/P]^* [P - C(t)]$$

Ini is a 'coefficient of innovation', and *Imi* is a 'coefficient of imitation. If there is no specific data for a user population, a generic set of values that summarizes the behavior of many communities is $Ini = 0.03$, $Imi = 0.38$ [10]. Time is in years.

It is not difficult to show that this model starts with a period of linear growth. With the standard coefficients, *C* will grow linearly at a rate of 3% of the potential customers per year – for about three years. At that point, about ten percent of the potential customers will have adopted the innovation. For the next six and one-half years, the growth is exponential, with a time constant of $1/(Imi - Ini)$. Thereafter, growth slows as the market saturates.

The combination of a Markov model description of user activities with innovation-diffusion for the number of digital library users provides a reasonable and readily generalizable approach to estimating future system user access and data distribution rates.

4. ACKNOWLEDGMENTS

This work was supported by CERES investigation funding.

5. REFERENCES

[1] Daellenbach, H. G., J. A. George, and D. C. McNickle. Introduction to Operations Research Techniques, Second Ed., Allyn and Bacon, Boston, MA, 1983.

[2] Thomas, R. J. New Product Development: Managing and Forecasting for Strategic Success, John Wiley & Sons, New York, NY, 1993.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
ACM 1-58113-345-6/01/0006.

Author Index

Tutorials

Sunday, June 24

Full-Day Tutorials:

1	Practical digital libraries overview / Edward A. Fox
2	Evaluating, using, and publishing eBooks / Gene Golovchinsky, Cathy Marshall, Elli Mylonas

Half-Day Tutorials:

Morning:		*Afternoon*	
3	Thesauri and ontologies in digital libraries (Part 1) / Dagobert Soergel	3	Thesauri and ontologies in digital libraries (Part 2) / Dagobert Soergel
4	How to build a digital library using open source software / Ian H. Witten	5	Hands-on workshop: Build your own digital library collections / Ian Witten, David Bainbridge
		6	Building interoperable digital libraries: A practical guide to creating open archives / Hussein Suleman

Tutorial Details:

Tutorial 1

Title:	Practical digital libraries overview
Presenter:	Edward A. Fox, Department of Computer Science, Virginia Tech
Email:	fox@vt.edu
Duration:	Full-day, Buck Mountain Room
Level:	Introductory / intermediate
Expected audience:	Medium (10-25)

Description: The tutorial will start with an overview of definitions, foundations, scenarios and perspectives. It will cover a variety of issues, including search, retrieval and resource discovery; multimedia/hypermedia; metadata (e.g., Dublin Core); electronic publishing; document models and representations; SGML and XML; database approaches; agents and distributed processing; 2D and 3D interfaces and visualizations; metrics; architectures and interoperability; commerce; educational and social concerns; and intellectual property rights, among others.

Presenter's bio: Dr. Edward A. Fox holds a Ph.D. and M.S. in Computer Science from Cornell University, and a B.S. from M.I.T. Since 1983 he has been at Virginia Polytechnic Institute and State University (VPI&SU, also called Virginia Tech), where he serves as Professor of Computer Science. He directs the Digital Library Research Laboratory, the Internet Technology Innovation Center at Virginia Tech, and varied R&D projects. He is general chair of the First ACM/IEEE Joint Conference on Digital Libraries. He is co-editor-in-chief of ACM Journal of Educational Resources in Computing (JERIC) and serves on the editorial boards of a number of journals. He has authored or co-authored many publications in the areas of digital libraries, information storage and retrieval, hypertext/hypermedia/multimedia, computational linguistics, CD-ROM and optical disc technology, electronic publishing, and expert systems.

Tutorial 2

Title:	Evaluating, using, and publishing eBooks
Presenter:	Gene Golovchinsky (FX Palo Alto Laboratory) and Cathy Marshall (Microsoft), Elli Mylonas (Scholarly Technology Group, Brown University)
Email:	cathymar@microsoft.com (C. Marshall), elli_mylonas@brown.edu (E. Mylonas)
Duration:	Full-day, Harrison/Tyler Room
Level:	Introductory / intermediate
Expected audience:	Medium (10-25)

Description: This tutorial is an introduction to eBooks. Presenters will discuss and compare existing hardware (devices such as the Softbook, the Rocket eBook, Palm Pilot, etc.) and their software, document representation formats (PDF, HTML, Open eBook Format, MS Reader etc.), the electronic publishing process, and the future of reading on such devices. A portion of the tutorial will be devoted to an introduction of the Open eBook Format Specification and how to apply it to create documents that can be read on any OEB-compliant reader. The tutorial will conclude with an open-ended discussion about the future possibilities for such devices, focusing on opportunities to overcome some limitations of paper books and documents.

Presenter's bio:

- **Elli Mylonas** is the Associate Director for Projects and Research at the Scholarly Technology Group, Brown University. Her current work includes a project to convert structured document formats to OEB. She was one of the founding members of the Perseus Project at Harvard University, an early hypertext for scholarly and pedagogical use. She has worked extensively with SGML and XML and OEB, and has spoken and published on markup, hypertext and academic computing projects.
- **Cathy Marshall** is an Architect and Senior Researcher in the eBooks Group at Microsoft, and a long-time participant in the international Hypertext, Digital Library, and WWW research communities. She is on the Board of Directors of the Electronic Literature Organization. Her research lies in the disciplinary interstices of computer science, social science, and the arts. See http://www.csdl.tamu.edu/~marshall
- **Gene Golovchinsky** is a Senior Research Scientist at FX Palo Alto Laboratory (FXPAL), where he is a member of the Mobile computing group. His research interests include user interface design (with an emphasis on information exploration and information retrieval), hypertext, and pen-based computing. Gene completed his Ph.D. at the University of Toronto in 1996. Prior to joining FXPAL, Gene had worked at GMD-IPSI in Darmstadt, Germany, at IBM, and at Kaiser Electronics. See http://www.fxpal.xerox.com/people/gene.

Tutorial 3

Title:	Thesauri and ontologies in digital libraries. Part 1: Structure and use in knowledge-based assistance to users; Part 2: Design, evaluation, and development
Presenter:	Dagobert Soergel, College of Information Studies, University of Maryland, College Park
Email:	ds52@umail.umd.edu
Duration:	Two half-days (same day), Monroe Room
Level:	Introductory [Part 1] , intermediate [Part 2]
Expected audience:	Medium (19-25) [Part 1] / Small (5-10) [Part 2]

Description:

[Part 1 - MORNING] This introductory tutorial is intended for anyone concerned with subject access to digital libraries. It provides a bridge by presenting methods of subject access as treated in an information studies program for those coming to digital libraries from other fields. It will elucidate through examples the conceptual and vocabulary problems users face when searching digital libraries. It will then show how a well-structured thesaurus / ontology can be used as the knowledge base for an interface that can assist users with search topic clarification (for example through browsing well-structured hierarchies and guided facet analysis) and with finding good search terms (through query term mapping and query term expansion - synonym expansion and hierarchic expansion). It will touch on cross-database and cross-language searching as natural extensions of these functions. The workshop will cover the thesaurus structure needed to support these functions: Concept-term relationships for vocabulary control and synonym expansion, conceptual structure (semantic analysis, facets, and hierarchy) for topic clarification and hierarchic query term expansion. It will introduce a few sample thesauri and some thesaurus-supported digital libraries and Web sites to illustrate these principles.

[Part 2 - AFTERNOON] This tutorial is intended for people who have a basic familiarity with the function and structure of thesauri and ontologies (such as acquired in Part 1 or in previous tutorials). It will introduce criteria for the design and evaluation of thesauri and ontologies and then deal with methods and tools for their development: Locating sources; collecting concepts, terms, and relationships to reuse existing knowledge; developing and refining thesaurus/ontology structure; software and database structure for the development and maintenance of thesauri and ontologies; collaborative development of thesauri and ontologies; developing crosswalks/mappings between thesauri/ontologies. In summing up, the tutorial will address the question of the amount of resources needed to develop and maintain a thesaurus or ontology.

Presenter's bio: Dagobert Soergel holds an M.S. equivalent in mathematics and physics (1964) and a Ph.D. in political science (1970), both from the University of Freiburg, Germany. He is Professor of Information Studies, University of Maryland, where he teaches courses in information retrieval, thesaurus development, expert systems, and information technology, and an information systems consultant. He has been a visiting professor at the universities of Western Ontario, Chicago, and Konstanz, Germany. Among other books, he has authored Organizing Information (1985), which received the American Society of Information Science Best Book Award, Indexing Languages and Thesauri: Construction and Maintenance (1974), and numerous papers. He has developed several thesauri, most recently the Alcohol and Other Drug Thesaurus (http://etoh.niaaa.nih.gov/AODVol1/Aodthome.htm) for which he chairs the advisory committee, and is developing TermMaster, a thesaurus management software package. In 1997 he received the American Society of Information Science Award of Merit.

Tutorial 4 (morning)

Title:	How to build a digital library using open-source software
Presenter:	Ian H. Witten, Department of Computer Science, University of Waikato
Email:	ihw@cs.waikato.ac.nz
Duration:	Half-day, Wilson Room
Level:	Intermediate
Expected audience:	Medium (10-25)

Description: This tutorial describes how to build a digital library using the Greenstone digital library software, a comprehensive, open-source system for constructing, presenting, and maintaining information collections. Collections built automatically include effective full-text searching and metadata-based browsing facilities that are attractive and easy to use. They are easily maintainable and can be rebuilt entirely automatically. Searching is full-text, and different indexes can be constructed (including metadata indexes). Browsing utilizes hierarchical structures that are created automatically from metadata associated with the source documents. Collections can include text, pictures, audio, and video, formed using an easy to use tool called the Collector. Documents can be in any language: Chinese and Arabic interfaces exist. Although primarily designed for Web access, collections can be made available, in precisely the same form, on CD-ROM or DVD. The system is extensible: software "plugins" accommodate different document and metadata types. The Greenstone software runs under both Unix and Windows, and is issued as source code under the GNU public license. Attendees will receive an extensive user manual and should learn enough to download the software and set up a digital library system. Those with programming skills should be able to extend and tailor the system extensively.

Presenter's bio: Ian H. Witten is Professor of Computer Science at the University of Waikato in New Zealand, and directs the New Zealand Digital Library project (where the Greenstone software originates). He has published widely in the areas of digital libraries, data compression, information retrieval, and machine learning. He is co-author of Managing Gigabytes: Compressing and Indexing Documents and Images (Second edition, Morgan Kaufmann 1999) and Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (Morgan Kaufmann 2000), as well as many journal articles and conference papers. He is a fellow of the ACM and of the Royal Society of New Zealand, and a member of professional computing, information retrieval, and engineering associations in the UK, USA, Canada, and New Zealand.

Tutorial 5 (afternoon)

Title:	Hands-on workshop: Build your own digital library collections
Presenter:	Ian H. Witten and David Bainbridge , Department of Computer Science, University of Waikato
Email:	ihw@cs.waikato.ac.nz (I. Witten), davidb@cs.waikato.ac.nz (D. Bainbridge)
Duration:	Half-day, Wilson Room
Level:	Intermediate
Expected audience:	Small (5-10)

Description: This is a hands-on laboratory-style workshop that follows on from the tutorial "How to build a digital library using open-source software." Attendees will first install the basic Greenstone system (described in the former tutorial) on their own computer. Then they will learn how to personalize its appearance, how to build their own collections, and how to take advantage of advanced features such as interactive phrase browsing. The primary goal is to enable attendees to create a collection of their own material that they bring along to the workshop, and leave the workshop with that collection (and others) installed on a digital library system on their own computer.

Presenter's bio:

- **For Ian H. Witten**, see above.
- **David Bainbridge** is a faculty member in Computer Science at the University of Waikato, New Zealand. An active member of the New Zealand Digital Library project, he has worked with several United Nations Agencies, the BBC and various public libraries. He holds a Ph.D. in Optical Music Recognition from the University of Canterbury, New Zealand where he studied as a Commonwealth Scholar. Since moving to Waikato in 1996 he has continued and broadened his interest in computer music research, which has received international press and TV coverage and was co-recipient of the Digital Libraries Vannevar Bush award in 1999. David has also worked as a research engineer for Thorn EMI in the area of photo-realistic imaging and graduated from the University of Edinburgh in 1991 as the class medallist in Computer Science.

Tutorial 6 (afternoon)

Title:	Building interoperable digital libraries: A practical guide to creating Open Archives
Presenter:	Hussein Suleman, Department of Computer Science, Virginia Tech
Email:	hussein@vt.edu
Duration:	Half-day
Level:	Introductory / intermediate
Expected audience:	Medium (10-25)

Description: The Open Archives Initiative (OAI) is dedicated to solving problems of digital library interoperability by defining simple protocols, most recently for the exchange of metadata. This tutorial is aimed at introducing individuals to the concepts underlying OAI and its Protocol for Metadata Harvesting, as well as providing sufficient information to allow attendees to almost immediately implement the standard on their own archives. In addition, attendees will be introduced to issues that need to be addressed when building new systems, either in the capacity of being providers of data, users of data, or both.

Presenter's bio: Hussein Suleman is a Ph.D. student working with Edward Fox at Virginia Tech. His research focus is on topics closely related to matters of interoperability. He is currently funded by NSF to work with the iLumina project (http://www.ilumina-project.org) which is building a federated digital library of resources for science and technology education, incorporating resources from the Computer Science Teaching Center (http://www.cstc.org). This and other projects utilize technology developed by the Open Archives Initiative. He served as part of the technical working group that produced the latest revision of the Open Archives Metadata Harvesting Protocol. In this capacity he implemented the standards on various platforms and also developed and actively maintains the Repository Explorer software that is used by the Open Archives community to rigorously test archives for compliance with the standards.

DEMONSTRATIONS

Interactive Visualization of Video Meta-Data
Howard D. Wactlar, Mark Derthick (Carnegie Mellon University)

A System for Adding Content-Based Searching to a Traditional Music Library Catalogue Server
Matthew J. Dovey (Kings College, London)

Content Management for Multi-Presentation Digital Museum Exhibitions
Jen-Shin Hong, Bai-Hsuen Chen, Jieh Hsiang, Tien-Yu Shu (National ChiNan University)

Hierarchical Document Clustering of Digital Library Retrieval Results
Christopher R. Palmer, Jerome Pesenti, Raul E. Valdes-Perez, Michael G. Christel, Alex G. Hauptmann, Dorbin Ng, and Howard D. Wactlar (Carnegie Mellon University)

Indiana University Digital Music Library Project
Jon W. Dunn and Eric J. Isaacson (Indiana University, Bloomington)

PERSIVAL: Categorizing Hidden-Web Resources
Panagiotis G. Ipeirotis, Luis Gravano, Mehran Sahami (Columbia University)

PERSIVAL; Personalized Search and Summarization over Multimedia HealthCare Information
Noemie Elhadad, Min-Yen Kan, Simon Lok and Sm aranda Muresan (Columbia University)

Stanford Encyclopedia of Philosophy
Edward N. Zalta, Uri Nodelman (Stanford University), Colin Allen (Texas A&M University)

Using the Repository Explorer to Achieve OAI Protocol Compliance
Hussein Suleman (Virginia Tech)
PERSIVAL: View Segmentation and Static/Dynamic Summary Generation for Echocardiogram Videos
Shahram Ebadollahi, and Shih-Fu Chang (Columbia University)

POSTERS

An Atmospheric Visualization Collection for the NSDL
Keith Andrew (Eastern Illinois University), Christopher Klaus (Argonne National Laboratory), Gerald Mace (University of Utah)

Building the Physical Sciences Information Infrastructure: A Phased Approach
Judy C. Gilmore, Valerie S. Allen (U.S. Department of Energy)

A National Digital Library for Undergraduate Mathematics and Science Teacher Preparation and Professional Development
Kimberly S. Roempler (Eisenhower National Clearinghouse, The Ohio State University)

A versatile facsimile and transcription service for manuscripts and rare old books at the Miguel de Cervantes Digital Library
Alejandro Bia (University of Alicante, Spain)

Breaking the Metadata Generation Bottleneck: Preliminary Findings
Elizabeth D. Liddy, Eileen Allen, Michelle Monsour, Jennifer Liddy (Syracuse University), Stuart Sutton, Anne Turner (University of Washington), Woojin Paik, Sarah Harwell (solutions-united.com)

Development of an Earth Environmental Digital Library System for Soil and Land-Atmospheric Data
Eiji Ikoma, Taikan Oki, Masaru Kitsuregawa (Institute of Industrial Science, The University of Tokyo)

Digital Facsimile Editions and On-Line Editing
Harry Plantinga (Computer Science, Calvin College)

DSpace at MIT: Meeting the Challenges
Michael Bass (Hewlett-Packard), Margret Branschofsky (Faculty Liaison, MIT)

Exploiting Image Semantics for Picture Libraries
Kobus Barbard, David Forsyth (University of California at Berkeley)

Feature Extraction for Content-Based Image Retrieval in DARWIN: Digital Analysis and Recognition of

Whale Images on a Network
Kelly R. Debure, Adam S. Russell (Eckerd College)

Guided Linking: Efficiently Making Image-to-Transcript Correspondence
Cheng Jiun Yuan, W. Brent Seales (University of Kentucky)

Integrating Distributed Digital Libraries by using CORBA, XML and Servlet
Wing Hang Cheung, Michael R. Lyu, Kam Wing Ng (The Chinese University of Hong Kong)

A National Digital Library for Undergraduate Mathematics and Science Teacher Preparation and Professional Development
Kimberly S. Roempler (Eisenhower National Clearinghouse, The Ohio State University)

Print to Electronic: Measuring the Operational and Economic Implications of an Electronic Journal Collection
Carol Montgomery, Linda Marion (Hagerty Library, Drexel University)

A Versatile Facsimile and Transcription Service for Manuscripts and Rare Old Books at the Miguel de Cervantes Digital Library
Alejandro Bria (Miguel de Cervantes Digital Library, University of Alicante, Spain)

The Virtual Naval Hospital: The Digital Library as Knowledge Management Tool for Nomadic Patrons
Michael P. D'Alessandro MD, Donna M. D'Alessandro MD, Mary J.C. Hendrix PhD (University of Iowa, Iowa City), CAPT Richard S. Bakalar MC USN (Naval Medical Information Management Center), LT Denis E. Ashley MC USNR (United States Navy Bureau of Medicine and Surgery)

Using Markov Models and Innovation Diffusion as a Tool for Predicting Digital Library Access and Distribution Rates
Bruce R. Barkstrom (NASA Langley Research Center)

Turbo Recognition: Decoding Page Layout
Taku A. Tokuyasu (University of California at Berkeley)

Panels

Panel 3C (90min): The Open Archives Initiative: Perspectives on Metadata Harvesting

Moderator: James B. Lloyd (University of Tennessee)

Panelists:

- Don Waters (Mellon Foundation)
- Tim Cole (Chair, Library Information Technology Committee, University of Illinois at Urbana-Champaign)
- Caroline Arms (Library of Congress)
- Donald Waters (Mellon Foundation)
- Jeffrey Young (OCLC)

Panel 4C (90min): Different Cultures Meet - Lessons Learned in Digital Library Development

Moderator: Ching-chih Chen (Professor, Graduate School of Library and Information Science, Simmons College)

Panelists:

- Hsueh-hua Chen (Chair, Department of Library and Information Science, National Taiwan University)
- Wen Gao (Deputy President, Graduate Schools, Chinese Academy of Sciences, Beijing)
- Von-Wun Soo (Professor of Computer Science, National Tsinghua University, Taipei)
- Li-Zhu Zhou (Chair, Department of Computer Science, Tsinghua University, Beijing)

Panel 5C (90min): Digital Library Collaborations in a World Community

Moderator: David Fulker (Unidata Program Center)

Panelists:

- Sharon Dawes (SUNY Albany)
- Leonid Kalinichenko (Institute of Informatics Problems, Russian Academy of Science, Moscow State University)
- Tamara Sumner (Center for LifeLong Learning and Design, Dept. of Computer Science and the Institute of Cognitive Science, University of Colorado)
- Constantino Thanos (DELOS Director, Consiglio Nazionale delle Ricerche, Istituto di Elaborazione della Informazione)
- Alex Ushakov (ChemQuest Project, Department of Chemistry and Biochemistry, University of Northern Colorado)

Panel 6C (90min): A Digital Strategy for the Library of Congress - Discussion of the LC21 Report and the Role of the Digital Library Community

Moderator: Alan Inouye (Computer Science and Telecommunications Board, National Academy of Sciences)

Panelists:

- Dale Flecker (Associate Director of Planning and Systems, Harvard University Library)
- Margaret Hedstrom (Associate Professor, School of Information, University of Michigan.)
- David Levy, Professor (Information School, University of Washington)

Panel 7C (90min): The President's Information Technology Advisory Committee's February 2001 Digital Library Report and its Impact

Moderator: Sally E. Howe (National Coordination Office for Information Technology Research & Development

Panelists:

- David Nagel (President AT&T Labs, Chair PITAC Digital Library Panel)
- Ching-chih Chen (Professor, Graduate School of Library and Information Science, Simmons College, and PITAC)
- Stephen M. Griffin (NSF, Digital Libraries Initiative)
- James H. Lightbourne (NSF, National SMETE Digital Library (NSDL) Program)
- Walter L. Warnick (Department of Energy, Director Office of Scientific and Technical Information)

Panel 8C (2hr): The National SMETE Digital Library Program

Moderator: Brandon Muramatsu (SMETE.ORG Project Director, University of California at Berkeley)

Panelists:

- James Lightbourne (National SMETE Digital Library (NSDL) Program,National Science Foundation)
- Marcia Mardis, Merit Networks, TeacherLib NSDL Project)
- Cathryn A. Manduca (University Corporation for Atmospheric Research)
- Flora P. McMartin (University of California at Berkeley)

Panel 11C (2hr): High Tech or High Touch?

Moderator: David Levy (Professor, Information School, University of Washington)

Panelists:

- Diane Nester Kresh (Director Public Services Collections, Library of Congress)
- Oren Etzioni (Associate Professor, Department of Computer Science and Engineering, University of Washington)
- William Arms (Professor, Computer Science Department, Cornell University)
- Barbara Tillett (Director, Integrated Library System Program Office, Library of Congress)

Panel 12C (90min): Digital Libraries Supporting Digital Government

Moderator: Gary Marchionini (Cary C. Boshamer Professor, School of Library and Information Science, University of North Carolina at Chapel Hill)

Panelists:

- Lawrence E. Brandt (Program Manager for Digital Government Research, National Science Foundation)
- Hsinchun Chen (University of Arizona)
- Anne Craig (Illinois State Library)
- Judith Klavans (Columbia University)

Panel 13C (90min): Digital Music Libraries - Research and Development

Moderator: Christine Brancolini (Director, Indiana University Digital Library Program)

Panelists:

- David Bainbridge (University of Waikato)
- Mary Wallace Davidson (William and Gayle Cook Music Library, Indiana University)
- Andrew P. Dillon (School of Library and Information Science, Indiana University)
- Matthew Dovey (Libraries Automation Service, University of Oxford)
- Jon W. Dunn (Digital Library Program, Indiana University)
- Ichiro Fujinaga (Peabody Conservatory of Music, Johns Hopkins University)
- Eric J. Isaacson (School of Music, Indiana University)
- Michael Fingerhut (IRCAM-Centre Pompidou)

Workshops

W1 Visual Interfaces to Digital Libraries - Its Past, Present, and Future (Roanoke Ballroom B)

W2 Technology of Browsing Applications (Roanoke Ballroom GH)

W3 Classification Crosswalks: Bringing Communities Together (The 4th Networked Knowledge Organization Sources/Systems (NKOS) Workshop) (Roanoke Ballroom EF)

w4 ~~Workshop on Digital Libraries in Asian Languages: Accessing, Indexing, and Organizing Digital Libraries in Asia~~ **(CANCELLED)**

W5 ~~Information Visualization for Digital Libraries~~ (CANCELLED)

(W1) Title: Visual Interfaces to Digital Libraries - Its Past, Present, and Future

Chairs:

Katy Börner, Assistant Professor
Information Science & Cognitive Science
Indiana University, SLIS
10th Street & Jordan Avenue
Main Library 019
Bloomington, IN 47405, USA
Phone: (812) 855-3256
Fax: (812) 855-6166
E-mail: katy@indiana.edu
http://ella.slis.indiana.edu/~katy

Dr Chaomei Chen, Reader
Department of Information Systems and Computing
Director, The VIVID Research Centre
Brunel University
Uxbridge UB8 3PH, UK
Tel: +44 1895 20 30 80
Fax: +44 1895 251 686
Email: chaomei.chen@brunel.ac.uk
http://www.brunel.ac.uk/~cssrccc2/

Location: Roanoke Ballroom B

Expected audience: The workshop is suitable for researchers, practitioners, and graduate students in the areas of information visualization, digital libraries, human-computer interaction, library and information science, and computer science.

Description: The primary aim of the workshop is to raise the awareness of several interconnected fields of research related to the design and use of visual interfaces to digital libraries, especially in information visualization, human-computer interaction, and cognitive psychology. This workshop also aims to stimulate participants to reflect on the state of the art in their own fields by identifying challenging issues concerning visual interfaces and thereby fostering a ultidisciplinary research agenda for future research and development.

Objectives:

- To provide a stimulating forum for researchers in information visualization and digital libraries to share their views, experiences, and plans.
- To raise the awareness of the state of the art in related fields of research.
- To identify a research agenda concerning the role of visual interfaces in digital library research.
- To exploit potentially useful theories, methodologies, and technologies.
- To establish long-term interdisciplinary collaboration between researchers from different fields.

Submission and Selection: You are invited to submit a 2-page position paper. The camera-ready copy of accepted papers can be up to 6 pages long.

Program Committee:

Ann Blandford, Middlesex University, UK
Kevin Boyack, Sandia National Laboratories, USA
Martin Dodge, University College London, UK
Xia Lin, Drexel University, USA
John MacColl, University of Edinburgh, UK
Sougata Mukherjea, Inktomi Corporation, USA
Sue O'Hare, Post Office Research Group, UK
Henry Small, Institute for Scientific Information, USA

Planned publications: Workshop participants will be invited to produce an extended version of their work for publication in an edited book in the Springer Book Series on Information Visualization to provide a comprehensive coverage of the topic to a wider audience.

For more information, connect to the WSs webpage: http://vw.indiana.edu/visual01/

[go to top of page] △

(W2) Title: The Technology of Browsing Applications

Chairs:

Nina Wacholder
Center for Research on Information Access
Columbia University
Email: nina@cs.columbia.edu
Phone: 212-939-7119
Fax: 212-666-0140

Craig Nevill Manning
Computer Science
Rutgers University
Email: nevill@cs.rutgers.edu
Phone: 917-202-7145
Fax: 801-760-7628

Location: Roanoke Ballroom GH

Expected audience: Information retrieval community, Natural language processing community, Publishers and other purveyors of document content, Managers of digital libraries.

Description: Phrase browsing applications provide information seekers with access to text content via structured lists of index terms. These lists provide a preview of the content of a collection. The index terms, which may be identified by a variety of techniques, are phrases that represent important concepts referred to in a document or collection of documents. The browsing system supports interactive navigation and organization of the phrases. The goal of this workshop is to bring together researchers interested in any aspect of phrase browsing technology, including, but not limited to, identification of index terms, techniques for hierarchical organization of the terms, implementation of efficient systems, usability of browsing applications, and techniques for evaluating this technology.

Selection process: We are soliciting long papers (up to 6 pages) and short papers (up to 2 pages). We also invite proposals for a panel discussion. Submissions will be reviewed by the Chairs of the workshop and a program committee. For more information, go to the Workshop's web page, http://www.cs.columbia.edu/~nina/browsingtechnologywkshop.html

Publications: Workshop Proceedings to be published electronically on the JCDL 2001 Web Site.

[go to top of page] △

(W3) Classification Crosswalks: Bringing Communities Together (The 4th Networked Knowledge Organization Sources/Systems (NKOS) Workshop)

Chairs:

Gail Hodge
Information International Associates, Inc./National Biological Information Infrastructure
312 Walnut Pl.
Havertown, PA 19083
Voice: 610/789-6769
Fax: 865/481-0390
Email:Gailhodge@aol.com

Paul Thompson
West Group
Email:Paul.Thompson@westgroup.com

Diane Vizine-Goetz
Senior Research Scientist
OCLC Office of Research
OCLC Online Computer
Library Center, Inc.
6565 Frantz Road
Dublin, OH 43017-3395,
USA
Email:vizine@oclc.org
Phone: +1-614-764-6084
Fax: +1-614-718-7519

Marcia Lei Zeng, Ph.D.
Associate Professor
School of Library and
Information Science
Kent State University
Kent,OH 44242-0001
(330)672-0009 (direct phone)
(330)672-2782 (school office)
(330)672-7965 (fax)
Email:mzeng@kent.edu

Web Site:http://nkos.slis.kent.edu/

Location: Roanoke Ballroom EF

Expected audience: The goal of the NKOS activity is to develop a community of researchers and developers who are working toward creating interactive Knowledge Organization Systems accessible over the Web. This includes thesaurus and ontology developers, digital library and information infrastructure developers, information scientists, and library professionals. NKOS Workshops have been held at the last three ACM DL meetings.

Description: Mapping between/among classification schemes is beneficial within an organization that has a number of implicit or explicit schemes, between organizations seeking to exchange information, and in a digital library context where collections are organized by different classifications. This cross scheme mapping could be done manually, but if many schemes are to be mapped, it may be desirable to provide automated tools to support the process. This workshop will present research and projects that identify the state-of-the-practice and outline the research agenda. Participants will also be encouraged to give short presentations on other NKOS-related activities.

Outline

Session 1: Technical Session on Classification Crosswalks
Session 2: Open Forum for Presentations of Activities and Discussion by Participants
Session 3: Continuing work on NKOS activities: a taxonomy of knowledge organization sources and an XML DTD for vocabulary mark-up

Selection process: First-come, first-served

Planned Publication(s): A brief report of the meeting will be published on the NKOS web site (http://nkos.slis.kent.edu) and in D-Lib Magazine.

[go to top of page] △

~~(W4) Workshop on Digital Libraries in Asian Languages: Accessing, Indexing, and Organizing Digital Libraries in Asia~~ CANCELLED

Chairs:

Professor Su-Shing Chen
University of Missouri-Columbia
Dept of Computer Engineering & Computer Science
University of Missouri-Columbia
Columbia, MO 65211
Tel: 573-882-5176
Fax: 573-882-8318
Email:chens@missouri.edu

Professor Ching-chih Chen,
Graduate School of Library and Information Science
Simmons College
300 The Fenway
Boston, MA 02115, USA
Tel: 617-521-2804
Fax: 617-521-3192
Email: chen@simmons.edu

This workshop has been cancelled by the workshop chairs.

[go to top of page] △

(W5) ~~Information Visualization for Digital Libraries~~ (CANCELLED)

Chairs:

Lucy Nowell, Chief Scientist
Synthesis, Analysis and Visualization of Information
Battelle / Pacific Northwest National Laboratory (PNNL)
Tel: (509)372-4295
Fax: (509)375-3641
Email: lucy.nowell@pnl.gov

Beth Hetzler, Chief Scientist
Synthesis, Analysis and Visualization of Information
Battelle / Pacific Northwest National Laboratory (PNNL)
Voice: (509)375-6690
Fax: (509)375-3641
Email:beth.hetzler@pnl.gov

This workshop has been cancelled by the workshop chairs.

[go to top of page] △

Student Volunteers

The period to accept Student Volunteers is closed

Thanks for your interest, an email of the final list of student volunteers will be emailed soon.

Student volunteers will get a free conference registration and will be able to attend some of the paper and panel sessions as well as, symposia, and tutorials.

Student volunteers will work approximately 20 hours. Duties will include tasks, such as answering questions from conference attendees, monitoring entrance to sessions, assisting with registration, and running miscellaneous errands.

This is a great opportunity to be a part of the cutting edge in the digital library research from both the ACM and IEEE communities. There are plenty of opportunities to interact with the best and most prominent people in the field! If you are close to graduation and looking for a job this is one of the best places to get recruited.

Examples of the SV tasks:

- run registration desk and distribute proceedings,
- check badges for different sessions,
- help with A/V equipment,
- distribute materials for workshops and tutorials.

Student volunteer benefits:

- No conference fees.
- Attend papers, panels.
- Conference proceedings.

Lodging Information:

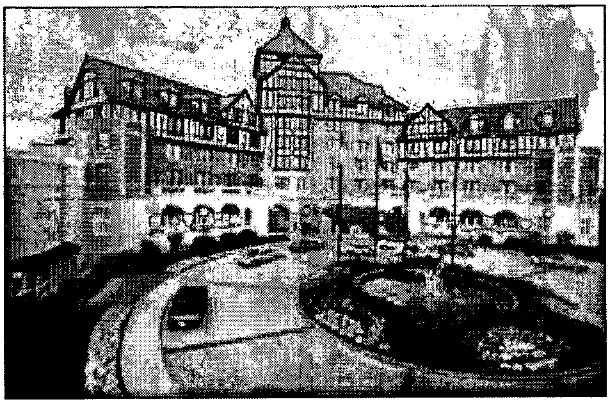
We will try to arrange for reduced rates at the conference hotel. We will also seek out other, less expensive hotels in the area and help interested students get together to share room costs.

Student volunteers may send an email to abdulla1@llnl.gov.

About Roanoke

***Nestled among the rolling Blue Ridge Mountains,* Roanoke** combines the convenience of a major city with a friendly small-town atmosphere. Minutes from the Blue Ridge Parkway, the Appalachian Trail and Smith Mountain Lake, Roanoke provides many opportunities for outdoor recreation. Local tourism centers round the natural beauty of the area, the Appalachian mountain culture, and the history of the American South.

JCDL01's host, the Hotel Roanoke and Conference Center, is part of that history. During Roanoke's days as a railroad center, the "white lady on the hill" was recognized internationally as one of the country's grand railroad hotels. The hotel was fully renovated in the 1990s and is now managed by Doubletree Hotels in partnership with Virginia Tech. The modern conference center incorporates flexible meeting spaces, ergonomic furniture, continuous break facilities, T-1 Internet access, and Ethernet networking. Taken as a whole, the facility is a charming combination of old southern style and new world technology. The hotel is only minutes away from the Roanoke Regional Airport.



Crossing the railroad tracks on an enclosed pedestrian bridge, the visitor enters a thriving revitalized downtown with restaurants, coffee shops and entertainment all available in easy walking distance. Downtown Roanoke is also home to the historic Farmer's Market, the oldest such market in continuous use in Virginia.
Produce, flowers, local crafts and mountain delicacies are all available on a stroll through the market, as is the friendly society of the vendors.

Local Trips

While at the conference, go here and visit Virginia Wineries, the Wilderness in Bottom Creek George Hike, or a recreation of the life in Virginia in the old times.

Travel

The Roanoke airport is served by US Airways, Continental, Delta, Northwest and United. US Airways and Delta have been selected as official carriers for JCDL01, and are offering discounted fares. Both US Airways and Delta offer a 10% discount on unrestricted round-trip coach fares and a 5% discount on certain other fares. For reservations on US Airways you or your travel agent can call the US Airways Group and Meeting Reservation Office toll-free at 877.874.7667 and refer to **Gold File #55661791**. For Delta reservations, call Delta Meeting Network Reservations at 800.241.6760 and refer to **File Number 172531A**.

Visitors may alternatively choose to arrive at Greensboro Airport (about 2 and half hours from the conference by car) or one of the airports in the metropolitan Washington, D.C. area (about 4 and half hours). There is no public transportation available from either airports, but those who choose to rent cars may find the excursion enjoyable. Roanoke is located in southwestern Virginia at the intersection of Interstate 81 (milepost 144) and US 220. Driving from the North, the route passes through the scenic Shenandoah Valley of Virginia, and includes such notable sites as Thomas Jefferson's Monticello, Natural Bridge, and the town of Lexington. The route southwest of Roanoke passes through the spectacular scenery of the Blue Ridge Mountains. Northwest US 460 follows the New River out of West Virginia and passes through Blacksburg, home of Virginia Tech, one of the conference sponsors.Travelers with leisure time are also encouraged to explore the Blue Ridge Parkway, America's premier scenic byway. Roanoke is at milepost 120 on the Parkway, between the Peaks of Otter and Mabry Mill.

- Reduced Airfare with Delta Airlines
- Reduced Airfare with US Air

Lodging

The Hotel Roanoke & Conference Center

(A Doubletree Hotel; see http://www.doubletreehotels.com/)
110 Shenandoah Avenue, Roanoke, VA 24016
Phone: 540-985-5900 Fax: 540-853-8290
1-800-222-TREE

Single or double rooms cost $94/night plus tax as long as the conference rate is available (i.e., until room block is exhausted or May 15, whichever comes first), if you mention "JCDL 2001". Hotel rooms on the "Club Floor" have individual Ethernet ports; if you are interested in one of these, please inform the Hotel when making your reservation. These are available on a first-come-first -served basis, so register as early as possible.

Additional lodging is available in the area, but only at some distance from the conference hotel. Hotel Roanoke runs a free shuttle to and from the Roanoke airport, leaving every half hour on the quarter hour.

The Patrick Henry Hotel

617 S. Jefferson Street • Downtown Roanoke, VA

www.patrickhenryroanoke.com

For reservations call 1-800-303-0988, Monday -Friday 8:30 AM - 5:00 PM EST

JCDL '01 welcomes the historic Patrick Henry Hotel as our second official lodging provider. From the caryatids decorating the overhung second floor to the oversized rooms with functioning windows, the Patrick Henry has the charm of a grand hotel in a convenient location about 8 blocks away from the Hotel Roanoke & Conference Center, along well-lit and level streets through the center of downtown Roanoke.

Rooms are available at conference rate from Sunday through Thursday nights: single rooms, $79/night; doubles, $89/night. Kings, suites and mini-suites are also available. Most rooms have kitchenettes. Complimentary transportation is provided to and from the Roanoke Airport. Please inform the reservations staff that you are with the "Joint Conference on Digital Libraries."

Holiday Inn Express

Gainesboro Rd. at Orange Ave.
Roanoke, VA 24016
phone: (540) 982-0100
fax: (540) 345-4551

Due to unprecedented attendance, JCDL has expanded to a third hotel. The Holiday Inn Express is a recently constructed facility within walking distance of the Hotel Roanoke. All rooms have inside access. Doubles and king rooms are available at a conference rate of $68 / night, including a complimentary breakfast bar. All rooms have coffee makers, hair dryers and irons; some have microwaves and refrigerators. **To make reservations, call (540) 982-0100 and mention JCDL.**

Directions: From I-81 or US 220 take I-581 into Roanoke. Get off at Exit 4W, take an immediate left onto Gainesboro Road and a right into the Holiday Inn Express parking lot. Air travelers: the Express has no shuttle. Limousine and taxi service are available at the airport, but please check this page before you leave in case we can arrange something better.

KEY CONFERENCE COMMITTEE MEMBERS			
General Chair	**Edward Fox**	Dept. of Computer Science, M/C 0106, Virginia Tech, Blacksburg, VA, 24061 USA	fox@vt.edu phone: +1 540 231 5113 fax: +1 540 231 6075
Program Chair	**Christine Borgman**	Dept. of Information Studies, UCLA, Los Angeles, CA, 90095 1520 USA	cborgman@ucla.edu phone: +1 310 825 6164 fax: +1 310 206 4460
Treasurer	**Neil Rowe**	Computer Science Dept., Naval Postgraduate School, Monterey, California, USA	rowe@cs.nps.navy.mil phone: +1 831 656 2462 fax: +1 831 656 2814
Student Volunteers Chair	**Ghaleb Abdulla**	Lawrence Livermore National Lab, USA	abdulla1@llnl.gov
Workshops Chair	**Marianne Afifi**	Center for Scholarly Technology Information Services Division University of Southern California Los Angeles, CA 90089-0182	afifi@usc.edu phone: +1 213 740 8817 fax: +1 213 740 7713
Webmaster	**Fernando Das-Neves**	Dept. of Computer Science, M/C 0106, Virginia Tech, Blacksburg, VA, 24061 USA	fdasneve@vt.edu phone: +1 540 231 2060
Local Arrangements	**Robert K. France**	Digital Libraries Research Lab, 2030 Torgersen Hall, Blacksburg, VA 24061-0368 USA	france@vt.edu phone: +1 540 231 6256
Registration	**Jim French**	University of Virginia	french@cs.virginia.edu
Tutorials Chair	**Jonathan Furner**	Information Studies, UCLA, Los Angeles, CA USA	jfurner@ucla.edu
Panel Chair	**Gene Golovchinsky**	FX Palo Alto Laboratory, Inc. 3400 Hillview Ave., Bldg. 4 Palo Alto, CA 94304 USA	gene@pal.xerox.com phone: +1 650 813 7361 fax: +1 650 813 7081

Networking Coordinator	**Unmil Karadkar**	Texas A&M University	unmil@csdl.tamu.edu phone: +1 979 845 4924
Demos Chair	**Ray Larson**	University of California at Berkeley, California, USA	ray@sims.berkeley.edu
NSDL Support	**Brandon Muramatsu**	University of California at Berkeley, California, USA	mura@smete.org
Posters Chair	**Craig Nevill-Manning**	Dept. of Computer Science, Rutgers U., Piscataway, New Jersey, USA	nevill@cs.rutgers.edu phone: +1 732 445 2379 fax: +1 732 445 0537
Sponsoring and Exhibiting	**Michael L. Nelson**	School of Information and Library Science, University of North Carolina, North Carolina, USA	**mln@ils.unc.edu** phone: +1 919 966 5042 fax:+1 919 962 8071
Publicity	**Edie Rasmussen**	School of Information Sciences University of Pittsburgh Pittsburgh, PA 15260	erasmus@mail.sis.pitt.edu phone: +1 412 624 9459 fax: +1 412 648 7001
A/V Coordinator	**Luis Francisco-Revilla**	Texas A&M University, Texas, USA	l0f0954@csdl.tamu.edu

Program Committee

Caroline Arms, Library of Congress, USA
William Arms, Cornell Univ., USA
Nicholas Belkin, Rutgers Univ., USA
Jose Luis Borbinha, Biblioteca Nacional, PT
Daniel Brickley, Univ. of Bristol, UK
Ching-Chih Chen, Simmons College
Hsinchun Chen, Univ. of Arizona, USA
Su-Shing Chen, Univ. of Missouri, USA
Sayeed Choudhury, Johns Hopkins, USA
Greg Crane, Tufts Univ., USA
Lorcan Dempsey, DNER, UK
Andrew Dillon, Indiana Univ., USA
Alison Druin, Univ. of Maryland, USA
Dale Flecker, Harvard Univ., USA
Edward Fox, Virginia Tech, USA
Robert France, Virginia Tech, USA
James Frew, UCSB, USA
Jonathan Furner, UCLA, USA
Richard Furuta, Texas A&M Univ, USA.
Hector Garcia-Molina, Stanford Univ., USA
Anne Gilliland-Swetland, UCLA, USA
Gene Golovchinsky, FXPAL, USA
Vasileios Hatzivassikoglou, Columbia Univ, USA.
Anne Kenney, Cornell Univ., USA
Judith Klavans, Columbia Univ., USA
Traugott Koch, Lund Univ. & DTV
Carl Lagoze, Cornell Univ., USA
Louis Gomez, Northwestern Univ., USA
Ray Larson, UC, Berkeley, USA
Greg Leazer, UCLA, USA
David M. Levy, Univ. of Washington, USA
Ee-peng Lim, Nanyang Technological Univ., SG
Clifford Lynch, Coalition for Networked Info., USA
Gary Marchionini, Univ. of NC, USA
Catherine Marshall, FXPAL, USA
Cliff McKnight, Loughborough Univ., UK
Alexa McCray, Nat'l Library of Medicine, USA
Ellie Mylonas, Brown Univ., USA
John Ober, California Digital Library, USA
Edie Rasmussen, Univ. of Pittsburgh, USA
Joyce Ray, IMLS, USA
Allen Renear, Univ. of Illinois at Urbana-Champaign, USA
John Richardson, UCLA, USA
Neil Rowe, Naval Postgraduate School, USA
Chris Rusbridge, Univ. of Glasgow, Scotland
Sherry Schmidt, Arizona State Univ.
Michael Seadle, Michigan St. Univ., USA
Dagobert Soergel, Univ. of Maryland, USA
Rebecca Wesley, Stanford Univ., USA
Robin Williams, IBM Almaden, USA

Conference Steering Committee

Nicholas Belkin, Chair, Rutgers University
Richard Furuta, Texas A&M University
Nabil R. Adam, Rutgers University
Erich J. Neuhold, GMD-IPSI
Gary Marchionini, University of North Carolina
Yelena Yesha, University of Maryland, Baltimore County
Sally Howe, National Coordination Office for IT R&D

JCDL Frequently Asked Questions

1. ***Q: What audio/visual equipment will be available?***
   A: There will be two of each of the following available:
   - VCR and TV
   - overhead projector and screen
   - 35 mm slide projector
   - microphone
   - flip chart stand and pad

   [redacted]

2. ***Q: Is there a website where I can check to see how many people have signed up for my workshop/tutorial/etc.?***
   A: Unfortunatley, there is no website accessible for such information. However, General Chair Edward Fox's assistant Chad Schennum (cschennu@vt.edu) will send out periodic e-mails to the appropriate listservs with updates.

3. ***Q: Will there be internet access in the room where I am giving my workshop/tutorial/etc.?***
   A: Yes indeed, there is T1 access in all meeting rooms being used for JCDL.

Attendees For JCDL 2001
The First ACM+IEEE Joint Conference on Digital Libraries

- **Number of Attendees Listed: 404**
- **Total includes all those who indicated "Yes" to being included in the Attendee List.**
- **The listings excludes any duplicate (or multiple) registrations made under the same client number and contact information.**
- **The list does not include individuals who purchased additional banquet tickets only.**

Karim B. Boughida
The Getty Center
Getty Research Institute
1200 Getty Center Drive
Los Angeles CA 90049
USA
Phone: 310-440-7411
kboughida@getty.edu

Shawn Bowers
Computer Science Department
Oregon Graduate Institute
20000 NW Walker Road
Beaverton OR 97006
USA
Phone: 503-748-7068
Fax: 503-748-1553
shawn@cse.ogi.edu

Brian Bruya
2100 Date St., #2405
Honolulu HI 96826
USA
Phone: 808-956-6030
Fax: 808-956-9228
bruya@hawaii.edu

Sam Byrd
3022 Crossfield Rd.

Richmond VA 23233
USA
Phone: 804-692-3761
Fax: 804-692-3603
jmtaylor@lva.lib.va.us

Chin-Wan Chung
373-1 Kusong-dong, Yusong-gu
Taejon
South Korea
Phone: 82-2-958-3316
chungcw@cs.kaist.ac.kr

Glenn Courson
316 N. Tilden St. Apt. C

Richmond VA 23221
USA
Phone: 804-692-3761
Fax: 804-692-3603
jmtaylor@lva.lib.va.us

Lois Delcambre
Computer Science Department
Oregon Graduate Institute
20000 NW Walker Road

Thomas R Elliott
2 Mountain Lake Court
Durham NC 27713
USA

Susan Gibbons
Rush Rhees Library
University of Rochester
Rochester NY 14627

Beaverton OR 97006
USA
Phone: 503-690-1689
Fax: 503-690-1553
lmd@cse.ogi.edu

Phone: 919-962-0502
tom_elliott@unc.edu

USA
Phone: 716-275-6320
Fax: 716-273-1032
sgibbons@rcl.lib.rochester.edu

Paul Gorman
3181 SW Sam Jackson Park Rd.
Portland OR 97201
USA
Phone: 503-494-4025
Fax: 503-494-4551
GORMANP@OHSU.EDU

Carol Hansen - Montgomery, Ph.D
Drexel University
33rd and Market Streets
Philadelphia PA 191042875
USA
Phone: 215-895-2750
Fax: 215-895-2070
montgoch@drexel.edu

Corey A Harper
1100 W Hwy 54 Bypass Apt 40
Chapel Hill NC 27516
USA
Phone: (919) 942-0688
harpc@ils.unc.edu

Chern Koh
Room B31
International House of Philadelphia
3701 Chestnut Street
Philadelphia PA 19104
USA
Phone: 215-823-2885
chernkoh@drexel.edu

Paul Marotta
481 Eighth Avenue #835

New York NY 10001
USA
pmarotta@newworldrecords.org

John Millard
206 King Library
Oxford OH 45056
USA
millarj@lib.muohio.edu

Katy Newton
4704 Calvert Rd., #3
College Park MD 20740
USA
Phone: 301-454-0910
katyn@glue.umd.edu

Ashwini Pande
1207 Snyder Lane
Apt. 1200 A
Blacksburg VA 24060
USA
aspande@vt.edu

Anne Ramsden
The Open University
Walton Hall
Milton Keynes
Buckinghamshire MK7 6AA
Afghanistan

Phone: +01908-858563
Fax: +01908-653571
a.ramsden@open.ac.uk

Jeremy Rowe
2120 S. Las Palmas Circle
Mesa AZ 85202
USA
Phone: 480-965-8622
Fax: 480-965-8698
jeremy.rowe@asu.edu

Brent Seales
1265 Clear Creek Road
Nicholasville KY 40356
USA
seales@dcs.uky.edu

Frank A. Settle
Department of Chemistry
Washington & Lee University

Lexington VA 24450
USA
Phone: 540-463-8616
Fax: 540-463-8878
fsettle@wlu.edu

Chris Sullivan
Dept of Chemistry
Washington & Lee Univ
Lexington VA 24450
USA
Phone: 540-463-8616
fsettle@wlu.edu

Raj Sunderraman
Dept. of Computer Science
Georgia State University
Atlanta GA 30303
USA
Phone: 404-463-9716
Fax: 404-651-2246
raj@cs.gsu.edu

Jean Marie Taylor
105 Queen Mary Court

Williamsburg VA 23188
USA
Phone: 804-692-3761
Fax: 804-692-3603
jmtaylor@lva.lib.va.us

Victoria Uren
KMI, The Open University

Milton Keynes MK7 6AA
Afghanistan
Phone: +44-908-65-8516
Fax: +44-1908-65-31-69

Donald J Waters
140 E 62nd Street
New York NY 10021
USA
Phone: 212-838-8400
djw@mellon.org

Paul Zarins
123J SSRC
Green Library Program Officer

Stanford CA 943056004
USA
Phone: 650-725-1028

.s.uren@open.ac.uk

Fax: 650-723-9348
pzarins@stanford.edu

Heidi Neoma Abbey
University of Connecticut Libraries
Thomas J. Dodd Research Center
405 Babbidge Road, Unit 1205
Storrs CT 06269
USA
Phone: 860.486.2993
Fax: 860.486.4521
heidi.abbey@uconn.edu

Ghaleb M. Abdulla
124 Lamonte Ln
Tracy CA 95377
USA
Phone: (409) 238-7554
abdulla1@llnl.gov

Marianne Afifi
10917 Barman Avenue
Culver City CA 90230
USA
Phone: 213 740 8817
Fax: 213 740 8460
afifi@usc.edu

Ralph Alberico
Carrier Library
James Madison University, MSC 1703
Harrisonburg VA 22801
USA
Phone: 540 568-3828
Fax: 540 568-6339
alberira@jmu.edu

Margaret C. Alessi
Library of Congress
101 Independence Ave SE
Washington DC 20540
USA
Phone: 202-707-7953
maal@loc.gov

Suzanne Lorraine Allard
2666 La Grange Road
Shelbyville KY 40065
USA
slalla0@pop.uky.edu

Robert B Allen
4121 E. Hornbake Library
Clis
University Of Maryland
College Park MD 20742
USA
Phone: 301-405-2052
Fax: 973-829-5981

Colin Allen
422 Mission Road
Santa Fe NM 87501
USA
Phone: 505 982 9783
colin-allen@tamu.edu

Miriam H. Allman
Tufts University
37 Beacon St. #52
Boston MA 02108
USA
Phone: 617-627-5455
Fax: 617-627-3002
miriam.allman@tufts.edu

a@glue.umd.edu

Micah Altman
G-4 LiHauer Center, North Yard
Cambridge MA 02138
USA
Phone: 617-496-3847
Fax: 509-562-0658
maltman@data.fas.harvard.edu

Teal Anderson
3601 Greenway Unit B1
Baltimore MD 21218
USA
Phone: 410-516-7745
Fax: 410-516-5170
teal@jhu.edu

Martha Anderson
6512 Sara Alyce Ct
Burke VA 22015
USA
Phone: 202-707-2598
mande@loc.gov

Jeff Anderson-Lee
UC Berkeley, Dept. of Computer Science,
387 Soda Hall #1776
Berkeley CA 94720
USA
Phone: 510-643-6447
Fax: 510-643-5775
jonah@eecs.berkeley.edu

Kenning Arlitsch
295 S. 1500 East, Rm. 468
Marriott Library
University of Utah
Salt Lake City UT 84112
USA
Phone: (801) 585-3721
Fax: (801) 585-3464
karlitsc@library.utah.edu

Caroline Arms
166 Duke of Gloucester St.
Annapolis MD 21401
USA
Phone: (202) 707-0105
Fax: (202) 707-0955
caar@loc.gov

William Y Arms
Cornell University, 4130 Upson Hall
Ithaca NY 14853
USA
Phone: 607-255-3046
Fax: 607-255-4428
wya@cs.cornell.edu

John Aubry
Central Park West at 79th St.
New York NY 10024
USA
Phone: 212 313-7413
Fax: 212 769-5009
jaubry@amnh.org

David Bainbridge
17 Ernest Road
Hamilton
New Zealand
davidb@cs.waikato.ac.nz

Vipul Bansal
1835 WillowTree Ln.
Apt 6B8
Ann Arbor MI 48105
USA
bansal@umich.edu

Bruce R. Barkstrom
21 Langley Blvd.,Mail Stop 420

Hampton VA 236812199
USA
Phone: 757-864-5676
Fax: 757-864-7996
b.r.barkstrom@larc.nasa.gov

kobus Barnard
University of California, Berkeley
Computer Science Division
387 Soda Hall
Berkeley CA 947201101
USA
Phone: (510) 642-4979
Fax: (510) 643-4816
kobus@cs.berkeley.edu

Michael J. Bass
Hewlett-Packard Company
Building 10-500 MIT
77 Massachusetts Avenue
Cambridge MA 02139
USA
Phone: 617.253.6617
Fax: 617.452.3000
bass@alum.mit.edu

Nicholas J. Belkin
4 Devoe Street
South River NJ 08882
USA
Phone: 732-932-8585
Fax: 732 932 6916
nick@belkin.rutgers.edu

Mamie J Bell
CDC
1600 Clifton Rd, NE MS-C04
Atlanta GA 30333
USA
Phone: 404-639-1598
Fax: 404-639-1160
mbell1@cdc.gov

Jezekiel BenArie
400 E. Randolph Drv.,Apt. #3322
Chicago IL 60601
USA
Phone: 312-996-2648
benarie@eecs.uic.edu

Nuala A. Bennett
University of Illinois at Urbana-Champaign
Urbana IL 61801
USA
Phone: 217-333-9048
Fax: 217-244-7764
nabennet@uiuc.edu

Gerry Bernbom
601 E. Kirkwood Ave
Bloomington IN 47405
USA
Phone: 812-855-9220
Fax: 812-855-3310
bernbom@indiana.edu

Alejandro Gabriel Bia
Colonia Romana 13

Ann Blandford
School Of Computer Science

Barbara A. Blummer
17100 Science Drive

1.1 Portal 2 7ºE

Alicante 03016
Spain
Phone: +34-600948601
Fax: +34-965909477
abia@dlsi.ua.es

Middlesex University
Bounds Green Road
London N11 2nq
United Kngdm
Phone: 44-208-362-6411
a.blandford@virgin.net

Bowie MD 20715
USA
Phone: 301-805-7539
bablumm@super.org

Christine L Borgman
2845 Medill Place
Los Angeles CA 90064
USA
Phone: 310-825-6164
Fax: 310-838-8256
cborgman@ucla.edu

Katy Borner
Information Science & Cognitive Science
Indiana University, SLIS
10th Street & Jordan Avenue
Bloomington IN 47405
USA
Phone: 812 855 3256
Fax: 812 855 6166
katy@indiana.edu

Rachael Bower
2136 Kendall Ave.
Madison WI 53705
USA
Phone: 608-262-6587
bower@cs.wisc.edu

Kevin W. Boyack
Sandia National Laboratories
P.O. Box 5800, MS-0318
Albuquerque NM 87185
USA
Phone: 805-844-7556
Fax: 805-844-2415
kboyack@sandia.gov

Thomas Boyd
Technical Development Manager
UCAR/DLESE
PO Box 3000
Boulder CO 803073000
USA
Phone: 303-497-2650
tboyd@ucar.edu

Kristine Brancolini
3113 Daniel Street

Bloomington IN 47401
USA
Phone: 812/855-3710
Fax: 812/856-2062
brancoli@indiana.edu

Margret G. Branschofsky
Bldg. 10-500, MIT
Cambridge MA 02139

Anne Sigrid Bressler
Shield Library
100 N.W. Quad

Jeff Bridgers
University of Maryland, Mckeldin Library
B0223

SA
Phone: 617-253-1293
Fax: 617-452-3000
margretb@mit.edu

Davis CA 956165292
USA
Phone: 530-752-1202
Fax: 530-452-8785
asbressler@ucdavis.edu

College Park MD 20742
USA
Phone: 301-405-9194
jeffrey_bridgers@umail.umd.edu

James A. H. Brooks
CAB International
Nosworthy Way
Wallingford
Oxon OXIO 8DE
United Kngdm
Phone: 44-0-1491-829448
Fax: 44-0-1491-833508
j.brooks@cabi.org

Peter Brophy
20 Bishopdale Road
Lancaster
United Kngdm
Phone: +44-1524-382301
p.brophy@mmu.ac.uk

Elizabeth W. Brown
100 West University Pkwy #6A
Baltimore MD 21210
USA
Phone: 410-516-6834
Fax: 410-516-8684
ebrown@jhu.edu

Michael Brown
773 Anderson Hall
Dept of Computer Science
University of Kentucky
Lexington KY 40506
USA
mbrown@dcs.uky.edu

Ilvio Bruder
A.-Einstein-Str. 21
Rostock 18059
Germany
ilr@informatik.uni-rostock.de

George Robert Buchanan
Middlesex University
Bounds Green Road
London
United Kngdm
g.buchanan@mdx.ac.uk

Robin Douglas Burke
Information Systems & Decision
Sciences,CSU Fullenton

Fullerton CA 92834

Joseph A Busch
42 Bonview St
San Francisco CA 941105105
USA
Phone: 415-778-3129

Judith Elaine Bush
1200 Villa St
Mountain View CA 940411100
USA
Phone: 1 (650) 691 2308

SA
Phone: 714-278-5513
Fax: 714-278-5940
rburke@fullerton.edu

Fax: 415-778-3131
jbusch@interwoven.com

judith_bush@notes.rlg.org

Don Byrd
57 South Street
Williamsburg MA 01096
USA
Phone: 413-545-3147
dbyrd@cs.umass.edu

Guoray Cai
3085 Williamsburg Drive
State College PA 16801
USA
Phone: 814-865-4448
Fax: 814-865-5604
cai@ist.psu.edu

Nadia Caidi
663 Spadina Avenue #3

Toronto ON M5S2H9
Canada
Phone: 416-978-4664
Fax: 416-971-1399
caidi@fis.utoronto.ca

Pascal V Calarco
7 North Boulevard
Apt. #7
Richmond VA 23220
USA
Phone: 804-342-7494
Fax: 804-828-0151
pcalarco@erols.com

Juan Pablo Casares
5527 Ellsworth Ave. #306
Pittsburgh PA 15232
USA
Phone: 412-682-4851
Fax: 412-268-1266
juan.casares@cmu.edu

Hugh Cayless
405 Misty Grove Circle

Morrisville NC 27560
USA
Phone: 919-843-6260
hugh_cayless@unc.edu

Michael Chiu-Lung Chau
Artificial Intelligence Lab.
Mis Dept, Univ. Of Arizona
Mcclelland Hall 430
Tucson AZ 85721
USA
Phone: 510-621-6219

Ching-Chih Chen
300 The Fenway
Boston MA 02115
USA
Phone: 617-521-2804
Fax: 617-521-3192
chen@simmons.edu

Chaomei Chen
Dept. Of Info Sys & Comp.
Brunel University
Uxbridge Ub8 3ph
United Kngdm
Phone: 44-1895-203-080
Fax: 44-1895-251-686

Fax: 510-621-2433
mchau@ai.bpa.arizona.edu

chaomei.chen@brunel.ac.uk

Hsinchun Chen
The University of Arizona
Main Campus
P.O. Drawer 40370
Tucson AZ 857170370
USA
hchen@bpa.arizona.edu

Sayeed Choudhury
DKC, MSE Library
Johns Hopkins University
Baltimore MD 21218
USA
Phone: 410 516 4930
Fax: 410 516 5080
sayeed@jhu.edu

Michael G Christel
Carnegie Mellon University
Weh 4212

PITTSBURGH PA 15213
USA
Phone: 412-268-7799
Fax: 412-268-5576
christel@cs.cmu.edu

Greg Colati
7 Plymouth St.

Arlington MA 02476
USA
Phone: 617.627.3631
Fax: 617.627.3002
gregory.colati@tufts.edu

jack colbert
800 Memorial
griffin GA 30224
USA
Phone: 770.412.4770
colbertj@mail.spalding.public.lib.ga.us

Timothy W. Cole
216 Altgeld Hall
1409 West Green Street
Urbana IL 61801
USA
Phone: 217-244-7837
Fax: 217-244-4362
t-cole3@uiuc.edu

Brian Cooper
575 S. Rengstorff Ave. #117
Mountain View CA 94040
USA
Phone: 650-961-9365
Fax: 650-725-2588
cooperb@stanford.edu

James W Cooper
IBM TJ Watson Research Center
PO Box 704
Yorktown Heights NY 10598
USA
Phone: 914-784-7285
jwcnmr@watson.ibm.com

Robert J. Cordaro
2050 Ashmere Drive
Charlottesville VA 22902
USA
Phone: 804-243-8629
Fax: 804-924-1431
cordaro@virginia.edu

Anne Craig
Illinois State Library
300 South Second Street
Springfield IL 62701
USA
Phone: 217-785-5607
Fax: 217-557-6737
acraig@ilsos.net

Gregory R Crane
Tufts University
Dept. Of Classics
Eaton Hall 124
Medford MA 02155
USA
Phone: 617-627-3830
Fax: 617-627-3032
gcrane@tufts.edu

Martha Crawley
1100 Pennsylvania Ave. NW, Suite 802
Washington DC 20506
USA
Phone: 202-606-5513
Fax: 202-606-1077
mcrawley@imls.gov

Arturo Crespo
Gates Building. Room 420

Stanford, CA 94305
USA
Phone: 650-723-9273
Fax: 650-725-2588
crespo@cs.stanford.edu

Michael P D'Alessandro
210 Lexington Avenue
Iowa City IA 52246
USA
Phone: 319-356-3394
Fax: 319-356-2220
michael-dalessandro@uiowa.edu

Courtney S. Danforth
University of Maryland, Mckeldin Library
2M100E
College Park MD 20742
USA
Phone: 301-405-9063
cd147@umail.umd.edu

Sherry L. Davids
10301 Baltimore Ave.,Rm. 110
Beltsville MD 20705
USA
Phone: 301-504-5729
Fax: 301-504-5471
sdavids@nal.usda.gov

Laurie Davidson
5850 Shellmound Way
Emeryville CA 94608
USA
Phone: 510-655-6200 x2404
Fax: 510-450-6350
davidson@iii.com

Mary Wallace Davidson
620 S. Park Ave.
Bloomington IN 47401
USA
Phone: (812) 855-2972
Fax: (812) 855-3843
mdavidso@indiana.edu

obert Davies
2b Fitzgerald Road

London SW14 8HA
United Kngdm
Phone: +44 (0) 208 876 3121
Fax: +44 (0) 870 512910
rob.davies@mdrpartners.com

Russell Davis
7927 Jones Branch Drive
Suite 600S
McLean VA 22102
USA
Phone: 703-448-0087
mtrask@e-numerate.com

Lynne Davis
p.o.box3000
Boulder CO 80307
USA
Phone: 303-497-8313
lynne@ucar.edu

Melissa Dawe
University of Colorado
38 Pine Street
Salinas CA 93901
USA
Phone: 303-492-1592
melissa.dawe@colorado.edu

Repke De Vries
Grensstraat 23

1091 SW Amsterdam 1091 SW
Netherlands
repke.de.vries@niwi.knaw.nl

Douglas B Dearie
Highland Technologies, Inc.
4831 Walden Lane
Lanham MD 20706
USA
Phone: 301-345-8200
doug_d@htech.com

Kelly R. Debure
4200 54th Avenue South
St. Petersburg FL 33712
USA
Phone: 813-864-7749-4246
deburekr@eckerd.edu

John R Deller
Michigan State U.
Dept. Electrical & Computer Engr.
2120 Engineering Building
East Lansing MI 48824
USA
Phone: 5173538840
Fax: 5173531980
deller@msu.edu

Dave Deniman
4842 Brandon Creek Dr.

BOULDER CO 80301
USA
Phone: 303-530-3874
Fax: 303-530-3874
dave@deniman.com

Mark Derthick
Human Computer Interaction Institute
Carnegie Mellon University

Hadhami Dhraief
Kuelfweg 5
Hannover 30419

Timothy DiLauro
M.S.E. Library
3400 N. Charles Street

5000 Forbes Avenue
Pittsburgh PA 152133891
USA
Phone: 412-268-8812
Fax: 412 268 1266
mad@cs.cmu.edu

Germany
Phone: +49 511 762 19714
Fax: +49 511 762 19712
dhraief@kbs.uni-hannover.de

Baltimore MD 21218
USA
timmo@jhu.edu

Andrew Dillon
2711 E 10th Street
Bloomington IN 47408
USA
Phone: 812-855-5113
Fax: 812-855-6166
adillon@indiana.edu

Andy A. Long
UC Berkeley, 5138 Etcheverry Hall
Berkeley CA 947201740
USA
Phone: 510-643-1819
Fax: 510-643-1822
adong@me.berkeley.edu

Maureen H Donovan
2372 Lytham Rd
Columbus OH 43220
USA
Phone: 614-457-3494
Fax: 614-292-1918
donovan.1@osu.edu

Maureen H Donovan
2372 Lytham Road
Columbus OH 43220
USA
Phone: 614-292-3502
Fax: 614-292-1918
donovan.1@osu.edu

Robert L. Dougherty
Merck & Co., Inc.
WP 42-3
Sumneytown Pike
West Point PA 19486
USA
Phone: 215-652-5280
bob_dougherty@merck.com

Mary Claire Dougherty
University Library
1935 Sheridan Road

Evanston IL 60208
USA
Phone: 847-467-1437
Fax: 847-491-8306
m-dougherty@northwestern.edu

Matthew J. Dovey
65 St Giles
Oxford
United Kngdm
Phone: +44 1865 278272

Carol L. Dowling
10301 Baltimore Ave.,NAL Bldg.,Rm. 011
Beltsville MD 20705
USA
Phone: 301-504-5178

Jessica Draper
3800 HBLL
Brigham Young University
Provo UT 84602
USA

Fax: +44 1865 278175
matthew.dovey@las.ox.ac.uk

Fax: 301-504-5213
cdowling@nal.usdagov

Phone: (801)378-1456
Fax: (801)378-8910
jessica_draper@byu.edu

Allison Druin
4106 Clagett Road
University Park MD 20782
USA
Phone: 301-405-7406
allisond@umiacs.umd.edu

Christopher M Dunavant
505B Milhurst
Blacksburg VA 24060
USA
Phone: 540-557-1200
Fax: 540-557-1210
dunavantc@vtls.com

Jon Dunn
Indiana University
Main Library E170
1320 E. 10th St.
Bloomington IN 47405
USA
Phone: 812-855-0953
Fax: 812-856-2062
jwd@indiana.edu

Shahram Ebadollahi
Columbia University,1312 S.W. Mudd, 500
West 120th Street,Mailbox DG
New York NY 10027
USA
Phone: 212-939-7155
Fax: 212-932-9421
se98@columbia.edu

Carol A. Ellerbeck
Soldier Field Road

Boston MA 02163
USA
Phone: 617-495-6745
Fax: 617-495-8948
cellerbeck@hbs.edu

Oren Etzioni
Department of Computer Science
University of Washington
Box 352350
Seattle WA 98195
USA
Phone: 206-685-3035
etzioni@cs.washington.edu

Daniel Faensen
Takustr. 9
Berlin
Germany
Phone: +49-30-83875121
Fax: +49-30-83875109
faensen@inf.fu-berlin.de

David Edward Fenske
3141 Chestnut St.
Philadelphia PA 19104
USA
Phone: 215-895-2475
Fax: 215-895-6378
fenske@drexel.edu

Serena Jardine Fenton
508 Yorktown Drive
Chapel Hill NC 27516
USA
Phone: (919) 968-1626
fents@ibiblio.org

Kristine Ferry
1802 Brookfield Dr.
Ann Arbor MI 48103
USA
ferryk@earthlink.net

Michael Fingerhut
IRCAM
1, place Igor Stravinsky
Paris 75004
France
Phone: 33-1-44-784853
mf@ircam.fr

Les J Finken
2710 Brookside Drive
Iowa City IA 522455407
USA
Phone: 319-335-5467
Fax: 319-335-5505
les-finken@uiowa.edu

Dale Flecker
Harvard University Library
1280 Massachusetts Ave, SUite 404

Cambridge MA 02138
USA
Phone: (617) 495-3724
Fax: (617) 495-0491
dale_flecker@harvard.edu

Edward A Fox
203 Craig Drive
Blacksburg VA 24060
USA
Phone: 540-231-5113
Fax: 540-231-6075
fox@fox.cs.vt.edu

Robert France
2502 Carolina Ave.

Roanoke VA 24014
USA
Phone: 540.231.6256
france@vt.edu

Luis Francisco-Revilla
306 Redmod Dr. #300
College Station TX 77840
USA
Phone: 1+(979)845-4924
l0f0954@csdl.tamu.edu

Paolo Frasconi
Dept of Systems & Comp Science
University of Florence
Via Di Santa Marta, 3
Florence I-50139
Italy
Phone: +39-055 479 6362
paolo@ieee.org

James C French
3044 Amberfield Trail
Charlottesville VA 22911
USA
Phone: 804-982-2213
Fax: 804-982-2214
french@cs.virginia.edu

James Frew
Computer Systems Laboratory
University Of California
At Santa Barbara
Santa Barbara CA 93106
USA
Phone: 805-893-7356
frew@ucsb.edu

Antje Fritz
c/o Foyer des Alpes
Rue des Alpes 17
CH-1211 Geneva
Geneva
Switzerland
Phone: 0041-22-7417742
Fax: 0041-22-7417705
a.fritz@dcaf.ch

Ichiro Fujinaga
1001 St. Paul, 7H
Baltimore MD 21202
USA
ich@jhu.edu

David Fulker
UCAR/Unidata & Dlese
P.O. Box 3000
Boulder CO 80307
USA
Phone: 303-497-8650
Fax: 303-497-8650
fulker@ucar.edu

Jonathan Furner
UCLA - GSE&IS
300 Young Drive North
Mail Box 951520
Los Angeles CA 90095
USA
Phone: (310) 825-5210
Fax: (310) 206-4460
j.furner@ucla.edu

Richard Furuta
Texas A & M University
Dept. Of Computer Science
3112 TAMU
College Station TX 778433112
USA
Phone: 979-845-3839
Fax: 979-847-8578
furuta@cs.tamu.edu

Joseph Futrelle
152 Computing Applications Building
605 E. Springfield Ave.
Champaign IL 61820
USA
Phone: (217) 265-0296
Fax: (217) 244-1987
futrelle@ncsa.uiuc.edu

Ralph Gabbard
3816 W Highway 231
Grencastle IN 46135
USA
Phone: 812-237-2580
Fax: 812-237-2567
libgabb@link2000.net

Hector Garcia-Molina
Stanford Univesity
Gates Computer Science
Stanford CA 94305
USA
Phone: 650-723-0685
Fax: 650-725-2588
hector@cs.stanford.edu

Glenn R. Gardner

stewart garnger

Gary Geisler

Library of Congress
101 Independence Ave SE
Washington DC 20540
USA
Phone: 202-707-7414
ggar@loc.gov

Flat 19
Beatty Court
Southampton SO19 8RQ
United Kngdm
Phone: +44 23 80 366819
gkh12@dial.pipex.com

4321 Pin Oak Dr
Durham NC 27707
USA
geisg@ils.unc.edu

Sarah E Giersch
2000 Perimeter Park Drive
Suite 160
Morrisville NC 27560
USA
Phone: 919.276.3424
Fax: 919.462.6450
sgiersch@eduprise.com

Judy C. Gilmore
DOE/OSTI
175 Oak Ridge Turnpike
Oak Ridge TN 37830
USA
Phone: 865-576-5600
Fax: 865-576-9357
gilmorej@osti.gov

Katy M. Ginger
UCAR/DLESE
PO Box 3000

Boulder CO 80307
USA
Phone: 303-497-8341
ginger@ucar.edu

Charles Dudley Girard
2400 Ashland Rd. Apt A1
Columbia SC 29210
USA
girard@cse.sc.edu

Gene Golovchinsky
Fx Palo Alto Labratory
3400 Hillview Ave.
Bldg. 4
Palo Alto CA 94304
USA
Phone: 650-813-7361
Fax: 650-813-7081
gene@pal.xerox.com

Marcos Andre Goncalves
504 Cedar Orchard Dr W
Blacksburg VA 240609150
USA
mgoncalv@vt.edu

Rodney M. Goodman
California Institute of Tecnology
1200 E. California Blvd.
Pasadena, Ca. 91125 M/C 136-93

Laura Gottesman
Library of Congress
101 Independence Ave SE
Washington DC 20540

Rebecca A. Graham
3400 N. Charles St
MSE Library
Baltimore MD 21218

Pasadena CA 91125
USA
Phone: 626-395-2239
Fax: 626-585-8798
rogo@micro.caltech.edu

Stewart Granger
Flat 19
Beatty Court
Anson Drive
Sothampton SO19 8RQ
United Kngdm
Phone: +44 23 80 366819
stewartg@dial.pipex.com

Valerie Jane Gregg
7808 Accotink Place
Alexandria VA 22308
USA
Phone: 703-292-4768
Fax: 703-292-9030
vgregg@nsf.gov

Jeffery M. Griffith
Baker Library
Harvard Business School
Soldiers Field
Boston MA 02163
USA
Phone: 617-384-7188

USA
Phone: 202-707-0650
lgot@loc.gov

Cheryl Graunke
222 Chestertown Street

Gaithersburg MD 20878
USA
Phone: 202-707-3603
Fax: 202-707-0115
cgra@loc.gov

Robert S. Gresehover
11100 Johns Hopkins Road
Laurel MD 20723
USA
Phone: 443-778-4818
Fax: 443-778-6614
robert.gresehover@jhuapel.edu

Benjamin M Gross
2342 Shattuck Avenue
#510

BERKELEY CA 94704
USA
Phone: 415-358-4302

USA
Phone: 410-516-8781
Fax: 410-516-5080
rgraham@jhu.edu

Maayan Greffet
Hebrew University of Jerusalem
Sirkis St. 8/8
Jerusalem 95436
Israel
Phone: 053-236772
mary@cs.huji.ac.ie

Stephen Griffin
National Science Foundation
4201 Wilson Blvd, Suite 1115
Arlington VA 22230
USA
Phone: 703-292-8930
sgriffin@nsf.gov

Abigail M. Grotke
Library of Congress
101 Independence Ave SE
Washington DC 20540
USA
Phone: 202-707-2833
abgr@loc.gov

Fax: 617-495-8948
jgriffith@hbs.edu

Aparna Gurijala
1551 I Sparpan Village
East Lansing MI 48823
USA
Phone: 517-355-3148
gurijala@egr.msu.edu

Julia Hamilton
51 Luzerne St
Rochester NY 14620
USA
Phone: (716) 722-2449
julia.hamilton@kodak.com

Amy Marie Hartson
410 S. Wilmington Street
Raleigh NC 27601
USA
Phone: (919) 546-6749
amy.hartson@pgnmail.com

bgross-acm@bgross.com

Jin Hyee Ha
Sookmyung Women's University,53-12
Cheongpadong 2ka,Yongsanku
Seoul 140-742
South Korea
Phone: 82-2-710-9371
Fax: 82-2-710-9276
jinhyee@sookmyung.ac.kr

Marcia Hanna
8725 JJ Kingman Rd
Suite 0944
Ft. Belvoir VA 220606218
USA
Phone: 703-767-8062
mhanna@dtic.mil

Alex Hauptmann
339 Stone Church Rd
Finleyville PA 15332
USA
alex@cs.cmu.edu

James Leonard Halliday
Indiana University,1201 E. 3rd Street,
Music Library
Bloomington IN 47405
USA
Phone: 812-855-8804
Fax: 812-855-3843
jhallida@indiana.edu

Corey A Harper
1100 W Hwy 54 Bypass Apt 40
Chapel Hill NC 27516
USA
Phone: (919) 942-0688
harpc@ils.unc.edu

Margaret Hedstrom
1415 Dixboro Road
Ann Arbor MI 48105
USA
Phone: 734-332-0031
Fax: 734-764-2475
hedstrom@umich.edu

Normandy Helmer
33925 Seavey Loop
Eugene OR 97405
USA
Phone: 541-346-1864
Fax: 541-346-1882
nhelmer@darkwing.uoregon.edu

Harriette Hemmasi
Indiana University Library
1320 East 10th Street
Bloomington IN 47405
USA
Phone: 812-855-3403
hhemmasi@indiana.edu

Geneva Henry
Rice University
Fondren Library -- Ms 44
Po Box 1892
HOUSTON TX 772511892
USA
Phone: 713-348-2480
Fax: 713.348.5699
ghenry@rice.edu

Steven Hensen
Box 90185
Duke University Library
Duke University
Durham NC 27708
USA
Phone: 919.660.5826
Fax: 919.660-5934
hensen@duke.edu

Bob Henshaw
UNC-CH
CB# 3420
05 Smith Hall
Chapel Hill NC 27599
USA
bhenshaw@email.unc.edu

Heather Hessel
9054 Carson Street
Apt. A

CULVER CITY CA 90232
USA
hhessel@ucla.edu

Linda L Hill
499 Mills Way
Goleta CA 93117
USA
Phone: 805-893-8587
Fax: 805-893-3045
lhill@alexandria.ucsb.edu

Christopher Dale Hodge
2330 Dunford Hall
UNiversity of Tennessee
Knoxville TN 37996
USA
Phone: (865) 974-7505
Fax: (865) 974-8655
hodge@tns.utk.edu

Gail M. Hodge
312 Walnut Place
Havertown PA 19083
USA
Phone: (610)789-6769
Fax: (865)81-0390
gailhodge@aol.com

Nancy J. Hoebelheinrich

Douglas Holland

Jen-Shin Hong

anford University, Cataloging
Dept.,Meyer Library,3rd Fl.,Mail Code:6004
Stanford CA 943056004
USA
Phone: 650-725-6843
nhoebel@stanford.edu

Missouri Botanical Garden
P.O. Box 299
St. Louis MO 63166
USA
Phone: 314-577-5158
Fax: 314-577-9590
doug.holland@mobot.org

#1 University Road, PULI, Nantao 545
Department of Computer Engineering
National ChiNan University
PULI
Taiwan Roc
jshong@csie.ncnu.edu.tw

Leah Houser
6565 Frantz Road

Dublin OH 43017
USA
Phone: 614-764-6000
Fax: 614-718-7627
leah_houser@oclc.org

Sally Howe
National Coordination Office
4201 Wilson Blvd Suite 690
Arlington VA 222300001
USA
Phone: 703-306-4722
Fax: 703-306-4727
howe@ccic.gov

Jeffrey C Huestis
467 S. Holmes Ave.
Apt. C
Kirkwood MO 63122
USA
Phone: 314-935-5951
Fax: 314-935-4045
huestis@library.wustl.edu

Beverly Hunter
130 Mossie Lane
Amissville VA 20106
USA
Phone: 540 937-4038
Fax: 540 9377892
bev@piedmontresearch.org

Mary Ide
WGBH Media Archives and Preservation
Center, 125 Western Ave.
Boston MA 02134
USA
Phone: 617-300-2368
mary_ide@wgbh.org

Eiji Ikoma
4-6-1, Komaba
Meguro, Tokyo 153-8505
Japan
Phone: 81-3-5452-6256
Fax: 81-3-5452-6457
eikoma@tkl.iis.u-tokyo.ac.jp

Alan Inouye
National Research Council
Computer Science and Telecommunications
Board
2101 Constitution Ave., N.W. (Harris

Panagiotis G Ipeirotis
Columbia University
Department Of Computer Science
1214 Amsterdam Avenue
NEW YORK NY 100277003

Eric Isaacson
School of Music
Indiana University

Bloomington IN 47405

uilding 560)
Washington DC 20418
USA
Phone: 202-334-2849
Fax: 202-334-2318
ainouye@nas.edu

USA
Phone: +1 (212) 939-7117
Fax: +1 (212) 666-0140
pirot@cs.columbia.edu

USA
Phone: 812-855-0296
Fax: 812-856-4170
isaacso@indiana.edu

peter jacso
322 Aoloa Street 709
Kailua HI 96734
USA
Phone: 808 956-5817
jacso@hawaii.edu

peter jacso
322 aoloa street 709
kailua HI 96734
USA
jacso@hawaii.edu

Mathew C. Jadud
Indiana University,Music Library, 1201 E.
3rd St.
Bloomington IN 47405
USA
Phone: 812-856-0026
mjadud@indiana.edu

Greg Janee
CS/Alexandria Digital Library
UC Santa Barbara
Santa Barbara CA 93106
USA
Phone: 805-893-8453
Fax: 805-893-3045
gjanee@alexandria.ucsb.edu

Thomas Jevec
57 E. Delaware Pl. #2902
Chicago IL 60611
USA
Phone: 312-642-7802
Fax: 312-642-7832
tej@lucent.com

Ruth Ann Jones
100 Library
East Lansing MI 488241048
USA
Phone: 517-432-3977
Fax: 517-432-4795
jonesr@msu.edu

Steve Jones
University Of Waikato
Dept Computer Science
Private Bag 3105
Hamilton
New Zealand

Paul Jones
University of North Carolina
Campus Box 3456
Chapel Hill NC 27599
USA
Phone: 919-962-7600

Min-Yen Kan
414 W 121st Street, #38
New York NY 10027
USA
Phone: 212 939 7111
Fax: 212 939 7110

ıone: +64-7-838-4490
Fax: +64-7-858-5095
stevej@cs.waikato.ac.nz

pj@unc.edu

knmnyn@hotmail.com

John Kane
National Agriculture Library
Rm. 013 10307 Baltimore Blvd.
Beltsville MD 20705
USA
Phone: 301-504-6400
Fax: 301-504-7473
jkane@nal.usda.gov

ByungHoon Kang
University of California, Berkeley
Computer Science Division
493 Soda
Berkeley CA 947201101
USA
Phone: (510) 643-4816
Fax: (510) 643-1534
hoon@cs.berkeley.edu

Unmil P Karadkar
F-105 Front St. #B
College Station TX 77840
USA
Phone: 979-845-4924
unmil@cs.tamu.edu

Patricia Ann Keaton
6071 Magnol Lane
Woodland Hills CA 91367
USA
Phone: 8187159732
keaton@micro.caltech.edu

Judith Kelly
1865 W. Braod St.
Athens GA 30606
USA
Phone: 706-369-6277
Fax: 706-583-2636
judy_kelly@oit.peachnet.edu

Caitlin S. Kelly
Lucent Technology
600 Mountain Ave
Room 3A-434
Murray Hill NJ 07974
USA
Phone: 908-582-2186
cskelly@lucent.com

Karon Kelly
P.O. Box 3000
Boulder CO 80302
USA
Phone: 303-497-2652
Fax: 303-497-1170
kkelly@ucar.edu

Virginia Kerr
436 Prairie Avenue
Wilmette IL 60091
USA
Phone: (847) 491-7786
Fax: (847) 491-7786
vkerr@northwestern.edu

m. khoo
2235 Arapahoe Avenue
Boulder CO 80302
USA
michael.khoo@colorado.edu

Kevin S. Kiernan
627 Kastle Rd
Lexington KY 40502
USA
Phone: 859-266-3353
Fax: 859-266-8533
kiernan@pop.uky.edu

Sung-hyuk Kim
Sookmyung Women's University,53-12
Cheongpadong 2ka, Yongsanku
Seoul 140-742
South Korea
Phone: 82-2-710-9371
Fax: 82-2-710-9276
ksh@sookmyung.oc.kr

Kyung Ok Kim
Sookmyung Women's University,53-12
Cheongpadong 2ka,Yongsanku
Seoul 140-742
South Korea
Phone: 82-2-710-9371
Fax: 82-2-710-9276
miffi@dreamwiz.com

Richard James King
4555 Overlook Ave
Code 5220
Washington DC 20375
USA
Phone: 202-767-7515
Fax: 202-767-3352
james.king@nrl.navy.mil

Bruce R Kingma
4-206 Center for Science Tech
Syracuse University
Syracuse NY 13244
USA
brkingma@syr.edu

Amy J. Kirchhoff
JSTOR / Princeton University
221 Nassau Street
Princeton NJ 08544
USA
Phone: 609-258-4604
Fax: 609-258-5778
amykir@princeton.edu

Nobuko Kishi
Tsuda College
Kodaira-Shi
Tokyo 187-8577
Japan
Phone: +81-423-42-5160
Fax: +81-42-342-5161
kishi@tsuda.ac.jp

Christopher Klaus
5200 North Lamar Blvd.
Bldg. G 203
Austin TX 78751
USA
Phone: (512)453-4010
klaus@anl.gov

Judith L. Klavans
511 Butler Library, Mailcode 1103
535 West 114th Street
New York NY 10027
USA
Phone: 212-854-7443
klavans@cs.columbis.edu

Vickie Lynn Kline
Schmidt Library
York College of PA

York PA 174057199
USA
Phone: 717-815-1459
Fax: 717-849-1608
vkline@ycp.edu

Traugott Koch
Tornavaegen 9 B
Lund
Sweden
traugott.koch@ub2.lu.se

Gerd Kortemeyer
Michigan State University
123 North Kedzie Hall
East Lansing MI 48824
USA
Phone: 517-432-5468
Fax: 517-432-5653
korte@lite.msu.edu

Harald Krottmaier
Inffeldgasse 16c
Graz 8010
Austria
Phone: 43-316-8735631
Fax: 43-316-8735699
HKROTT@IICM.EDU

Harald Krottmaier
Inffeldgasse 16c
Graz
Austria
Phone: 43-316-8735631
Fax: 43-316-8735699
HKROTT@IICM.EDU

John A. Kunze
1325 Josephine St
Berkeley CA 94703
USA
Phone: 415-502-6660
jak@ckm.ucsf.edu

Carl J. Lagoze
Cornell University, 4130 Upson Hall
Ithaca NY 14853
USA
Phone: 607-255-6046
Fax: 607-255-4428
lagoze@cs.cornell.edu

Jean Laleuf
Box 1910
Brown University

PROVIDENCE RI 02912
USA
Phone: 401-863-7658
jrl@cs.brown.edu

Ann M. Lally
Eller College of Business and Public
Adm.-McClelland Hall 430
Tucson AZ 85721
USA
Phone: 520-621-6219
Fax: 520-621-2433
alally@bpa.arizona.edu

Cliff Lampe
2228 Stone Rd

Ronald L Larsen
6608 Woodstream Drive

Ray R Larson
2602 Beach Head Way

Ann Arbor MI 48105
USA
Phone: 734-936-3779
cacl@umich.edu

Chirstopher Lee
2456 Stone Rd.

Ann Arbor MI 48105
USA
calz@umich.edu

David M Levy
University of Washington
Suite 370, Mary Gates Hall

Seattle WA 981952840
USA
Phone: 206-616-2545
Fax: 206-616-3152
dmlevy@u.washington.edu

James B. Lloyd
Special Collections
Hoskins Library
Univ of Tenn
Knoxville TN 37996
USA

Seabrook MD 20706
USA
Phone: 301-405-2978
Fax: 301-314-2625
rlarsen@deans.umd.edu

Jacqueline Lesch
1200 E. Colton Ave.
Redlands CA 92373
USA
Phone: (909) 335-5268
Fax: (909) 307-6952
lesch@institute.redlands.edu

Elizabeth D. Liddy
Syracuse University
4-206 Center for Science & Technology
Syracuse NY 13244
USA
Phone: 315-443-5484
liddy@sur.edu

Raymond A Lorie
1272 Echo Valley Drive
San Jose CA 95120
USA
Phone: 408-927-1720
lorie@almaden.ibm.com

Richmond CA 94804
USA
Phone: 510-642-6046
ray@sherlock.berkeley.edu

Hilde Levine
600 Mountain Ave.
Room 3A-416
Murray Hill NJ 07974
USA
Phone: 9085824114
hlevine@lucent.com

Mingchun Liu
Information Systems Lab,School of
IEEE,Nanyang Technological
University,Nanyang Avenue
Singapore 639798
Singapore
Phone: 65-790-5639
p147508078@ntu.edu.sg

Karen Christina Lund
P.O. Box 235
Clinton MD 20735
USA
Phone: 202-707-0156
Fax: 202-707-3764

. .one: 865/974-4480
Fax: 865/974-0560
jlloyd@utk.edu

Marilyn R. Lutz
University of Maine, Fogler Library,
Orono ME 04469
USA
Phone: 207-581-1658
Fax: 207-581-1653
lutz@maine.edu

John MacColl
Darwin Library
The King's Buildings
Mayfield Road
Edinburgh
United Kngdm
Phone: +44 131 650 7275
Fax: +44 131 650 6702
john.maccoll@ed.ac.uk

Susan Manus
101 Independence Avenue, SE
LM 110

Washington DC 20540
USA

Michael R. Lyu
Computer Science Department
The Chinese University of Hong Kong

Hong Kong
Hong Kong
Phone: (+852)26098429
Fax: (+852)26035024
lyu@cse.cuhk.edu.hk

Elizabeth K. Madden
Library of Congress
101 Independence Ave SE
Washington DC 20540
USA
Phone: 202-707-4578
emad@loc.gov

Gary Marchionini
108 Hanover Pl

Chapel Hill NC 27516
USA
Phone: 919-966-3611

klun@loc.gov

David W MacCarn
WGBH Educational Foundation
125 Western Avenue
Boston MA 02134
USA
Phone: 617-492-2777
dave_maccarn@wgbh.org

Daniela M. Maestro
1919 M. Street NW, Suite 200
Washington DC 20036
USA
Phone: 202-912-1000 x470
Fax: 202-912-0772
dmaestro@conservation.org

Marcia Mardis
4251 Plymoth Rd.,Suite 2000
Ann Arbor MI 48105
USA
Phone: 734-764-9430
mmardis@merit.edu

Phone: 202-707-3741
Fax: 202-707-0621
sman@loc.gov

march@ils.unc.edu

Joseph Dennis Marek
355 Innisbrooke Ave.
Greenwood IN 461429216
USA
Phone: 317-887-3177
marek@sam.on-net.net

Richard Marisa
315 North Geneva Street
Ithaca NY 14850
USA
Phone: 607 255 7636
rjm2@cornell.edu

Mary Marlino
P.O. Box 3000
Boulder CO 80307
USA
Phone: 303-497-2656
Fax: 303-497-8336
marlino@ucar.edu

Cathy Marshall
856 Castro Street
San Francisco CA 94114
USA
Phone: 425-705-9057
cathymar@microsoft.com

Christie Mawhinney
8600 Rockville Pike
Bethesda MD 20894
USA
Phone: 301-496-9136
christie_mawhinney@nlm.nih.gov

Dave McArthur
2000 Perimeter Park Dr., Suite 160
Morrisville, NC 27560

Morrisville NC 27560
USA
Phone: 310-450-3493
dmcarthur@collegis.com

Sally Hart McCallum
Library of Congress
Network Development and MARC
Standards Office
Washington DC 20450
USA
Phone: 202-707-5119
Fax: 202-707-0115
smcc@loc.gov

Cavan McCarthy
School of Library and Information Science
Room 3087 Main Library
University of Iowa
IOWA CITY IA 52242
USA
Phone: (319) 335-5716
Fax: (319) 335-5374
cavan-mccarthy@uiowa.edu

Kathleen McKeown
CS Dept., Columbia University
1214 Amsterdam Avenue
New York NY 10027
USA
Phone: 212-939-7118
Fax: 212-666-0140
mckeown@cs.columbia.edu

Cliff McKnight
Dept of Information Science
Loughborough University
Ashby Road
Loughborough LE11 3TU
United Kngdm
Phone: +44-1509-223050
Fax: +44-1509-223053
c.mcknight@lboro.ac.uk

Daniel Patrick McShane
Alderman Library
P. O. Box 400108
Charlottesville VA 22904
USA
Phone: 804.924.3198
dpm5h@virginia.edu

Weiyi Meng
Dept. Of Computer Science
Thomas J. Watson School
P.O. Box 6000
Binghamton NY 139026000
USA
Phone: 607-777-4311
Fax: 607-777-4822
meng@cs.binghamton.edu

Fred A Miller
120 Wooden Shoe Court North
Christiansburg VA 240731275
USA
Phone: 540-951-8991
Fax: 540-557-1210
millert@vtls.com

John S. Miller
422 Anschutz Library
1301 Hoch Auditoria Drive
University of Kansas
Lawrence KS 66045
USA
Phone: 785-864-3894
Fax: 785-864-5380
jsmiller@ku.edu

Stephen Miller
Main Library 4th Floor
University of Georgia
Athens GA 30602
USA
sdmiller@uga.edu

David Millman
206 W 99 St
New York NY 10025
USA
Phone: 212-854-4284
Fax: 212-662-6442
dsm@columbia.edu

William E Moen
Po Box 311068
School Of Library & Info. Sci.
Univ. Of North Texas
Denton TX 76203
USA
Phone: 940-565-3563
Fax: 940-565-3101
wemoen@unt.edu

Norliza Mohd-Nasir
School of Computing Science
Middlesex University
Bounds Green Road
London N11 2NQ
United Kngdm
y.theng@mdx.ac.uk

Sally Anne Mubbard
1200 Getty Center Drive
Suite 1100
Los Angeles CA 900491688
USA
Phone: 310-440-6684
sshubbard@getty.edu

Adrienne Muir
Department of Information Science
Loughborough University
Loughborough
United Kngdm
Phone: +44 1509 223064
Fax: +44 1509 223053
a.muir@lboro.ac.uk

Sougata Mukherjea
2315 N First Street
San Jose CA 95131
USA
Phone: 408-570-8082
sougatam@bea.com

Brandon Muramatsu
3115 Etcheverry Hall
Berkeley CA 947201750
USA
Phone: 510-643-1817
Fax: 510-643-1822
mura@smete.org

Catherine Murray-Rust
121 Valley Library
Oregon State University
Corvallis OR 97330
USA
Phone: 541-737-8527
Fax: 541-737-3453
catherine.murray-rust@orst.edu

Elli Mylonas
Box 1841 - CIS
Brown University
Providence RI 02912
USA
Phone: 401-863-7231
Fax: 401-863-9313
elli_mylonas@brown.edu

Jocelyne NANARD
161 rue Ada

Montpellier 34192
France
Phone: 33-46-741-8516
jnanard@lirmm.fr

John Lewis Needham
45 East 30th Street
New York NY 10016
USA
Phone: 212- 532 -4717
needham@ebrary.com

Michael L. Nelson
MS 158
Hampton VA 23681
USA
Phone: 757-864-8511
m.l.nelson@larc.nasa.gov

Erich J Neuhold
Gmd-Ipsi
Dolivostr. 15
Darmstadt D-64293
Germany
Phone: 49-6151-869-802
Fax: 49-6151-869-969
neuhold@darmstadt.gmd.de

Michael Neuman
Georgetown University, 9520 Prospect St.
NW
Washington DC 20057
USA
Phone: 202-687-6283
Fax: 202-687-8367
neuman@georgetown.edu

Heike Dr. Neuroth
SUB
SSG-FI
Papendiek 14
37073 Goettingen
Germany
Phone: +49-(0)551 393866
neuroth@mail.sub.uni-goettingen.de

Craig Nevill-Manning
578a 30th Street
San Francisco CA 94131
USA
Phone: 415 845 7145
Fax: 801 760 7628
craig@nevill-manning.com

Katy Newton
4704 Calvert Rd., #3

College Park MD 20740
USA
Phone: 301-454-0910
katyn@glue.umd.edu

Glen Newton
Building M-55 Montreal Road
Ottawa ON KIA 052
Canada
Phone: 613-990-9163
Fax: 613-952-8246
glen.newton@nrc.ca

Lorraine Normore
OCLC, Office of Research
6565 Frantz Road
Dublin OH 43017
USA
Phone: 614-761-5263
Fax: 614-764-2344
normorel@oclc.org

Lucy Terry Nowell
40 Edgewood Drive
Richland WA 99352
USA
Phone: (509)946-5761
Fax: (509)375-3641
lucy.nowell@pnl.gov

John L. Ober
300 Lakeside Dr., 6th Floor
Oakland CA 94612
USA
john.ober@ucop.edu

Jia-Yu Pan
240 Melwood Ave. Apt. D1
Pittsburgh PA 15213
USA

Bill Parod
Northwestern University
1935 Sheridan Road
NU Library 2East

Joy Paulson
107 E State St
Apt. 306
Ithaca NY 14850

. .ione: 412-268-7123
jypan@cs.cmu.edu

Gordon Paynter
C/-Dept. of Computer Science
University of Waikato
Private Bag 3105
Hamilton
New Zealand
Phone: +64-7-838-4021
Fax: +64-7-838-4155
paynter@cs.waikato.ac.nz

William Robert Pendleton
Indiana University,1201 E. Third Street,
Music Library
Bloomington IN 47405
USA
Phone: 812-855-3472
Fax: 812-855-3843
wpendlet@indiana.edu

Karen E. Pettigrew
The Information School University of
Washington,Box 352840, Mary Gates Hall
Seattle WA 981952840
USA
Phone: 206-281-9277

Evanston IL 60208
USA
Phone: 847-491-5368
bill-parod@northwestern.edu

Doug Pearson
2711 East 10th Street
Bloomington IN 47408
USA
Phone: 812-855-3846
dodpears@indiana.edu

Saverio Perugini, Jr
123 Main Campbell Hall - #3
Virginia Tech
Blacksburg VA 24060
USA
Phone: 540-232-6318
sperugin@vt.edu

W Harry Plantinga
Dept. Of Comp Sci.
Calvin College
3201 Burton S.E.
Grand Rapids MI 49546
USA

USA
Phone: 607-255-7950
Fax: 607-255-0318
jp243@cornell.edu

Anselmo Peñas
Dpto. Lenguajes y Sistemas, UNED
Ciudad Universitaria, s/n
Madrid 28040
Spain
Phone: +34-91 398 7750
Fax: +34 91 398 6535
anselmo@lsi.uned.es

Michael J. Petro
367 Scarsdale Road
Tuckahoe NY 10707
USA
Phone: 732-562-3992
Fax: 732-981-9334
m.petro@ieee.org

Richard Pollard
School Of Info. Sci.
Univ. Of Tennessee, Knoxville
804 Volunteer Blvd.
Knoxville TN 379964330
USA

Fax: 206-616-3152
kpettigr@u.washington.edu

Phone: 616-957-6860
hplantin@calvin.edu

Phone: 865-974-8026
Fax: 865-974-4967
richard-pollard@utk.edu

James Powell
1300 Torgersen Hall
Virginia Tech
Blacksburg VA 24061
USA
Phone: 540-231-6927
Fax: 540-231-9265
jpowell@vt.edu

Allison L Powell
55-E Barclay Place Ct.
Charlottesville VA 22901
USA
Phone: 804-971-1181
alp4g@cs.virginia.edu

Cecilia M. Prewston
PO Box 8310

Emeryville CA 94662
USA
Phone: 510-547-3207
Fax: 510-658-4976
cecilia@well.com

Edie Rasmussen
135 North Bellefield Avenue
School of Information Sciences
University of Pittsburgh
Pittsburgh PA 15260
USA
Phone: 412-624-9459
Fax: 412-648-7001
erasmus@mail.sis.pitt.edu

Andreas Rauber
Tu Wien
Institut F. Softwaretechnik
Faveritenstr. 9-11/188
Vienna A-1040
Austria
Phone: +43-1-58801-4126
Fax: +43 1 5040532
rauber@ifs.tuwien.ac.at

Joyce Ray
1100 Pennsylvania Ave., NW Ste.802
Washington DC 20506
USA
Phone: 202-606-5384
Fax: 202-606-1077
jray@imls.gov

John Reuning
140 BPW Club Rd., Apt. D9
Carrboro NC 27510
USA

Sarah Reuning
140 BPW Club Road, Apt. D9
Carrboro NC 27510
USA

Glenn J. Ricci
Library of Congress
101 Independence Ave SE
Washington DC 20540

jreuning@ibiblio.org

Phone: 919-960-7022
sreuning@email.unc.edu

USA
Phone: 202-707-1831
gric@loc.gov

Rita C. Richardello
300 Day Hill Road

Windsor CT 06095
USA
Phone: 860-285-7763
Fax: 860-298-9555
rrichardello@1limra.com

William Ryan Richardson
430C Harding Avenue
Blacksburg VA 24060
USA
Phone: 540 200-1860
ryanr@vt.edu

Tracy L Riggs
EECS Department
493 Soda Hall
Berkeley CA 94720
USA
Phone: (510) 643-4816
Fax: (510) 643-1534
tracyr@cs.berkeley.edu

Michael Robertson
Rochester Institute of Technology,90 Lomb
Memorial Dr.(Wallace Library)
Rochester NY 146235604
USA
Phone: 716-475-2565
Fax: 716-475-7007
marwml@rit.edu

Lisa M. Robinson
Michigan State University
100 Main Library
East Lansing MI 488241048
USA
Phone: 517-432-1645
Fax: 517-432-4796
robin179@msu.edu

Kimberly S. Roempler
1929 Kenny Road

Columbus OH 43210
USA
Phone: 614-688-3485
Fax: 614-292-2066
roempler@enc.org

Perry D. Roland
University of Virginia,Digital Library R&D,
Alderman Library
Charlottesville VA 22904
USA
Phone: 804-982-2702
Fax: 804-924-1431
pdr4h@virginia.edu

Bruce Rosenstock
1104 Auburn Drive
Davis CA 95616
USA
bbrosenstock@ucdavis.edu

Timothy B. Rowe
Department of Geological Sciences
The Univesity of Texas
Austin TX 78712
USA
Phone: 512-232-5512
Fax: 512-471-9425
rowe@mail.utexas.edu

Neil C Rowe
Naval Postgraduate School
Code Cs/Rp
833 Dyer Road
Monterey CA 93943
USA
Phone: 831-656-2462
rowe@cs.nps.navy.mil

Mary Rowlatt
County Hall, Room BG 09/010
Duke Street
Chelmsford CM1 1 LX
United Kngdm
Phone: +44 (0)1245 436524
Fax: +44 (0) 1245 257634
maryr@essexcc.gov.uk

Hava Rubenson
1200 Villa Street
Mountain View CA 94041
USA
Phone: 650-691-2289
hava_rubenson@notes.rlg.org

Adam Russell
6219 Palma del Mar, Apt. 402
St. Petersburg FL 33715
USA
Phone: (727) 867 3118
russelas@eckerd.edu

Jeff Rydberg-Cox
5317 W. 80th Street
Prairie Village KS 66208
USA
Phone: 816-235-2560
Fax: 816-235-1308
rydbergcoxj@umkc.edu

Lynetta S. Sacherek
Oregon Health Sciences Univ
3181 SW Sam Jackson Park Road
Mail Code: BICC
Portland OR 972013098
USA
Phone: 503-494-0467
Fax: 503-494-4551
sacherek@ohsu.edu

Manuel Sanchéz
Apdo.Correos 99
Alicante 03080
Spain
Phone: +34-965909567
Fax: +34-965909477
manuel.sanchez@cervantesvirtual.com

Kurt Sanftleben
4928 Breeze Way
Montclair VA 22026
USA
sanftlebenka@tecom.usmc.mil

Chad Schennum
604 Cambridge Road
Blacksburg VA 24060
USA
cschennu@vt.edu

Francois Schiettecatte
F S Consulting Inc
326 N. Charles, Suite 300
Baltimore MD 21201
USA
Phone: 410-625-2080
Fax: 410-625-2081
francois@fsconsult.com

Rudi Schmiede
Inst. of Sociology
Residenzschloss
D-64283 Darmstadt
Germany
Phone: +49 6151 16 2809
Fax: +49 6151 16 6042
schmiede@ifs.tu-darmstadt.de

Michael S. Seadle
100 Library
Michigan State University
East Lansing MI 48824
USA
Phone: 517-432-0807
Fax: 517-432-4795
seadle@msu.edu

Jill Sexton
CB# 3918 Wilson Library
University of North Carolina at Chapel Hill

Chapel Hill NC 27514
USA
kuhn@metalab.unc.edu

Elizabeth J Shaw
Rm.626 IS Building,135 N. Bellefield
Pittsburgh PA 15260
USA
Phone: 412-624-9455
Fax: 412-648-7001
ejashaw+@pitt.edu

Tina Shrader
Arizona State University Libraries
PO Box 871006

Tempe AZ 85287
USA
Phone: 480-965-9806
Fax: 480-965-1043
tina.shrader@asu.edu

Jeffrey J Simon
600-700 Mountain Avenue
RM 3D-592
Murray Hill NJ 07974
USA
Phone: 908.582.3757
Fax: 908.582.5417
jjsimon@lucent.com

Andre Skupin
Department Of Geography
University Of New Orleans

New Orleans LA 70148
USA
askupin@uno.edu

Laura Smart
565 Gayley Avenue #605
Los Angeles CA 90024
USA
lsmart@ucla.edu

Alan Smeaton
School of Computer Applications

Lisa Smith
1005 Briarwood Blvd.

Gordon W. Smith
Calif State Univ

Dublin City University
Glasnevin
Dublin 9
Ireland
Phone: 353-1-7005262
asmeaton@compapp.dcu.ie

Arlington TX 76013
USA
Phone: 817-272-3731
Fax: 817-272-5797
smith@library.uta.edu

401 Golden Shore
Long Beach CA 90802
USA
Phone: 562-951-4263
Fax: 562-951-4925
gwsmith@calstate.edu

Dagobert Soergel
University Of Maryland
School of Information
Studiees
College Park MD 207420001
USA
Phone: 301-405-2037
Fax: 301-314-9145
d552@umail.umd.edu

Ingeborg Torvik Solvberg
IDI, Norwegian University of Science and
Technology
Trondheim 7491
Norway
Phone: +47-73-596027
Fax: 47-73-59-1733
ingeborg.solvberg@idi.ntnu.no

Shaung Song
6102 Etoheverry
Berkeley CA 94720
USA
Phone: 510-643-8146
Fax: 510-643-1822
shaung@newton.berkeley.edu

Von-Wun Soo
101 Section 2, Kuan Fu Road
Hsin Chu 30043
Taiwan Roc
Phone: 886-357-31068
Fax: 886-357-31068
soo@cs.nthu.edu.tw

Lisa Spiro
5430 Edith St.
Houston TX 77096
USA
Phone: (713) 348-2594
lspiro@rice.edu

Ed Sponsler
1200 E. California Blvd.
MS 1-43
Pasadena CA 91125
USA
Phone: 626-395-3401
Fax: 626-431-2681
eds@library.caltech.edu

Scott M. Stevens
508 Troutwood Dr.
Pittsburgh PA 15237
USA

Amy Stucki
Bringham Young University
2840 Hall

Hussein Suleman
331 Main Campbell Hall

Phone: 412-268-7796
Fax: 412-268-1266
sms@cs.cmu.edu

Provo UT 84601
USA
Phone: 801-378-9194
amy_stucki@byu.edu

Blacksburg VA 24060
USA
hussein@vt.edu

Kristen Summers
Highland Technologies,Inc.,4831 Walden Ln.
Lanham MD 20706
USA
Phone: 301-306-2827
Fax: 301-306-8201
ksummers@htech.com

Tamara Sumner
3370 34th Street #D
Boulder CO 80301
USA
Phone: 303-492-2233
Fax: 303-492-2844
sumner@colorado.edu

Tamara Sumner
Dept. of Computer Science
Campus Box 430
University of Colorado at Boulder
Boulder CO 80309
USA
Phone: 303-492-2233
Fax: 303-492-2844
sumner@colorado.edu

Yin Leng Theng
School of Computing Science
Middlesex University
Bounds Green Road
London N11 2NQ
United Kngdm
Phone: +44-0208-362-6926
Fax: +44-0208-362-6411
y.theng@mdx.ac.uk

Deborah M. Thomas
Library of Congress
101 Independence Ave SE
Washington DC 20540
USA
Phone: 202-707-5963
deth@loc.gov

Helen Ruth Tibbo
School of Information and Library Science
201 Manning Hall CB#3360
University of North Carolina at Chapel Hill
Chapel Hill NC 275993360
USA
Phone: 919.962.8063
Fax: 919.962.8071
tibbo@ils.unc.edu

Mary Elizabeth Tiles
555 University Avenue #1100
Honolulu HI 96826
USA
Phone: 808 9463634

Barbara B. Tillett
101 Independence Ave., SE
Washington DC 205404010
USA
Phone: 202-707-4714

Brian Tingle
California Digital Library
University of California, Office of the President
1111 Franklin St., 7th Floor

...tiles@hawaii.edu

Fax: 202-707-4719
btil@loc.gov

Oakland CA 94607
USA
Phone: 510-987-0443
brian.tingle@ucop.edu

Taku Tokuyasu
UC Berkeley
Computer Science Division
493 Soda Hall
Berkeley CA 947201776
USA
Phone: (510) 642-8468
tokuyasu@cs.berkeley.edu

Elaine G Toms
Faculty of Information Studies
University of Toronto
140 St. George St.
Toronto ON M5S 3G6
Canada
Phone: 902-494-2452
Fax: 902-494-2451
toms@fis.utoronto.ca

Carl Townsend
12310 Sunrise Valley Drive
M/S P59
Reston VA 20191
USA
Phone: 703-755-5654
townsenc@nima.mil

Anthony Troncale
77th @ Central Park West
New York NY 10024
USA
Phone: 212-769-5421
Fax: 212-769-5009
troncale@amnh.org

Mark I. Turner
Highland Technologies, Inc.
Lanham MD 20706
USA
Phone: 301-306-2826
Fax: 301-306-8201
mturner@htech.com

Raul Valdes-Perez
Computer Science Dept.,Carnegie Mellon
University
Pittsburgh PA 15213
USA
Phone: 412-268-7127
Fax: 412-268-5575
valdes@cs.cmu.edu

Ellen M Voorhees
Nist
100 Bureau Drive
Stop 8940
Gaithersburg MD 208998940
USA

Nina Wacholder
Rutgers SCILS
4 Huntington ST.

New Brunswick NJ 08901
USA

Howard Wactlar
Carnegie Mellon University
School of Computer Science
5000 Forbes Ave, Wean 5216
Pittsburgh PA 15213
USA

. .ione: 301-975-3761
Fax: 301-975-5287
ellen.voorhees@nist.gov

Fax: 7329326916
nina@scils.rutgers.edu

Phone: 412-268-2571
Fax: 412-268-7458
wactlar@cmu.edu

Stephanie J Wagaman
9 South Commerce Way
Allentown PA 180178916
USA
Phone: 610-758-8700
Fax: 610-758-9700
wagamans@oclc.org

Christine Ann Walker
500 Belmont Road
Bettendorf IA 52722
USA
Phone: 563-441-4095
Fax: 563-441-4080
cwalker@eiccd.cc.ia.us

Christine Ann Walker
500 Belmont Road
Bettendorf IA 52722
USA
Phone: 563-441-4095
Fax: 563-441-4080
cwalker@eiccd.cc.ia.us

Colleen Byrne Wallace
Library of Congress
101 Independence Ave SE
Washington DC 20540
USA
Phone: 202-707-3449
cowa@loc.gov

James Ze Wang
Penn State/Stanford
PNC Assistant Professor
School of IST, 504 Rider I
University Park PA 16801
USA
jwang@ist.psu.edu

Jewel Hope Ward
116 B Cheek St
Carrboro NC 27510
USA
Phone: 919-961-0782
wardj@ils.unc.edu

Charles Richard Ward
7500 Promontory Court
Wilmington NC 28412
USA
Phone: 919-392-6561
ward@uncwil.edu

Beth Warner
University of Kansas,1450 Jayhawk Blvd.
Lawrence KS 660457535
USA
Phone: 785-864-4999
bwarner@ku.edu

Walter Warnick
19901 Germantown RD
Germantown MD 20874
USA
Phone: 856-576-0644
warnick@science.doe.gov

eter J Wasilko
3 Meadowbrook Drive
Ossining NY 105622916
USA
Phone: 914-941-5705
futurist@cloud9.net

John Weatherley
506 Hapgood St
Boulder CO 80302
USA
jweather@ucar.edu

mischa weiss-lijn
11 charleston Street
London
United Kngdm
m.weiss-lijn@cs.ucl.ac.uk

Gary S Wesley
431 Gates
Stanford CA 94305
USA
gary@db.stanford.edu

Rebecca Wesley
Stanford University
3650 Ross Road
Palo Alto CA 94303
USA
rwesley@stanford.edu

Claudia V. Weston
10301 Baltimore Ave., Rm. 013
Beltsville MD 207052351
USA
Phone: 301-504-6358
Fax: 301-504-7473
cweston@nal.usda.gov

Paul Robert Wheatley
31 Bainbrigge Road,
Headingley
Leeds LS6 3AD
United Kngdm
Phone: 0113 2335830
p.r.wheatley@leeds.ac.uk

Robert Wilensky
University Of California
Eecs Computer Sci Diviison
719 Soda Hall #1776
Berkeley CA 947201776
USA
wilensky@cs.berkeley.edu

Connie Wiley
Defense Technical Info Center
8725 John J. Kingman Rd.
Ste. 0944
Ft. Belvoir VA 22060
USA
Phone: 703-767-9112
Fax: 703-767-9119
cwiley@dtic.mil

Perry Willett
520 S. Jordan
Bloomington IN 47401
USA
Phone: 812-335-9546

Diana W. Williams
2645 Mulberry Lane
Greenville NC 27858
USA
Phone: 252-328-2771

Robin Williams
IBM Almaden Research Center
650 Harry Road
San Jose CA 95120
USA

Fax: 812-855-8068
pwillett@indiana.edu

Fax: 252-328-4834
williamsdi@mail.ecu.edu

Bonita Wilson
1895 Preston White Dr., Ste 100
Reston VA 20191
USA
Phone: 703-620-8990
Fax: 703-620-0913
bwilson@cnri.reston.va.us

Ian Hugh Witten
P O Box 3105
Hamilton
New Zealand
ihw@cs.waikato.ac.nz

Michael Wright
3370 34th Street, Apt D.
Boulder CO 80301
USA
Phone: 303-497-8654
mwright@vcar.edu

Cheryl Wright
P.O. Box 4892
Incline Village NV 89450
USA
Phone: 650-786-5202
Fax: 775-832-6698
cheryl.wright@sun.com

Yejun Wu
3422 RUTGERS STREET
HYATTSVILLE MD 20783
USA
Phone: 301-405-2033
Fax: 301-314-9145
wuyj@wam.umd.edu

Jia-Long Wu
6102 Etoheverry Hall
Berkeley CA 94720
USA
Phone: 510-643-8146
Fax: 510-643-1822
jialong@newtown.berkeley.edu

David Yaron
Department of Chemistry
Carnegie Mellon University
4400 Fifth Ave.
Pittsburgh PA 15213
USA
Phone: 412-268-1351
Fax: 412-268-1061
yaron@chem.cmu.edu

Jeffrey Young
4805 Olentangy Blvd.
Columbus OH 43214
USA
jyoung@oclc.org

Cheng Jiun Yuan
700 Woodland Avenue
Apt #F-212
Lexington KY 405083473
USA
yuan@dcs.uky.edu

Lei Zeng
7736 Salem Circle
Hudson OH 44236
USA
Phone: 330-672-2782
Fax: 330-672-7965
zeng@slis.kent.edu

Lesli Zimmerman
8978 Fox Lake Drive
Knoxville TN 37923
USA
Phone: 865-974-9427
lesli@aztec.lib.utk.edu

U.S. Department of Education
Office of Educatonal Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE
(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title:	ACM/IEEE-CS Joint Conference on Digital Libraries 2001		
Authors:	ACM Publications		
Corporate Source:	ACM Publications	Publication Date:	2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announces in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reporduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options below and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY ACM Inc TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY,HHAS BEEN GRANTED BY TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
●	○	○
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only.	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only.

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproducation from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Name (Signature):	Deborah Cotton	Position/Title:	Copyright & Permissions
Organization/Address:	ACM Publications, 1515 Broadway, 17th fl., New York, NY 10036		
Telephone:	212-626-0652	FAX:	212-869-0481
E-MAIL Address:	cotton@hq.acm.org permissions@acm.org	Date:	11/30/01