This paper addresses several questions about digital libraries. What kinds of communities will digital library technology produce? The Web seems much more popular than electronic journals--does this mean that surfing will replace literature reading, and that "nerds" building HTML hierarchies will supplant publishers? Will this mean that the universities will lose control of the quality of what their students read? Will the ability to do more research in one's dorm room mean that students will not talk to one another at all, that they will talk to people somewhere else in the world, or that they will talk to their roommates more than ever, perhaps about how to use the computer system? Digital information threatens our ideas of locality: will the association of students with a particular university, let alone university library, survive the Web? Might online references and online multimedia lectures produce the 'virtual university of the United States' and if so would that be desirable? Universities serve a variety of social functions which the Web can augment or diminish, depending on people's actions. The Web also may threaten ideas of quality in scholarship. This paper addresses potential consequences of the change to digital information, and suggests that universities can cope by being more proactive in their use of the Web for reward and communication. (Author)
Session #4  Patterns of Usage

Digital Libraries: A Unifying or Distributing Force?

Michael Lesk
Division Manager
Computer Science Research
Bellcore

Digital Libraries: A Unifying or Distributing Force?

Michael Lesk

For presentation at Scholarly Communication and Technology, a conference sponsored by the Andrew W. Mellon Foundation, Atlanta, Georgia (April 24, 1997).
Abstract

What kinds of communities will digital library technology produce? The Web seems much more popular than electronic journals. Does this mean that surfing will replace literature reading, and that nerds building HTML hierarchies will supplant publishers? Will this mean that the universities lose control of the quality of what their students read? Will the ability to do more research in one's dorm room mean that students do not talk to one another at all, that they talk to people somewhere else in the world, or that they talk to their roommates more than ever, perhaps about how to use the computer system?

Digital information threatens our ideas of locality: will the association of students with a particular university, let alone university library, survive the Web? Might we find that online references and online multimedia lectures would produce the 'virtual university of the United States' and if so would we want that? Universities serve a variety of social functions which the Web can augment or diminish, depending on our actions. The Web also may threaten our ideas of quality in scholarship. This paper addresses potential consequences of the change to digital information, and suggests that universities can cope by being more proactive in their use of the Web for reward and communication.

Introduction

There are several future trends that everyone seems to agree upon. They include

- widespread availability of computers for all college and university students and faculty;
- general substitution of electronic for paper information;
- library purchase of access to scholarly publications, rather than physical copies of them.

Early steps in these directions have been followed by many libraries. Much of this has taken the form of digitization. Unfortunately some of the digitized material is not used as much as we would like. This may reflect the choice of the material to convert; realistically 19th century books which have never been reprinted or microfilmed may have been obscure for good reasons and will not be used much in the future. But some more general problems with the style of much electronic library material suggest that the difficulties may be more pervasive.

The Web

The primary means today whereby people gain access to electronic material is over the World Wide Web. The growth of the Web is amply documented at http://www.cyberatlas.com and similar sites. Predictions for the number of Web users worldwide in the year 2000 run up to 1 billion [Negroponte 1995]; students have the highest Web usage of any demographic group, with about 40% of them in 1996 showing medium or high Web usage; and people have been predicting the end of paper libraries since at least 1964 [Samuel 1964]. Web surfing appears to be substituting for TV viewing and CD-ROM purchasing, taking its share of approximately 7 hours per day that the average American spends dealing with media of all forms. Advertisers are lining up to investigate Web users and find the best way to send product messages to them [Hoffman 1996].
The table below shows the growth of Web hosts just in the last three years (from Cyberatlas and Network Wizards):

<table>
<thead>
<tr>
<th>Date</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-96</td>
<td>9,472,224</td>
</tr>
<tr>
<td>Jul-95</td>
<td>6,642,000</td>
</tr>
<tr>
<td>Jan-95</td>
<td>4,852,000</td>
</tr>
<tr>
<td>Jul-94</td>
<td>3,212,000</td>
</tr>
<tr>
<td>Jan-94</td>
<td>2,217,000</td>
</tr>
<tr>
<td>Jul-93</td>
<td>1,776,000</td>
</tr>
<tr>
<td>Jan-93</td>
<td>1,313,000</td>
</tr>
<tr>
<td>Jul-92</td>
<td>992,000</td>
</tr>
</tbody>
</table>

Online Journals and the Web

Following the move of information to digital form, there are many experiments with online journals. Among the best known projects of this sort are the TULIP project of Elsevier [Hunter 1996] and the CORE project of Cornell, the American Chemical Society, Bellcore, Chemical Abstracts, and OCLC. These projects achieved more or less usage, but none of them approached the degree of epidemic success shown by the Web. The CORE project, for example, logged 87,000 sessions of 75 users, but when we ended access to primary chemical journals at Cornell, nobody stormed the library demanding the restoration of service. You can imagine what would happen if the Cornell administration were to cut access to the Web.

In the CORE project (see Entlich 1997), the majority of the usage was from the Chemistry and Materials Science departments. They provided 70% of active users and 86% of all sessions with the journals. There are various other departments at Cornell which use chemical information (Food Sciences, Chemical Engineering, etc.) but make less use of the online journals. Apparently the overhead of starting to use the system and learning its use discouraged those for whom it was not their primary interest. Many of the users printed out articles rather than read them online; about one article was printed for every four viewed, and people tended to print an article rather than flip through the bitmap images. People accessed articles through both browsing and searching, but they read the same kinds of articles they would have read otherwise, rather than changing their reading habits.

Some years ago the CORE project had compared the ability of people to read bitmaps compared with reformatted text, and found that people could read screen bitmaps just as fast as new text [Egan 1991]. Yet, in the actual use of the journals, the readers did not seem to like the page images. The Scepter interface provided a choice of page image or text format, and readers only looked at about one image page in every four articles. This suggests that despite assertions by some chemists in early interviews that they particularly liked the layout of ACS journal pages, for viewing online they prefer reformatted text to images of those pages, even though they can read either at the same speed. The Web-like style is preferred for online viewing.
Perhaps it is not surprising that the Web is more popular than scientific journals. After all, *Analytical Chemistry* has never had the circulation among undergraduates of *Time* or *Playboy*. But the Web is not being used only to find out sports scores or other non-scholarly activities (30% of all *Alta Vista* queries are about sex) [Weiderhold 1997]. The Web is routinely used by students to access all kinds of information needed in classroom work or for research. When I taught a course at Columbia, the students complained about reading assigned on paper, much preferring the reading which was available on the Web. The Web is preferred not just because it has recreational content but also as a way of getting scholarly material.

The convenience of the Web is obvious. If I need a chart or quote from a Mellon Foundation report, I can bring it up in a few tens of seconds at most on my workstation. If I need to find it on paper, and it isn't in my office, I'm faced with a few minutes to visit the Bellcore library, and probably a few weeks since like most libraries they are cutting back on acquisitions and will have to borrow it from somewhere else. The Web is so convenient that I frequently use it even to read publications that I do have in my office.

Web use is greeted so enthusiastically that volunteers have been typing in (or scanning) out-of-copyright literature on a large scale, as for example for Project Gutenberg. The figure below shows the number of books *added* to the Project Gutenberg archive each year in the 1990s; by comparison in the entire 1980s only two books were entered.

![Project Gutenberg texts](image)

By comparison, some of the electronic journal trials seem disappointing. Some of the reasons that digital library experiments have been less successful than they might have been involve the details of access. Whereas Web browsers are by now effectively universal on campuses, the specific software needed for the CORE project, as an example, was somewhat of a pain for users to install and use. Many of the electronic library projects involve scanned images which are difficult to manipulate on small screens, and they have rarely involved material which was designed for the kind of use that is common on computer systems. By contrast, most HTML material is written with the knowledge of the format in which it will be read and is adapted to that style. I note anecdotal complaints even that Acrobat documents are not as easy to read as normal Web pages.
Web pages, in particular, may have illustrations in color, and even animations, beyond the practical ability of any conventional publisher. Only one in a thousand pages of a chemical journal, for example, is likely to have a color illustration. Yet most popular web pages have color (although the blinking colored ad banners might be thought to detract rather than help Web users). Also, Web pages need not be written to the traditional standards of publishing—the viewgraphs that represent the talk associated with a scholarly paper may be easier to read than the paper itself.

This suggests that the issue with the popularity of the Web compared with digital library experiments is not just content or convenience but also style. In the same way that *Scientific American* is easier to read than traditional professional journals, Web pages can be designed to be easier for students to read than the textbooks they buy now. Reasons might include the way material is broken into fairly short units, each of which is easy to grasp; the informal style; the power of easy cross-referencing, so that details need not be repeated, the extreme personality shown by some Web pages, and the use of illustrations as mentioned before. Perhaps some of these techniques, well known to professional writers, could be encouraged by universities for research writing.

The attractiveness of the newer Web material also suggests that older material will become less and less read. In the same way that vinyl records have suddenly become very old, or that TV stations refuse to show black-and-white movies, libraries may find that the 19th century material in many libraries disappears from the view of the students. Mere scanning to produce bitmaps, resulting in material which can not be searched and which does not look like newly written text, may produce material that although more accessible than the old volumes, is still not as welcome to students as new material. How much conversion of the older bitmaps can be justified? Of course many vinyl recordings are reissued on CD, and movies are colorized, but libraries are unlikely to have resources to do much updating. How will we be able to present the past in a way that students will be willing to use? Perhaps this will be a golden age for scholars as nearly the entire world supply of reference books will have to be rewritten for HTML.

**Risks of the Web**

Of course, access to Web pages typically does not involve the academic library or bookstore at all. What does this mean for the future of access to information at a university? There are threats to various traditional values of the academic system.

**Quality.** Much of the material on the Web is junk; Gene Spafford refers to Usenet as a herd of elephants with diarrhea. Are students going to come to rely on this junk as real? Would we stop believing that slavery or the Holocaust really happened if enough followers of revisionist history put up a predominance of web pages claiming the reverse?

**Loyalty.** It has already been a problem for universities that the typical faculty member in surface effect physics, for example, views his or her colleagues as the other experts in surface effect physics around the world, rather than the other members of the same physics department. Will the Web now mean that this is true of undergraduates as well? Will University of Michigan undergraduates read web pages from Ohio State? Can the Midwest survive that?

**Shared experience.** Santayana wrote that it didn't matter what books students read as long as they all read the same thing. Will the great scattering of material on the Web mean that few undergraduate will be able to find somebody else who has been through the
same courses reading the same books? When I was an undergraduate I once had a friend who would look at people's bookshelves and recite the courses they had taken. This will become impossible.

**Diversity.** Since we can always fear two contradictory dangers, perhaps the ease of getting a few well-promoted Web site will mean that fewer sources are read. If nobody wants to waste time on a Web site that does not have cartoons, fancy color pictures and animation, then only a few well-funded organizations will be able to put up web sites that get an audience. Again, the United States publishes about 50,000 books each year, but produces less than 500 movies. Will the switch to the Web increase or decrease the variety of materials read at a campus?

**Equality of access** If computers are needed to find information, will this produce barriers for people who lack money, good eyesight, or some kinds of interface-using skills? Universities want to be sure that all students can use whatever information delivery techniques are used; is the Web acceptable to at least as wide a span of students as the traditional library?

**Recognition.** Traditionally faculty obtain recognition and status from publishing in prestigious journals. High-energy physicists used to get their latest information from *Physical Review Letters*; today they rely on Ginsparg's preprint bulletin board at Los Alamos National Laboratory. Since this is not referred, how do people select what to read? Typically, they choose papers by authors they have heard of. So the effect of the switch to electronic publishing is that it is now harder for a new physicist to attract attention.

A broader view of threats posed by electronics to the university, not just those arising from digital library technology, has been presented by Eli Noam [Noam 1995]. Noam worries more about video tapes and remote teaching via television, and the possibility that commercial institutions might attempt to supplant universities, offering cheap education based entirely on electronic technologies. Should they succeed in attracting enough customers to force traditional universities to lower tuition costs, the financial structure of present-day higher education would be destroyed. Noam recommended that universities emphasize personal mentoring and one-to-one instruction to take the greatest advantage of physical presence.

Similarly, Van Alstyne and Brynjolfsson [Van Alstyne 1996] have warned of 'balkanization' caused by the preference of individuals to select specialized contacts. They point to past triumphs involving cross-field work, such as the history of Watson and Crick, trained in physics and zoology respectively. In their view, search engines can be too effective, since letting people read only exactly what they were looking for may encourage overspecialization.

As an example of the tendency towards seeking collaborators away from one's base institution, the figure below shows the tendency of multi-authored papers to come from more than one institution. It was made by taking the first issue each year from the *SIAM Journal of Control and Optimization* (originally named *SIAM Journal of Control*) and counting the fraction of multi-authored papers in which all the authors came from one institution. The results was averaged over each decade. Note the drop in the 1990s. There has also, of course, been an increase in the total number of multiauthored papers (in 1965 the first issue had 14 papers and every paper had only one author; in 1996 there were 17 papers and only two were single-authored). But few of the multiple-authored papers today came from only one research institution.
Of course, there are advantages to the new technology as well, not just threats. And it is clear that the presence of the Web is coming, whatever universities do -- this is the first full paper I have written directly in HTML, rather than prepared for a typesetting language. Much of the expansiveness of the Web is all to the good; for many purposes access to random undergraduate opinions, and certainly to their fact-gathering, may well be preferable to ignorance. It is hard to imagine students or faculty giving up the speed with which things can be accessed from their desktops, anymore than we will give up cars because it is healthier to walk or ecologically more desirable to ride trains. How, then, can we ameliorate or prevent the possible dangers elaborated before?

University Publishing

Bellcore, like many corporations, has a formal policy for papers published under its name. These papers must be reviewed by management and others, reducing the chance that something sufficiently erroneous to be embarrassing, or something which poses a legal risk to the corporation, will appear. Many organizations do not yet have any equally organized policy for managing their web pages (Bellcore does have such a policy, dealing with an overlapping set of concerns). Should universities have rules about what can appear on their web pages? Should such rules distinguish between what goes out on 'personal' or 'organizational' pages? Should the presence of a page on a Harvard web page connote any particular sign of quality, similar to the appearance of a book under the Harvard University Press imprint? Perhaps a university should have an approved set of pages, providing some assurance of basic correctness, decency of content, and freedom from viruses; then people wishing to search for serious content might restrict their searches to these areas.

The creation of a university web site as the modern version of a university press or a journal offers a sudden switch back from publishers to the universities as the providers of information. If a university were to provide a refereed, high-prestige section of its web site, could it attract the publication that now goes to journals? The effect of this would be to provide a way for students to find quality material, and to build institutional loyalty and shared activities among the members of the university community. Perhaps the easiest way of doing this would be to make tenure depend on contribution to the university website, instead of contributions to journals.

The community could even be extended beyond the faculty. Undergraduate papers could be placed on a university web site; one can easily imagine different parts of the site for different
genres ranging from the research monograph to the quip of the day. This would let all students participate and get recognition, so long as there is some quality control imposed on this part of the site and that presence on it is recognized as an honor.

In addition to supporting better quality, a university web site devoted to course reading could make sure that a diversity of views is supported. Online reading lists, just like paper reading lists, can be compiled to avoid the problem of everyone relying on the same few sites. This would help, for example, if many of the search engines start making money by charging people to be listed higher in the list of matches (a recurrent rumor, but perhaps an urban legend). It would also push students to look at sites which perhaps lack fancy graphics and animation.

One could even imagine that excessive reliance on a university web site could produce too much inbreeding. If we lost the publications that now provide general prestige in favor of university web sites, will it be possible for a professor at a less prestigious university to put an article on the Harvard or Stanford web site? If not, how will anyone ever move up? I do not perceive this as likely to be a problem anytime soon; the reverse (a total lack of organizational identification) is more likely.

It is likely that web sites of this sort would not include anonymous contributions. The net is somewhat overrun right now with untraceable postings that often contain annoying or inflammatory material, ranging from the merely boring commercial advertising to the deliberately outrageous political posting. Having a place which did not allow this kind of material might help to civilize the Web and make it more productive.

Information Location

Some professors already provide Web reading lists, corresponding to the traditional lists of paper material. The average Columbia course, for example, has 3000 pages of paper reading (with an occasional additional audiotape in language courses). The lack of quality on the Web means that it will become more important for faculty to provide guidance to undergraduates about what to read there.

More important, it will be necessary for faculty to teach the skill of looking purely at the text of a document and making a judgment as to its credibility. Much of our ability to evaluate a paper document is based on the credibility of the publisher. On the Web, students will have to judge by principles like those of paleography. What do we know, if anything, about the source? Is there a motive for deception? How does the wording of the document read — credibly or excessively emotionally? Do facts that we can check elsewhere agree with other sources?

The library will also gain a new role. Universities should provide a training service for how to search the Web, and the library is the logical place to provide that. Partly this is a result of the training of librarians in search systems, which are rarely studied formally by any other groups. In addition, the librarians are the only hope to keep the alternative old information sources in front of students until most of them are converted, which will take a while.

The art of learning to retrieve information may also bring students together. I once asked a Columbia librarian whether the advent of computers and networks in the dormitory rooms was creating a generation of introverted nerds lacking social skills. She replied that it was the reverse. In the days of card catalogs students were rarely seen together; each person searched
the cards alone. Now, she said, she frequently saw groups of two or three students at the OPAC terminals, one explaining to the others how to do something. Oh, I said, so you're improving the students social skills by providing poor human interface software. Not intentionally, she replied. Even with good software, however, there is still a place for students helping each other find things, and universities can try to encourage this.

Much has been written about the 'information rich' vs. the 'information poor' and the fear that once a machine costing several thousand dollars is needed to gain information, poor people will be placed at a still greater disadvantage in society than they are today. In the university context, money may not be the key issue, since many university libraries provide computers for general use. However, some people face non-financial barriers to the use of electronic systems. These may include limited eyesight or hearing (which of course also affect the use of conventional libraries). More important is perhaps the difficulty that some users may have with some kinds of interface design. This ranges from relatively straightforward issues such as color-blindness, to complex perceptual issues involving different kinds of interfaces and their demands on different individuals. So far we do not really know whether some users will have a need for something other than whatever becomes the standard information interface; in fact we do not know whether some university students in the past had particular difficulties learning card catalogues.

Libraries may also be a good place to teach aspects of collaboration and sharing that will grow out of references as hyperlinking replaces traditional citation. Students are going to use the Web to cooperate in writing papers as well as finding information for them. The ease of including (or pointing to) the work of others is likely to greatly expand the extent to which student work becomes collaborative. Learning how to do collaborative work effectively and fairly is an important skill students can acquire. In particular, the desire to make attractive multimedia works, which may need expertise in writing, drawing, and perhaps even composing music, will drive us to encourage cooperative work. Given the start of this effort with quoting references, the library may be a place to teach cooperative software.

Students could also be encouraged to help organize all the information on the local web site. Why should a student's web page prefer local resources? Perhaps because some kind of academic credit is created for doing that. University web sites, to remain useful, will require constant maintenance and updating. Who is going to do that? Realistically, students

New Creativity

There is a wide rush of new presentation modes on the Web. We are going to see applets implementing animation, interactive games, and many other new kinds of presentation modes. The flowering of creativity in this should be encouraged. In the early days of television and of movies, the amount of equipment involved was beyond the resources of amateurs, and universities did not play a major role in their development. By contrast, universities are important in American theatre and classical music. The Web is also an area in which equipment is not really a limitation, and universities have a chance to play a role.

This represents a chance for the university art and music departments to join forces with the library. Just as the traditional tasks of preparing reading lists and scholarly articles can move onto a university web site, so can the new media. The advantage of doing this with the library is that we can actually save the beginnings of a new form of creativity. We lack the first email message; nobody understood that it was worth saving. Much of early film (perhaps half the
movies made before 1950) no longer survives. 1950s television is mostly gone for lack of recording devices. In an earlier age, the Elizabethans did not place a high value on saving their dramatic works; of the plays performed by the Admiral's Men (a competitor to Shakespeare's company) we have only 10% or 15% today. We have a chance not to make the same mistake with innovative Web page designs, providing that such pages are supported in some organized way, rather than on computers in individual student dorm rooms.

Recognizing software as a kind of scholarship is a change for the academic community. The National Science Foundation tends to say "we don't pay for software, we pay for knowledge," drawing a sharp distinction between the two. Even computer science departments have sometimes said that you can't get a PhD for writing a program. The new kinds of creativity will need a new kind of university recognition. Will we have honorary web pages instead of honorary degrees? We need undergraduate course credit and tenure consideration for web pages.

Software and data are new kinds of intellectual output which are not traditionally considered creative. Traditionally, for example, the design of a map was considered copyrightable; the data on the map, although representing more of the work, were not considered design and not protectable. In the new university publishing model, data should be a first-class item, whose accumulation and collection is valuable and leads to reward.

Switching to honoring a web page rather than a paper does have consequences for style, as discussed above. Web pages also have no size constraints; in principle there is no reason why a gigabyte could not be published by an undergraduate. Universities will need to develop both tools and rules for summarizing and accessing very large items, as needed.

Conclusion

To preserve access to quality information while also preserving some sense of community in a university, the academic institutions should take a more active view of their web sites. By using the Web as a reward, and as a way of building links between people, universities could serve a social purpose as well as an information purpose. The ample space and low cost of Web publishing provide a way to extend the intellectual community of a university, and to make it more inclusive. This may encourage students and faculty to work together, maintaining a local bonding of the students. The goal is to use university web publishing, information searching mechanisms, and rewards for new kinds of creativity to build a new kind of university community.

REFERENCES:


For additional information about the conference, or The Andrew W. Mellon Foundation's scholarly communication initiatives, please contact Richard Ekman. For additional information about ARL or this web site contact Patricia Brennan, ARL Program Officer at (202) 296-2296.