ED 454 859 IR 058 150

AUTHOR	Caplan, Priscilla
TITLE	International Metadata Initiatives: Lessons in Bibliographic Control.
PUB DATE	2000-11-00
NOTE	20p.; In: Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web (Washington, DC, November 15-17, 2000); see IR 058 144.
AVAILABLE FROM	For full text: http://lcweb.loc.gov/catdir/bibcontrol/caplan_paper.html.
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Cataloging; *Internet; *Metadata; *Standards; User Needs (Information); Users (Information); Visual Aids
IDENTIFIERS	Dublin Core; Electronic Commerce; *Electronic Resources; MARC

ABSTRACT

This paper looks at a subset of metadata schemes, including the Text Encoding Initiative (TEI) header, the Encoded Archival Description (EAD), the Dublin Core Metadata Element Set (DCMES), and the Visual Resources Association (VRA) Core Categories for visual resources. It examines why they developed as they did, major point of difference from traditional library cataloging, and what advantages they offer to their user communities. It also discusses challenges to implementers of these schemes and possible developments. It goes on to identify some commonalities among these cases and to attempt to generalize from these some lessons for developers of metadata element sets. It concludes by suggesting the importance of looking carefully at emerging schemes being developed by publishers in support of electronic commerce and rights management and thinking seriously about the implications of commodity metadata upon tradition bibliographic apparatus. (Contains 18 notes.) (MES)

International Metadata Initiatives: Lessons in Bibliographic Control

Priscilla Caplan
Assistant Director for Digital Library Services
Florida Center for Library Automation
2002 NW 13 St., Suite 320
Gainesville, FL 32609

A blooming garden, traversed by crosswalks, atop a steep and rocky road

Final version

Library historians are likely to see the1990s as a decade of particular excitement, creativity and change. It will certainly be known for the rise of the World Wide Web, and as the decade that the Digital Library was invented. It may also be known for an almost explosive proliferation of metadata schemes. The first draft version of the Text Encoding Initiative (TEI) Guidelines, including the definition of the TEI header, was distributed in 1990. The first version of the FGDC Content Standard for Digital Geospatial Metadata was released in 1994. The workshop that drafted the original Dublin Core Metadata Element set was held in 1995. The alpha version of the Encoded Archival Description (EAD) was released in 1996. The Core Categories for Visual Resources version 2.0 was released by the Visual Resources Association in 1997. The Data Documentation Initiative was established in 1995 and released an XML version of the DDI metadata standard for social science data resources in 1997. The learning resources community produced both the Dublin Core-based Gateway to Educational Materials (GEM) element set in 1998 and the IMS Meta-data Specification in 1999. And so on; this list is only a sampling. In the metadata garden, truly a thousand flowers are blooming.

This has been a mixed blessing for libraries, presenting (as most innovations do) both opportunities and challenges. On the positive side, it has given us new options for describing materials that are poorly served by the AACR2/MARC suite of standards, and it has created a renewed sense of intellectual excitement in resource description. At the same time, these new formats have placed new burdens on the library profession. There are mature, well-developed tools for creating and managing traditional cataloging [1]. There is in fact an entire industry dedicated to its support -- the integrated library system is after all integrated around the multipurpose bibliographic database. Now suddenly we are confronted by content standards with no syntax and with data structures that we have no systems to support. Suddenly we are charged with supporting any number of schemes, not to mention maintaining registries of them and crosswalks between them. Suddenly there is an expectation we can control and give access to metadata created by organizations outside of our own library community.

This paper looks at a subset of these metadata schemes in more detail: the TEI header, EAD, DCMES, and VRA Core. It looks at why they developed as they did, major points of difference from traditional library cataloging, and what advantages they offer to their user communities. It also discusses challenges to implementers of these schemes and possible future developments. It goes on to identify some commonalties among these cases, and to attempt to generalize from these some lessons for developers of metadata element sets. It concludes by suggesting we also look carefully at emerging schemes being developed by publishers in support of electronic commerce and rights management, and think seriously about the implications of commodity metadata upon our traditional bibliographic apparatus.

The Text Encoding Initiative (TEI) Header

The TEI Header is a good place to begin because it is basically bibliographic, in a narrow sense of the word. Encoded texts are fundamentally like books in a way that art slides, museum objects and satellite data are not. Many texts marked up according to the TEI guidelines are based on printed books for which AACR2/MARC catalog records exist. The developers of the TEI specification were well aware of libraries and the principles of bibliographic description. Under these circumstances it is not unreasonable to ask why the TEI header was developed at all. Why didn't the Text Encoding Initiative rely on library catalog records, and put their energies towards modifying traditional cataloging to better accommodate TEI-encoded texts?

The answer, to a large extent, was a matter of workflow. The TEI drafters envisioned that the same individuals who marked up electronic texts would be creating metadata for them, and that these individuals would not be librarians but rather humanities scholars. These scholar/encoders might be experts in their own areas but they could not be expected to learn cataloging rules, so the TEI guidelines quite deliberately do not require any cataloging knowledge. On the other hand, the drafters equally deliberately designed the header to provide a trained cataloger the information he would need to create a good cataloging record. [2] The header areas are based on ISBD, but rules for obtaining and representing the content are not prescribed.

Similarly, scholar/encoders could be expected to know SGML markup, so it was natural to represent the metadata content in SGML rather than MARC. Using SGML in turn allowed the metadata to be embedded in the TEI document itself, processed by the same software, and searched within the same retrieval system. In theory, if a standalone record was required, the header could be used to algorithmically create a MARC record for importing into the library's catalog system.

In fact, events did not turn out exactly as envisioned. Most TEI texts are created not by scholar/encoders, but by the staff of projects or electronic text centers closely associated with libraries. In many cases header data is created or reviewed and revised by librarians. This has led to a perceived need on both sides to bring the header more in line with traditional cataloging. Librarians have found the "leg-up" provided by the header to be of limited usefulness. A study by the CC:DA Task Force on Metadata and the Cataloging Rules analyzed the TEI header as a source of cataloging data and concluded, not surprisingly, that the data is directly usable only to the extent that the encoder followed cataloging rules. [3] Automatically derived MARC records are problematic for this reason, and cannot be integrated into library catalogs without review.

At the same time, a desire to support searching across multiple collections, or even to share TEI data between institutions, has provided an impetus for more consistency in both content guidelines and encoding practices. The Oxford Text Archive sponsored a meeting in the fall of 1997 which acknowledged both the need for greater compatibility with traditional cataloging and greater consistency in practice between electronic text centers, resulting in the draft of a guide to good practice. The following year a workshop on TEI and XML in Digital Libraries held at the Library of Congress charged a working group to "recommend some best practices for TEI header content and review the relationship between the Text Encoding Initiative header and MARC", resulting in a draft TEI/MARC Best Practices document. [4] (Interestingly, most studies of the TEI header have focused on its compatibility with traditional cataloging and its usefulness in relation to a library catalog system. Whether the TEI conventions, designed to be useable by scholar/encoders, are more or less useful than traditional cataloging for scholar/users, has not to my knowledge been studied.)

The TEI Header is not, of course, directly analogous to a catalog record, and supports a number of data categories which cannot be mapped to MARC or can only be loosely mapped to note fields. The change history section (<revisionDesc>) provides a structured way to log changes made to an electronic text, including date, responsible party, and nature of change. The elements for describing the source on which a TEI text is based (the <sourceDesc> within the <fileDesc>) allows a detailed and richly content-designated description which goes far beyond the MARC 534, particularly for non-print sources such as the spoken word, audio or video recordings. The encoding description (<encodingDesc>) section provides a place for lengthy and detailed description of the encoding of the electronic file, including data about the project which created it, the purpose for which it was created, transcription practices followed, editorial decisions made, and SGML tagging applied.

The encoding description area of the header is notable because it supports a function not addressed by IFLA's Functional Requirements for Bibliographic Records (FRBR) [5]: the ability to make use of the resource. FRBR describes four generic user tasks that catalog records must support: to find materials that correspond to the user's stated search criteria, to identify an entity, to select an entity appropriate to user needs, and to acquire or obtain access to the entity. This "bibliographic" approach to metadata has been contrasted to the approach taken by computer scientists, which puts more emphasis on the management of data, including support of data use, data sharing, data security, and data integrity functions. [6] The TEI header contains elements of both traditions, treating the electronic text as both an object to be discovered and a data file to be used and managed over time.

In sum, the TEI header contains bibliographic information supporting resource discovery, and data management portions supporting use of the resource. Historically, the progression of the bibliographic portion of the TEI header has been toward greater consistency in encoding, greater compatibility with traditional library cataloging, and greater syntactical congruence with MARC. This makes sense in the context of an integrated information system, serving the user who may be interested in any and all versions of a work, including printed texts, electronic reproductions, and TEI encoded representations. At the University of Michigan, TEI headers are actually generated from MARC records, and in a Library of Congress implementation, bibliographic fields are left out of the header altogether. It may be that the TEI header will evolve to carry only minimal bibliographic description, with the bulk of the section being replaced by an external MARC record. The MARC record could then point to a TEI header containing detailed encoding, profile and revision information, much as collection-level AMC records are used to point to more detailed EAD finding aids today.

The Encoded Archival Description (EAD)

The developers of the EAD had both MARC and the TEI header available to them as models. Unlike the TEI header, however, the EAD was designed as an electronic finding aid to resources that would not necessarily be available in electronic form. While the EAD can be used to describe web-accessible collections, its primary purpose is to improve awareness of archival holdings in all formats.

The archival community had been using the MARC AMC format for some time to give high level access to archives and manuscript collections. Archivists found, however, that AACR2 was inadequate for archival description, and adopted instead Steven Hensen's *Archives, Personal Papers and Manuscripts* (APPM) for content rules. The principles of bibliographic description apply even less to finding aids, as Daniel Pitti has pointed out. [7] Bibliographic description represents a published item; archival description represents a fonds, or organically generated collection. Bibliographic description emphasizes physical characteristics; archival description emphasizes intellectual structure and content. Bibliographic description supports finding, identification, selection and access; archival description is evidentiary and must document provenance and original order. A distinction of practical importance is that bibliographic description is typically brief, stylized and flat. Archival description is typically lengthy, narrative, and deeply hierarchical, making SGML, and later XML, a more suitable transport syntax than MARC.

Although archivists generally follow principles for archival description in their local finding aids, and although the General International Standard Archival Description (ISAD(G)), a set of general rules for archival description, was adopted by the International Council of Archives in 1994, there is no ruleset equivalent to APPM specifically for finding aids. In the absence of an existing content standard, the developers of FindAid, the predecessor from which the EAD evolved, solicited examples of paper finding aids from the community. As repositories tended to contribute only the samples they considered their best, a de facto corpus of best practices was acquired and used as the basis for developing the FindAid DTD. While the EAD was designed to accommodate the range of practice that was found, it was also developed in the hope of encouraging common practices regarding data content.

Paralleling the TEI standard which has a header preceding the encoded text, the EAD is divided into two parts, a bibliographic header (<eadHeader>) and the marked up finding aid itself (<findAid>). [8] The finding aid describes the collection and the header describes the finding aid, reminding us that one man's metadata is another man's data. The header in turn has sections for describing the original finding aid (for example, its author and title), describing the encoded version (for example, whether it was created by OCR, retyping, etc.), and recording a revision history.

The EAD has been rapidly, widely and internationally embraced, particularly by university archives and special collections departments within academic libraries. Certainly one source of this success is the ability of the EAD to accommodate existing archival practice, rather than forcing practice to conform to the constraints of a data format or syntax. [9] Because of this congruence, it has been possible to convert paper finding aids with some success, and something of a cottage industry has arisen in providing vended conversion services. The EAD also appears to be filling a void in tools for detailed collection description, as institutions are applying it to collections of all sorts, not only those controlled archivally.

Adoption of the EAD has been notably slower outside of academic institutions. State archives, for example, still rely almost exclusively on collection-level MARC records. A meeting of the Southeastern Archives and Records Conference in 1999 concluded that the "EAD is not useful unless there is substantive information at lower levels, and most state records series have only box inventories, with the frequent exception of governor's records." [10] The SGML-based structure also requires specialized editing tools and software for search and display that can present a barrier to implementation at smaller institutions. While the scholar/encoders of the TEI might be expected to have a research interest in SGML, most archivists would have no reason to be familiar with this encoding apart from the EAD. RLG and the Society of American Archivists (SAA) have been proactive in sponsoring intensive training, which is almost a prerequisite to implementation.

A major strength of the EAD -- its ability to represent complex finding aids with a high degree of content designation, while accommodating a wide range of local practices -- can also be a drawback. Although a relatively small number of tags are required, the tagset itself is extensive, and every repository must arrive at its own set of guidelines for which tags to use, how they may be used, and how data may be represented within them. Because the EAD does not include and is not directly correlated with established content rules, this allows for some creativity, and there is wide variation in practice. Widespread implementation of the EAD has been followed almost immediately by the desire for union access. In 1998 RLG launched its Archival Resources service, a union catalog of distributed collection guides. Using a registry of contributors and a customized harvester, the service collects and indexes EAD and non-EAD finding aids. In 1998 and 1999, the Digital Library Federation undertook a project to develop a Distributed Finding Aid Search System (DFAS). DFAS implemented Z39.50 search and retrieval across distributed EAD repositories as an alternative to the union catalog approach. Both Archival Resources and DFAS found the diversity in encoding practice to be a major problem. A report of the DFAS project concluded, "Our research has highlighted the problems caused by the lack of standardization in the application of EAD to finding aids, yet that lack of standardization is not easily overcome given the diversity of the underlying documents." [11]

The EAD has very clearly encouraged archivists to conceptually reexamine the logic, structure and content of their finding aids. In some cases has inspired repositories to reengineer their finding aids for more effective web-based use. It appears that the next phase in EAD development will be the establishment of common guidelines, including Z39.50 profiles and Best Practices for encoding particular types of finding aids. Changes in the EAD DTD itself may be necessary as use expands beyond the academic community that invented it, and as more experience is gained in representing collections of both digital and non-digital content.

Dublin Core Metadata Element Set (DCMES)

The DCMES is unusual among metadata element sets in the generality of its application and use. In contrast to other schemes which target particular types of materials and particular user communities, DCMES can be, and probably has been, used to describe nearly any type of information resource.

Like the TEI header and the EAD, the DCMES has evolved in unexpected ways. Though originally envisioned as a mechanism for encouraging authors to supply metadata for their own publications, the vast majority of use is from projects associated with libraries, cultural heritage institutions and government agencies. Originally intended to support description and discovery of what Clifford Lynch has called the "dark matter", or largely invisible content, of the Web, DCMES has found a multiplicity of other applications. It has been particularly useful in support of interoperability -- retrieval across multiple existing metadata stores. In this capacity DCMES has been used as a minimal set of commonly understood access points for cross-domain searching, as a common extract format for creating union catalogs, and as a searchable entry point to local files of more complex metadata. An emerging use in Open Archives and related initiatives is as the basis of an extract format for harvesting metadata from dissimilar repositories.

The most astonishing thing about DCMES is the pervasiveness of its adoption. Although the Dublin Core website maintains a list of DCMES-based projects, this barely hints at the number of implementations worldwide using or somehow based on the Core. One reason this is possible is because DCMES allows, even encourages, the use of local extensions. The basic model is to use DCMES elements where they apply, and supplement them with domain- or application-specific elements where needed. XML namespaces, which are supported in RDF and in the emerging specification for XML Schema, provide a practical means for implementing this type of combination.

Ironically, this same use highlights a weakness of the DCMES as a building block for other metadata schemes. Much has been made of the "Lego"(tm) model, in which Dublin Core elements can be snapped into other schema as appropriate. However, Legos (tm) require an extreme degree of precision, and for this approach to work, DCMES elements should be related to a data model that can be precisely described. The DCMES, however, developed organically, and attempts to apply a more rigorous data model after the fact have had to contend with inconsistencies already present in the element set. This has led to some tension between the need to maintain stability for existing implementers on the one hand, and the desire to move the element set towards greater logical consistency on the other.

Practically, most projects using DCMES have found its lack of specificity to be a problem. DCMES 1.1 gives only the broadest description of semantic categories; there are no rules for how to determine or represent content, and only the most general guidelines are available in draft status from the website. As a consequence, projects using DCMES for resource description find it necessary to develop their own conventions, a difficult and time-consuming endeavor. A working group charged with developing a user guide found that although librarians tended to be frustrated by the lack of a ruleset, other sectors were not, and there was no general consensus that common content rules were either necessary or desirable. In any case, the huge diversity of applications argues against canonical guidelines. The expectation is that communities sharing particular resource needs will get together to develop domain-specific rules. However, this is itself an arduous process. The CIMI Guide to Best Practice, for example, now available in version 1.1, took three years, a testbed implementation and extensive community review to accomplish. [12]

Most projects have also found a need for some refinement of the very general DCMES semantics, ordinarily referred to as "qualification". An initial set of qualifiers formally approved by the Dublin Core Metadata Initiative (DCMI) is in final draft status at the time of this writing. (Exactly what this means in terms of compliance for applications is still somewhat unclear.) Applications and communities are encouraged to develop their own qualifiers, as with extensions, and to submit these to the DCMI for review and approval. However, much of the apparatus required to support consistent and confident use of qualifiers is still outstanding, including clear guidelines for constructing valid qualifiers; a registry identifying approved, not-yet-approved-but-valid, and invalid-but-needed-by-some-community qualifiers; and a mechanism for approving new qualifiers.

The DCMES began as a grass-roots movement, independent of existing organizations and without a formal structure to manage it. Over time the DCMI has evolved alongside the DCMES to be "responsible for the development, standardization and promotion of the Dublin Core metadata element set." [13] However, the very diversity of Dublin Core implementers makes it extremely difficult to achieve consensus on any but the most basic issues. Also, apart from a skeletal directorate, participation is largely a volunteer effort. Unlike other standards discussed here, the DCMES is not a program of any larger organization that sees maintenance of DCMES as part of its core mission. And, although DCMI procedures are modeled after the W3C, the DCMI, unlike W3C, has no formal membership with the corporate commitment and financial support that entails. The most critical factor in the future of DCMES is whether a working organization can be achieved to manage the change process and to produce the documentation, support structures, and policies required by an international community of implementers holding very little in common.

Visual Resources Association (VRA) Core Categories for Visual Resources

Like the DCMES, the VRA Core was conceived as a core set of elements that particular applications could enhance with additional elements as needed. In contrast to the very comprehensive Categories for the Description of Works of Art (CDWA), the VRA Core was designed as a moderate set of elements which, if commonly supplied, would support the sharing of data for visual materials. Historically, catalogs or databases of visual materials tended to be institution-specific, using locally-defined data elements, formats and authorities. There was a great redundancy of effort, as every institution cataloged their own collections of slides based upon the same works of art. Widespread Internet use brought increasing pressure to share data, not only to help users find materials but also to create an environment in which works could be cataloged only once. As one of the drafters of the VRA Core described it, "The point of the exercise ... was to develop a set of elements that all visual resources curators could use to share information about the works of art so that they would not have to repeat the research process for each work represented in his/her collection." [14]

The VRA Core is still evolving fairly rapidly, moving towards a more generalized and more flexible model of the visual materials universe with each version. Version 1.1, which was never widely implemented, was at heart based on the collection of art slides, the basic model being an art object that was not held by the cataloging collection and a slide of the art object that was. The original Core consisted of descriptive elements, or "categories", to describe the object, the creator of the object, and the surrogate. Version 2.0 was a deliberate attempt to generalize the element set to accommodate non-art objects and to give greater weight to the surrogate, the term itself generalized to "visual document". VRA Core 2.0 defined 19 "Work description categories" and another nine "Visual Document description categories". Both version 1.1 and 2.0 attempted to accommodate the practical experience of catalogers of visual materials, that in describing a slide or photograph in their collections, they were simultaneously describing the original work in some other medium. However, this pairing breaks down very quickly into far more complex relationships: not only can there be multiple representations of the same work (a slide, a photo, a digital image), but some of these may be surrogates of others (the digital image is made from the photo), while some may be works in their own right (the photo was taken by a well-known artist). Works may exist as parts of wholes (a stained glass window in a building) or as parts of collections; visual documents may exist in collections, may encompass multiple works (a photo of two buildings), and so on.

Version 3.0 acknowledges this complexity by abandoning the attempt to separately describe works and various representations of them; it simply defines 17 categories that can be used as appropriate. Although it retains the conceptual distinction between a work and representations of the work (now called "images"), it embraces the "1:1 principle" popularized by Dublin Core, that a single set of metadata elements should describe a single entity, and it assumes that records describing images will be linked to the related works. It also incorporates the Dublin Core concepts of elements and element refinements, or "qualifiers"; for example, what in version 2.0 were independent categories for Creator and Role are in version 3.0 one category Creator with the qualifier Role.

Version 3.0 is too recent for widespread implementation, although many of the concepts it represents were previewed in Harvard's Visual Image Access (VIA) application. However, in general the VRA Core has been gratefully, even hungrily, received by visual resources curators. One attraction has certainly been its latitude in addressing both an original work and a derivative surrogate, something very difficult to accomplish in traditional library cataloging. Other attractions have included its focus on visual materials with distinct categories for concepts such as measurements, material and technique, and its flexibility in accommodating local cataloging practices. It has been found to be applicable to works of architecture, non-art images, and other domains beyond the art history slide collection. A paper by Marcia Lei Zeng describes how the VRA Core 2.0 was chosen over MARC and Dublin Core for cataloging a museum collection of historical costumes. [15]

On the other hand, while the VRA Core has been used for describing materials in institutional collections with some success, the visual resources community has some way to go toward achieving the goal of sharing information about works. The lack of standard cataloging rules and common authority schemes for content presents a major barrier to interoperability. Not only is the historical insularity of visual resources collections reflected in any number of purely local vocabularies and classification schemes, but, as visual materials cover a wide range of territory from pottery shards to buildings, a large number of specialty thesauri are in use. Inconsistency in the use of authorities was noted as the major problem in VISION, a testbed for the VRA Core version 2 developed by the VRA, the Getty Information Institute, and the Research Libraries Group. The testbed also revealed a main concern of participants was mapping to and from local databases, a sign that, for early implementers at least, local structures were still their foremost concern. The rapid evolution in the structure of the VRA Core indicates the community has yet to develop an underlying data model to support the complex relationships these materials exhibit. The future of the VRA Core probably depends less upon further improvements in the Categories themselves than on whether their existence serves as a catalyst for the development of a shared data model and content rules.

Functional Requirements for Metadata Records

It could be argued that beyond being intended for electronic description of information resources, the metadata schemes discussed above have little in common. They have different intended users and different intended uses. They are to varying degrees "bibliographic", in terms of being designed to support the Functional Requirements for Bibliographic Records. Some are defined in terms of DTDs, while others are semantic categories independent of syntax. In fact, none of these schemes are exclusively (and some not even primarily) intended for controlling electronic resources: to varying extents they describe paper and artifactual resources as well as digital. One should therefore be cautious about making inferences regarding lessons for bibliographic control of the web. Nonetheless a few generalizations are cautiously offered.

For starters, in no case did the actual creators, users and uses of these schemes turn out to be just what their developers anticipated. Metadata takes on a life of its own. Metadata schemes need to be seen as organic creations evolving in response to a changing environment, with the implication that a mechanism for effecting and controlling this evolution needs to exist. Ideally, such a mechanism is perceived as legitimate and authoritative, has a well-defined structure and process, gathers broad input from affected communities, and controls the rate of incremental change to ensure it is neither too fast for implementers to accommodate nor so slow as to present a barrier to effective use. It is arguable whether any of these emerging metadata schemes have managed to put such a mechanism into place, and it will be interesting to see whether and how these develop. In fact, it will be interesting to see whether the mechanisms governing change to traditional cataloging that have proved sufficient for a universe of print and other fixed, physically distributed media are in fact sufficient to accommodate the control of electronic resources in the rapidly changing network environment.

Another feature shared by all of these schemes is that none of them include or are based on rules for determining and representing content. What we have learned from this is that metadata schemes without content rules are not very useable. Implementers are forced to expend significant time and effort developing their own local guidelines to ensure some consistency in content and encoding within their own resource description projects. As usage becomes more widespread the desire arises to share metadata or to implement union search over a number of repositories, at which point the plethora of local guidelines immediately becomes a hurdle to overcome. In the next phase of maturity, implementers struggle communally to work out common use profiles to guarantee some minimal level of interoperability. Implementers' agreements in turn raise problems of their own: how are they publicized, who officially "owns" them, how are they maintained over time, how to accommodate (or prevent) the development of multiple competing profiles?

In the case of TEI, we see a movement towards greater conformance with traditional library cataloging, while the EAD is serving as an impetus for the development of content rules for finding aids, and there is some hope that the VRA Core will do the same for visual resources. The approach of the DCMES, with its many diverse user groups, has been to encourage the development of community-specific (as opposed to implementation-specific) guidelines and to encourage the use of existing authorities designated with a "scheme" qualifier. In all of these cases, but perhaps most intriguingly with DCMES, what we have been seeing, if we've been paying attention, is the re-invention of cataloging. For example, an extensive exchange concerning the nature of the distinction between Creator and Contributor took place on the main Dublin Core discussion list in the spring of 1999; it explored with some nuance the need to capture primary intellectual responsibility. On the negative side, we can see these communities slowly, painfully and with many false starts rediscover principles that librarians have understood all along. On the positive side, it will be constructive to learn from what they retain and what they throw away, because they are directly confronting what is necessary and feasible to meet the needs of users in the Internet environment.

One of the areas where guidelines are most needed is in how to handle works that are known to be available in different file formats (e.g. LaTeX and PDF) or different manifestations (e.g. a photograph and a digital image made from it). The IFLA model of work, expression, manifestation, and item is useful in untangling multiple versions conceptually, as is the principle of "1:1" espoused by DCMES and the VRA Core. However, it is by no means clear how to apply these practically, as discussions in both communities makes evident, or what mechanisms in supporting systems or in metadata schemes themselves might be required to effect reasonable retrieval and display in this context. [16]

Functions of metadata beyond resource discovery and identification appear to be especially important for electronic resources. The question here is which of these functions are best supported in descriptive schema and which in separate, complementary schemes. Restrictions on access and use, for example, can be seen not so much as a property of a resource but of the intersection of resource, user and use; most emergent descriptive schemes have shied away from extensive recording of rights and permissions, leaving these to other systems. Similarly, there has been quite a bit of activity in defining element sets for administrative and technical metadata useful in managing and preserving digital data over time, including sets defined by the RLG PRESERV, CEDARS, NEDLIB and CURL initiatives.

On the other hand, metadata schemes focused on complex electronic resources tend to include information needed to actually use the resource. The TEI header, for example, allows for lengthy description of encoding practices, and the Data Documentation Initiative (DDI) DTD for describing social science datasets, contains a "data files description" section for a detailed description of the format, size and structure of the datafiles. Metadata documenting the creation and maintenance of the metadata itself also appears to be an important and legitimate need, especially as SGML/XML-based formats encourage lengthier descriptions, maintained over time. Mechanisms for ensuring the authenticity of metadata will almost certainly be required.

Another apparent point of commonality seems to be an inclination to move information about agents (human and legal) into separate files, defined by separate metadata element sets. In the archival sphere, work is proceeding on an SGML encoding of archival authority records based on the International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAR(CPF)). In contrast to LC name authority records, which primarily identify the authorized form of name and contain little other information, these records "describe fully the attributes of the creator needed to appreciate the context of creation of a body of archival documents." [17] By developing an SGML encoding for ISAAR(CPF), archivists will be able to divorce the capture and maintenance of contextual information from the description of the archival entity in the EAD itself. The VRA Core has also evolved away from carrying detailed information about the creator. Version 1.1. included a set of categories pertaining to the creator, including Creator, Nationality and Culture. In version 2.0, Nationality and Culture were redefined to apply to the work, with the recommendation that these data as applied to creators should be recorded in an auxiliary authority file. The DCMI is considering development of an "Agent Core", a structured set of metadata elements such as affiliation and address which properly pertain to the agent (Creator, Contributor, Publisher) as opposed to the resource. This proposal is congruent with an RDF data model where the value of the property "Creator", for example, is itself a resource with properties of its own.

If anything is clear from this it is that the metadata environment is becoming increasingly complicated for both the information provider and the information seeker. Not only are there more metadata schemes for different types of resources, but these schemes rely upon both implementers' agreements to restrict practice and upon local extensions to broaden it. In addition, metadata records created at different times by different agencies and located in different places may have to be integrated at various points of use. In traditional bibliographic environments, the primary form of coordination required has been between descriptive bibliographic records and authority files for names and subjects. Both bibliographic and authority files tend to be under a library's direct control, and headings are either stored redundantly or there are direct links between the two types of records. Despite the relative simplicity of this model, a huge amount of effort goes into its maintenance, and library systems seem rarely to do exactly what one would want. It remains to be seen how record creation, maintenance, retrieval and use will perform in this far more complex environment.

And now for something completely different...

To this point most of our efforts have been related to metadata schemes that have been developed by librarians, archivists, curators and other information professionals, or by government agencies or research initiatives such as the Data Documentation Initiative, the Federal Geographic Data Service, and the National Biological Infrastructure Initiative. To date we have not focused much attention on schemes coming out of the publishing community. However, these may ultimately have the greatest impact on traditional bibliographic description.

Several international efforts have been proceeding more or less simultaneously. The best-known in the library community is probably the INDECS (Interoperability of Data in E-Commerce Systems) project. INDECS was funded by the European Commission and supported by major trade associations representing record companies, music publishers, film companies, and book and journal publishers. The goal of the project was to create a framework for electronic trading of intellectual property rights in all media, and the primary product was a metadata model which is due for release in final form this summer. (Although the original project officially ended in March 2000, its work may be carried on by a not-for-profit membership organization.)

The INDECS model is essentially a semantic model for describing intellectual property, the parties that create and trade it, and the agreements that they make about it. The assumption is that many different metadata schemes will be developed and used by specific industries (for example, music and book publishers), and that it must be possible for this metadata to be exchanged between industries and reused in different contexts for global electronic commerce to thrive. INDECS attempts to distill the potentially infinite range of descriptive elements pertaining to rights into a defined set of generic, universally applicable attributes and values. Data can be exchanged between domain-specific metadata schemes if they follow or can be mapped to the INDECS data dictionary. The example often given is that the different schemes may recognize screenplay adapters, translators or musical arrangers, but translated to

INDECS, these are all specific examples of a generic category (contributor agent role) and value (modifier).

INDECS principles impose other constraints on metadata schemes as well. Because rights can be traded at any level of the IFLA model (works, expressions, manifestations, items) good descriptive metadata will not conflate these levels, and will provide for extensive, explicit linking between them. Because virtually any element of descriptive metadata can be an element of a rights agreement (except titles), the values of elements must be strictly authority-controlled and stored as unique, coded values. Because rights agreements depend on metadata, the authority for any item of metadata must be securely identified.

While work was proceeding on the INDECS framework, the Association of American Publishers (AAP) developed over a short period in 1999 a metadata element set for exchanging product information for the book trade. Called the Guidelines for Online Information Exchange (ONIX), the specification was released in version 1.0 in January, 2000. ONIX was a direct response to the enormous growth in online booksales, which has resulted in a need for publishers, booksellers and distributors to create and exchange vastly expanded metadata for saleable items. The introduction to ONIX 1.0 points out that "Books with cover images, descriptions, reviews and additional information online outsell books without that information eight to one."

The same month that ONIX 1.0 was published, the EPICS Data Dictionary version 3.02 was released by EDItEUR, an international book and serials industry standards group. EPICS was developed as a joint project of EDItEUR, the Book Industry Communication (BIC) in the UK, and the Book and Serials Industry Communication (BASIC) in the US. Like ONIX, EPICS is a metadata specification designed for exchanging product information, motivated partly by the rise of Internet bookselling, and covering bibliographic, promotional and trade information.

Work immediately began to unite the two efforts. A new version of ONIX, consistent with EPICS and intended for both U.S. and European implementation, was released in May 2000 under the name ONIX International 1.01.[18] EPICS has been redefined as a more comprehensive data dictionary of which ONIX is a book industry subset, and the broader EPICS is being expanded to other areas, starting with audio-visual materials. Both schemes will be maintained by EDItEUR under the direction of a single international steering committee. These are extremely fast-moving standards and it is likely there will be additional developments between the time of this writing and the Bicentennial Conference on Bibliographic Control. The comments below are based on EPICS 3.02.

The EPICS data dictionary was developed coterminously with the INDECS project and has increasingly adopted the INDECS data model; it is seen as one of the first INDECS-compliant metadata standards. It should be possible to map EPICS elements to generic INDECS elements. Also in keeping with the INDECS model, EPICS requires precise and granular identification of all data elements in the scheme, supports both text and authority-controlled codes for nearly all data values, and allows extensive relationships between the described object and other objects to be specifically recorded.

It is interesting to compare the semantics of the more bibliographic elements of EPICS with those of traditional library cataloging. For a title, for example, it is possible to identify the type of title (title on piece, constructed title, alternative title, ISSN key title, etc.) and various subelements within a title, such as "title prefix" (which would be noted as non-filing characters in a 245), title, subtitle, etc. The semantic overlap is imperfect but substantial. However, there are major differences in the conceptual structure. An EPICS title cannot carry a statement of responsibility such as carried in the 245 subfield c, which is in fact information strictly pertaining to contributors and their roles. Similarly a former title, classified as a title in MARC (grouped in the 24x block), would be classified in EPICS as an element pertaining to a related object.

There is also a difference in the approach to content rules. The AACR2 approach is to take care in the selection of recorded data. There are extensive rules governing the selection of main entry, the justification of added entries, the chief source of information for title, etc. The publishers' approach is to allow the recording of any data so long as the nature of the data is explicitly recorded. The names and roles of all contributors can be recorded, as can the names and types of all titles. Selection of the most appropriate contributor or title for a particular purpose is not a function of the creator of the metadata, but rather of the user of the metadata (most likely a computer program). On the other hand, the publishers lean more strongly towards the use of coded values from named authority lists for the representation of content.

One reason that these approaches to resource description differ is because the underlying functional needs for the metadata differ. The publishing community is far more concerned with marketing and with managing intellectual property rights, while the library community has a need to manage huge inventories over a very long period of time. Nonetheless, both communities need to support end-user discovery and identification of information resources, and there is great overlap in the user tasks that must be supported by basic bibliographic data.

Much of our attention to date has been focused on what we might call specialty metadata schemes. While this has helped to increase our sophistication and understanding of metadata issues in general, and while it has surely enhanced access to important categories of materials, I would suggest that it is time to look through the other end of the telescope and begin thinking about basic bibliographic metadata as a commodity, produced and exchanged by a number of communities in order to serve a number of purposes. We are already in an environment where readers are as familiar with amazon.com as with their library catalogs. We are already in an environment where libraries purchase catalog records from any number of sources, from OCLC's PromptCat to our approval plan vendors. We will soon be in an environment where most metadata is exchanged in XML: the publishers have already adopted it, and library systems are moving in that direction. In this context it makes very little sense to think that libraries, publishers, booksellers, distributors and vendors will all be creating incompatible, non-reusable bibliographic metadata.

I am not sure myself what it means to think about commodity metadata. Perhaps that is something that can be explored in the context of the Bicentennial Conference. However I do urge librarians to take a serious and objective look at the metadata schema emerging in the publishing community with the long-term goal of maximizing the interchangeability of data. It will not be enough to simply develop mappings between these schemes and MARC. Experience with the TEI header and with crosswalks from DCMES to MARC has shown that simply mapping from a semantic or syntactical element in one scheme to a comparable element in another does not guarantee the usability of the converted metadata.

I suggest that we work proactively with publishers to establish enough commonality between our respective rulesets to allow meaningful exchange and reuse of metadata. Can we establish common authority lists? (For example, libraries use MARC relator codes and ONIX International uses contributor role codes -- is there a compelling reason for these to be different?) Are there content rules which, if shared, would substantially benefit both communities? Are there content rules in traditional library cataloging that don't make enough of a functional difference to insist upon? I also suggest that we evaluate the additional metadata elements designed to support the book trade for their potential use in our own systems. Does content such as author biographies, book review excerpts, and dust jacket summaries provide useful access points for retrieval, or help a user select an item appropriate to his needs (both end user tasks to be supported in FRBR)?

In sum, I suggest that the key question as we enter the new millenium is not bibliographic control of Web resources, but rather bibliographic control of both digital and non-digital resources in the Web environment. I expect that the Web environment will be characterized by the development of competing search engines and retrieval models, a proliferation of commercial and non-commercial bibliographic services, and the dominance of XML as a transport syntax for both data and metadata. Evidence indicates that successful metadata schemes must be flexible enough to accomodate unexpected users and uses, must have responsive mechanisms for change, must be based upon or work in conjunction with shared content rules, and must allow clear relationships to be established between different works and manifestations. While refining specialty metadata schemes, we should also work towards the development of a system of commodity metadata that will enable economic exchange, reuse and repurposing of metadata for current trade publications in all media.

ACKNOWLEDGEMENTS

Many thanks to John Price-Wilkin and David Martin for their reading of and helpful comments on sections of this paper.

NOTES

1. This paper will use the term "traditional cataloging" to refer to resource description based on a suite of rules including ISBD, AACR2, LC rule interpretations, LC name authority, and the

MARC formats for bibliographic data. I do this not to reflect a value judgement for or against traditional cataloging, but only because some short-hand term is needed.

2. "It is the intention of the developers, however, to ensure that the information required for a catalogue record be retrievable from the TEI file header, and moreover that the mapping from one to the other be as simple and straightforward as possible." C.M. Sperberg-McQueen and Lou Burnard, eds., Guidelines for Electronic Text Encoding and Interchange (TEI P3) (Chicago; Oxford : Text Encoding Initiative, c1994.) p.137. http://www.uic.edu/orgs/tei/p3/.
3. Committee on Cataloging: Description and Access, Task Force on Metadata and the Cataloging Rules. Final Report. August 21, 1998. http://www.ala.org/alcts/organization/ccs/ccda/tf-tei2.html.
4. TEI/MARC "Best Practices", November 25, 1998 Draft. http://www.lib.umich.edu/libhome/ocu/teiguide.html.
5. International Federation of Library Associations and Institutions. Functional Requirements of Bibliographic Records: Final Report. September 1997. http://www.ifla.org/VII/s13/frbr/frbr1.htm#1
6. Kathleen Burnett, Kwong Bor Ng and Soyeon Park. "A comparison of the two traditions of metadata development." Journal of the American Society for Information Science 50(13):1209-1217, 1999.
7. Daniel V. Pitti. "Encoded Archival Description: An Introduction and Overview." D-Lib Magazine, 5(11) November 1999. http://www.dlib.org/dlib/november99/11pitti.html
8. Actually there are three sections; an optional section can be included to supply a more "publisher-friendly" title page than the header provides.
9. You can almost hear the surprised delight in early testimonials to the EAD, such as this quote from a talk by Susan von Salis, Schlesinger Library at Radcliffe College, to the RLG Forum in Toronto, 1997. "As I mentioned, most finding aids include common components such as provenance, scope and contents, and access restrictions. So..... the DTD includes these 'parts' as its elements! Markup itself is simply a matter of wrapping the correct tags around the proper text." http://www.lib.umb.edu/newengarch/InternetResources/vonsalisrlg/index.html
10. Report of the Emerging Descriptive Standards Group, Southeastern Archives and Records Conference, Columbia SC, May 23-25, 1999. http://www.state.sc.us/scdah/sarc41999.htm
11. MacKenzie Smith. "DFAS: The Distributed Finding Aid Search System." D-Lib Magazine, 6(1) January 2000. http://www.dlib.org/dlib/january00/01smith.html
12. Consortium for the Computer Interchange of Museum Information. Guide to Best Practice: Dublin Core, version 1.1, April 2000. Available from http://www.cimi.org/standards/index.html#FIVE.
13. Renato Iannella and Rachel Heery. Dublin Core Metadata Initiative - Structure and Operation. April 1999. http://purl.org/dc/about/DCMIStructure-19990531.htm
14. Lynda S. White. "Creating the VRA Core: The Critical Issues." VRA Bulletin, v.25, no.4 (Winter 1998), p. 34-40.
15. Marcia Lei Zeng. "Metadata Elements for Object Description and Representation: A Case Report from a Digitized Historical Fashion Collection Project". Journal of the American Society for Information Science 50(13):1193-1208, 1999.
16. Bernhard Eversburg summarized the principle of 1:1 with the following verse:

Make metadata one to one,
just one per item, is the task.
Rather less,
more's a mess!
"But what's an item", now you ask?
If that's in doubt, do none.

http://www.mailbase.ac.uk/lists/dc-general/1999-04/0117.html Nonetheless, after extensive debate over whether Ansel Adams or the scanning technician is the Creator of a digitized Adam's photo, the answer appears to be that the Creator is in the eyes of the beholder.

17. International Council on Archives. ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families. Ottawa : The Secretariat of the ICA Ad Hoc Commission on Descriptive Standards, 1996. http://dobc.unipv.it/obc/add/infap/archdes/isaar_e.html
18. ONIX International Version 1.01. http://www.editeur.org/onixfiles.html

Library of Congress
January 23, 2001
Comments: lcweb@loc.gov

Conference Home Page

What's new

Greetings from the Director for Cataloging

Topical discussion groups

LC21: A Digital Strategy for the Library of Congress

Conference program

Speakers, commentators, and papers

Conference sponsors

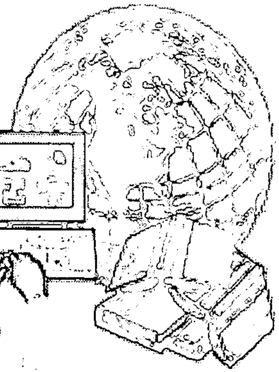
Conference discussion list

Logistical information for conference participants

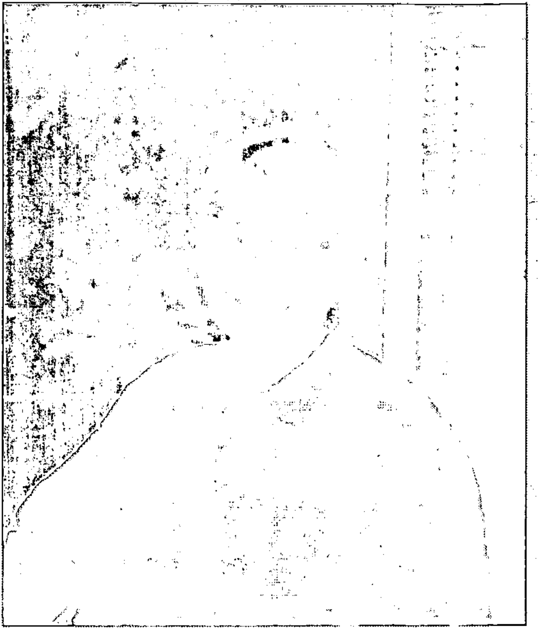
Conference Organizing Team

Bicentennial Conference on Bibliographic Control for the New Millennium

Confronting the Challenges of Networked Resources and the Web



sponsored by the Library of Congress Cataloging Directorate



Priscilla Caplan

Assistant Director for Digital Library Services
Florida Center for Library Automation
2002 NW 13 St., Suite 320
Gainesville, FL 32609

International Metadata Initiatives: Lessons in Bibliographic Control

About the presenter:

Priscilla Caplan has been at the Florida Center for Library Automation since Aug. 1999. Previously, she served as Assistant Director for Library Systems, University of Chicago Library, from Aug. 1993-July 1999, and as Head, Systems Development Division, Office for Information Systems, Harvard University Library, from July 1985-July 1993. Her professional activities include being co-chair of the Dublin Core (DC) Standardization Working Group (1999-) and member of the DC Advisory Committee (1998-); chair, National Information Standards Organization (NISO), Standards Development Committee (1997-) and member of the NISO Board of Directors (1998-); Lecturer, Dominican University, School of Library and Information Science (July 1998-July 1999) ; Director, CUIP Digital Library, Chicago Public Schools/University of Chicago Internet Project (Nov. 1997-July 1999); member of the Digital Library Federation, Architecture Committee (1998-1999); and member (1991-1993, 1993-1995 terms) and chair (1995-1996) of MARBI. Caplan has written extensively on metadata and related issues which have been published in The Cybrarian's Manual 2, D-Lib Magazine, Public Access Computer Systems Review, The Serials Librarian, and Cataloging & Classification Quarterly

Full text of paper is available

Cataloging Directorate Home Page

Library of Congress Home Page

Summary:

The decade of the 1990s saw the development of a proliferation of metadata element sets for resource description. This paper looks at a subset of these metadata schemes in more detail: the TEI header, EAD, Dublin Core, and VRA Core. It looks at why they developed as they did, major points of difference from traditional (AACR2/MARC) library cataloging, and what advantages they offer to their user communities. It also discusses challenges to implementers of these schemes and possible future developments. It goes on to identify some commonalties among these cases, and to attempt to generalize from these some lessons for developers of metadata element sets. It concludes by suggesting we also look carefully at emerging schemes being developed by publishers in support of electronic commerce and rights management, and think seriously about the implications of commodity metadata upon our traditional bibliographic apparatus.