DOCUMENT RESUME

ED 454 869 IR 058 160

AUTHOR Arms, Caroline R.
TITLE Some Observations on Metadata and Digital Libraries.
PUB DATE 2000-11-00
NOTE 22p.; 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/arms_paper.html.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Access to Information; Cataloging; Comparative Analysis; *Electronic Libraries; *Information Retrieval; *Metadata; Models; User Needs (Information)
IDENTIFIERS American Memory Project (Library of Congress); *Electronic Resources

ABSTRACT

This paper describes experiences in gathering together metadata from heterogeneous sources for the American Memory project of the Library of Congress, particularly for the collections digitized and cataloged at other institutions. It also reflects on several initiatives to develop rich structured metadata schemes for specific domains and to find simple approaches to support resource discovery across domains. Trends and commonalities are identified, and influences among metadata schemes are explored. Highlights include: differences in digital libraries; objectives for metadata and expectations of users; community-specific metadata models and schemas; metadata for cross-domain discovery; types, formats, and genres of digital content; metadata for search and metadata for display; how users search in digital libraries; searching by topic, originator, date range, place, or type; and improved tools to support access to resources in digital libraries. A table compares search buckets for metadata for the American Memory and Alexandria Digital Library projects. (Contains 34 notes.) (MES)

Some Observations on Metadata and Digital Libraries

Caroline R. Arms
Information Technology Services
Library of Congress
Washington, D.C. 20540

Final version

> "By the year 2000, information and knowledge may be as important as mobility. We are assuming that the average man of that year may make a capital investment in an "intermedium" or "console"--his intellectual Ford or Cadillac--comparable to the investment he makes now in an automobile, or that he will rent one from a public utility that handles information processing as Consolidated Edison handles electric power. In business, government, and education, the concept of "desk" may be primarily a display-and-control system in a telecommunication-telecomputation system--and its most vital part may be the cable ("umbilical cord") that connects it, via a wall socket, into the procognitive utility net."

J. C. R. Licklider. *Libraries of the Future*. M.I.T. Press, 1965

The words above may not be those in use today, but the prescience of these sentences from thirty-five years ago is amazing. At the time, computers were used by "collecting data and writing a computer program, having the data and program punched into cards, delivering the cards to a computer center in the morning, and picking up a pile of 'printouts' in the afternoon." Time-sharing computing systems with typewriting terminals for remote users were just emerging from the research laboratory for practical use, as was the use of cathode-ray devices as terminals. The book, *Libraries of the Future*, was based on a two-year study sponsored by the Council on Library Resources and carried out, under Licklider's leadership, by a group of engineers and psychologists from Bolt Beranek and Newman, Inc. (BBN) and the Massachusetts Institute of Technology (MIT) starting in late 1961. Charged by the Council to explore how developing technologies might shape libraries in the year 2000, the group envisioned a much closer "interaction with the fund of knowledge"(1) than print libraries can support. They saw the fund of scientific knowledge directly available not only to scientists but to their experiments; they envisioned the ability to feed research results directly back into the fund of knowledge. Licklider would surely be delighted to see the systems for accumulating and using genome resources today. This vision was so different from the library of the early 1960s that it seemed appropriate to use a different term; the term chosen was "procognitive system." In the year 2000, the Internet, with all the information resources to

which it provides access, serves as the "procognitive utility net" that Licklider predicted.(2) Communities, institutions, and individuals have been building digital libraries as they work towards the vision they share with Licklider and his colleagues for a richer, closer interaction with the fund of knowledge.

Libraries have always supported interactions with the fund of knowledge, interactions that come in many shapes and sizes. Libraries support scholarly communication and formal education; they also help people find facts, figures, and tax regulations. Interacting with knowledge is what lifelong learning is all about. Users of American Memory register delight when they find pictures of the town where they grew up or recognize a family member in a picture, sound recording, or letter. Genealogical research is immensely popular. Contributions to the fund of knowledge come from many sources, including individuals as amateurs. Vast numbers of informative web pages represent personal contributions to the fund of knowledge by enthusiasts: railroad buffs; music-lovers; naval history mavens; watchers of birds, badgers, and other creatures; and many more.(3) I will not argue that the World Wide Web is a digital library (although some do). I will limit my concept of a digital library to collections of resources in digital form assembled for a particular community or purpose and managed with an intention of ongoing accessibility and usability. With this loose definition, perhaps what distinguishes a digital library from a set of documents or web pages is the existence of some formalized, structured metadata (data about data) to provide organized access to a body of resources.

User communities are building domain-specific digital libraries with domain-specific metadata schemas and guidelines.(4) This should be no surprise. Domain-specific controlled vocabularies and abstracting and indexing services are not a new phenomenon and the full-text databases that some of these services have developed into are certainly digital libraries. Even within the traditional cataloging community, descriptive practices vary for classes of material. Practices developed originally for cataloging books have been adapted and extended over 150 years (if one considers the plans developed by Sir Anthony Panizzi for organizing books in the British Library as the starting-point for modern bibliographic practice). As Elaine Svenonius points out in her recent book, *The Intellectual Foundation of Information Organization*, these practices "have been jolted in the twentieth century by information explosions, the computer revolution, the proliferation of new media, and the drive toward universal bibliographic control. How they have withstood these jolts, where they have remained firm, where they have cracked, and where cracked how they have been repaired or still await repair is a dramatic -- and instructive -- history for those interested in organizing information intelligently." (5) For example, in response to the jolt of incorporating non-book media, the Anglo-American Cataloguing Rules have been supplemented by manuals for other classes of material: *Archives, Personal Papers, and Manuscripts* (APPM); *Graphic Materials* (GIHC); and *Archival Moving Image Materials* (AMIM). An archival collection of personal papers is typically cataloged in a single collection-level record. Very different descriptive and organizational practices have been developed by archivists to organize and describe collections at whatever finer level of granularity is deemed appropriate.

Varying descriptive practices have taken into account not only the observed intellectual needs of the traditional users of the resources, but also more pragmatic factors: the mission and capabilities of the traditional custodian; economic realities (manpower and funding); technical realities (tools available to help custodial institutions prepare and users to take advantage of metadata); the pattern of updating

required; and the physical nature of the artifacts themselves. The nature of today's bibliographic systems make allowance for synergies with (or are constrained by exigencies of) inventory control and are shaped by the importance in the published literature of relationships expressed by shared author, title, and subject headings. For archivists, two of the key factors shaping practice are the importance of the integrity of a collection as a whole and the sheer impracticality of describing individual items in detail. Although catalogs in book form were replaced by card catalogs early in the twentieth century, the document form of an archival finding aid has remained useful. For museums, the importance of detailed information about provenance and the historical and creative context for individual items has led to yet another set of practices. Many of these factors do not change simply because reproductions or surrogates can be created in digital form or when today's analogs are created in digital form. Even among the communities with preservation of the cultural record in their mission, there will continue to be heterogeneity of descriptive practice in digital libraries, and for good reason. One challenge is to identify the pragmatic factors that have changed or will inevitably change and adapt to them. Underlying principles grounded in users' needs should still guide practice.

What is different in a digital library?

The networked world of cyberspace and the development of advanced computational tools are shifting the balance among factors that shaped past descriptive practices. Although developed for a different purpose, the model Lawrence Lessig introduced in *Code and other Laws of Cyberspace* sheds light on this balance (for this author, at least). Lessig suggests that an individual's behavior is constrained by *law, architecture, market, and norms*.(6) He stresses the fundamental differences between the architecture of physical space (where distance, weight, and walls set limits) and the architecture of cyberspace (manifested in software, network equipment, and protocols). The four factors are clearly interdependent; indeed, in response to societal norms, laws are passed to regulate architecture and the market. Turn to the business section of any newspaper, and it becomes obvious that the architecture of cyberspace is changing the market (and the overall economic environment) in which businesses, individuals, and libraries operate. Cyberspace offers new cost structures, users from new communities, users from traditional communities with new expectations, and new tools to serve those users. Clearly, understanding of the surrounding "architecture" and "market" will guide the development of future practices and systems for preparing and using metadata in digital libraries.

The architecture and market for print publishing has been relatively stable and libraries are adapted for that environment. For digital libraries, the environment is likely to be in a state of change for the foreseeable future; research and experimentation will be ongoing. For a thoughtful analysis of some of the metadata issues warranting research, see *Metadata for Digital Libraries: a Research Agenda*, developed by a joint task working group established under the auspices of the European Union and the National Science Foundation.(7) The current article should not be seen as an attempt to develop an overarching theoretical or technical framework or as a comprehensive overview, but as observations from the trenches of American Memory, a production digital library system that is also an experiment. The integration of heterogeneous content, including content and metadata prepared by other institutions, into American Memory has provided a close look at practical hurdles in the path to Licklider's vision. It has also stimulated an appreciation for how varying descriptive traditions can contribute to achieving that

vision and a not infrequent sense of frustration at the difficulty of incorporating new tools to build better services and enrich the interaction with the fund of knowledge.

Objectives for metadata and expectations of users

In the IFLA *Functional Requirements for the Bibliographic Record* four objectives are listed for the bibliographic record for an entity: to enable a user to *find, identify, select,* and *obtain access* to the entity described. Elaine Svenonius suggests that it is helpful to distinguish between the objective of *finding* or *locating* a particular entity (known item) and the *collocation* objective, which allows a user to find sets of entities, for instance works by the same author or on the same topic.(8) She also suggests that a *navigation* objective be added to reflect the wish of users to find other works related to a given work.(9) The Dublin Core Metadata Initiative(10) describes the Dublin Core metadata set as facilitating *discovery*; this term is often applied to the process of looking for resources in a broad cross-domain universe. In a digital library, metadata may be needed to help the user not only *select* an item appropriate for the current purpose but also to *use* it. Such metadata may include structural metadata that permits navigation among components of a single intellectual "object" (for example, to skip to a particular segment of a digital sound recording) or information about terms and conditions in a machine-parsable form that can be used to limit access to authorized users. Metadata is also required to manage content and to record information that will support future preservation activities. In this article, the focus is on metadata that leads users to resources in digital libraries (*discovery, finding, collocation, navigation*) and lets them choose resources for the task at hand (*identification, selection*).

The networked world has changed the expectations of information users by removing physical constraints. In the physical world, few question the practice of storing maps or manuscripts separately from books or locating the specialists who help users access and use the resources in different areas of a library or in different libraries. In the networked world, the digital libraries hold the potential to brings resources of many types together to the user's laboratory or desktop (or even palmtop). Some users will want first to cast a wide net as they trawl for information and then to draw it tight and precisely around the most relevant and usable resources for a particular purpose. The wide net calls for interoperability and commonality across domains and custodial communities, while assessment of fitness for a particular purpose often calls for rich and domain-specific metadata.

One characteristic of a digital library is the accessibility of the content to the user. In the traditional library, subject to the physical constraints of weight and distance, the user who selects lots of items that turn out not to be fit for the purpose at hand will incur a high cost (in effort, if not in monetary terms). If users can delve straight into the content to aid the act of selection, some metadata conventionally recorded to support selection may be less essential. Users and builders of digital libraries also recognize the potential for navigation among related items. This feature was the fundamental function underlying the initial success of the World Wide Web. Online, citations can lead directly to cited works and works can link directly to datasets or models they describe. Reference links and thumbnail images change the architecture for research and information seeking. Users can navigate using relationships between metadata records and relationships between resources from their desktop. The functions that metadata must support may not change in a digital library, but, in the new architecture and with new tools, it seems

likely that metadata practices must evolve to provide the functionality cost-effectively in the new market.

Content is also accessible for building services in a digital library. Metadata or surrogates can be derived or extracted from content files. Today, a thumbnail is essentially part of the descriptive record for an image. In the collection of negative strips from the Farm Security Administration, many of the images were never captioned; access to the uncaptioned images within American Memory is primarily through navigation. Contact sheet displays of thumbnails are ordered using the processing number sequence on the negatives (an identifier that implies chronological order, at least for negatives on a single strip). Automatic summarization and analysis of other forms of content is an active research area. The value of including basic metadata in the header of text marked up in SGML or XML is also well recognized. Text conversion projects following the Text Encoding Initiative (TEI) guidelines often use a mapping between elements in the header section of the marked up file to fields in a MARC catalog record. In some cases, the TEI header is derived from the catalog record; in other cases, a catalog record is derived from the header. For material published in digital form, the publishing community will maintain basic bibliographic metadata for its own purposes, including promotion and business-to-business dealings with booksellers. It will be inexpensive to include it in file headers. For example, the Open eBook Publication Structure specification includes tags for "publication metadata."(11) In a digital library, searchable text can minimize the need for metadata (particularly when items do not merit the expense of individual cataloging). American Memory has a popular collection, *American Life Histories*, for which the only item-specific metadata is a title (a display string needed for result lists) and an identifier; the text itself is searched to provide intellectual access.

Creators of non-text materials also recognize the value in embedding such metadata within the files. Today, cameras can record the date and time of an exposure. If there is not already a camera with built-in GPS to record location, there will be very soon. Proposals for new digital file formats, such as JPEG2000 (12) (for images) and MPEG-7 (13) (primarily for sound and video) include the ability to embed descriptive metadata within the file.

Community-specific metadata models and schemas

The educational community has been active in exploring the functional needs of educators for finding and using instructional materials. Resources are available because of economic incentives to establish and manage online learning environments and government efforts to improve education. Some information tasks call for rich metadata. For example, a teacher may be looking for material to help him explain a very specific topic (say cell division in an embryo) in a particular class. The teacher wants to be confident that the material makes appropriate assumptions about what the students already know, has already been used successfully in the classroom, will occupy an appropriate amount of class time, and will work on the equipment available. This level of specificity may be available in a service built by and for educators. A metadata schema for instructional resources has been developed by the IMS Global Learning Consortium, Inc., a global consortium with members from educational, commercial, and government organizations.(14)

Another area with specialized metadata needs is that of geospatial resources. The ability to build maps

dynamically or relate geospatial facts from distributed sources of information is enhanced by commonality of metadata. Such capabilities can support government tasks such as emergency management, city planning, and tracking air quality or global climate change. The Federal Geographic Data Committee has developed a Content Standard for Digital Geospatial Metadata.(15) There is worldwide standardization activity in this area in relation to metadata schemas and to the content standards for elements. The activity involves government, commercial and educational sectors. As with the instructional metadata schemas, the functionality required by the creator and user communities, rather than traditional libraries, is driving these activities.

The Visual Resources Association has developed a two-level hierarchical model for describing objects or visual works (such as paintings, sculptures, or buildings) and images of those works. A single set of metadata elements, the VRA Core Categories 3.0 (16), can be applied to the works and to the images. This approach follows the so-called "1-1" principle, distinguishing characteristics of the image surrogate clearly from characteristics of the work. This principle emerged from the Dublin Core community but has provoked considerable controversy. It can not be applied in existing "flat" bibliographic systems without creating awkwardness and confusion for users through multiplicity of records and a burden on cataloging processes by requiring replication of work-level information in records for each surrogate image. A system that takes advantage of the two-level structure must allow searching across all records but present results that pull in work records automatically when a "hit" is at the image level. The metadata schema used by the Art Museum Image Consortium (AMICO) has a similar hierarchical structure.(17) The Visual Information Access (VIA) system at Harvard University uses a three-level hierarchy.(18) Although the Encoded Archival Description (EAD) standard has been used primarily for describing collections of papers and records that have not been digitized, it too provides a hierarchical structure for description at different levels.(19) The EAD metadata structure has been used effectively as the basis for digital library services, for example at the University of California, Berkeley(20) (and now at the California Digital Library) and Duke University.(21)

A more complex conceptual model has been proposed by the Documentation Standards Group International Committee for Documentation of the International Council of Museums (ICOM-CIDOC).(22) The CIDOC model is an object-oriented reference model that expresses a much more complex knowledge universe than the simple relationship of a descriptive record to a resource or even of a hierarchical structure of related descriptions and resources. It allows for descriptions not only of works, images, document, (conceptual objects) and objects (physical entities), but also of people (actors), places, and periods (time-spans). This model will form the basis for the Cultural Materials Initiative digital library project at the Research Libraries Group. Full use of such a model will integrate gazetteers, biographical dictionaries, and encyclopedias, going well beyond the traditional use of authority records and thesauri.

One example of a digital library enriched by integrating the use of reference resources such as biographical dictionaries and the Thesaurus of Geographic Names is *Perseus*, housed at Tufts University.(23) In this digital library for the study of the ancient world, the books "talk" to each other. Names of places and people been identified automatically in the text and can be used as links to related information elsewhere in the corpus. This includes automated disambiguation of different places with the

same name (e.g., Springfield) and of references to people who share names with places (e.g., Lincoln). Licklider would have been delighted. In 1965, he complained that "when it comes to organizing the body of knowledge, or even indexing and abstracting it, books by themselves make no active contribution at all."(24)

Some of these rich metadata schemas or models have great potential for enriching the interaction with knowledge for users through collocation and navigation using the relationships expressed in the models and supporting knowledge organization systems, such as gazetteers, name authority files, and thesauri. The models are, however, unfamiliar and will require new tools to implement and deploy. The web sites that present them usually have FAQ (Frequently Asked Question) pages that emphasize that most features are optional. There is plenty of evidence that, unless there is strong economic motivation, introductory guidelines and basic tools are needed before complex standards gain broad acceptance. Part of the genius shown by Tim Berners-Lee in his original standards for the World Wide Web was the simplicity of the specifications. The conceptual model could be explained on one slide and fleshed out in three more. Implementing a server was a few lines of code; a young programmer built a graphical interface (Mosaic) as a side project and the rest is history. The other aspect of Berners-Lee's genius was the instant integration of legacy content by supporting earlier Internet protocols now seldom mentioned (gopher, wais, news). Real-world experience with the simple (e.g. HTML) has led iteratively to better understanding of which extensions to its functionality are most essential. Complex metadata schemas present a challenge for those with valuable legacy metadata to migrate or metadata maintained in rich, but different schemas. It is relatively easy to "dumb down" metadata records from a rich schema into a simpler one, for purposes of interoperable retrieval, while still maintaining the full richness in a master system. Transforming from one rich schema to another is usually more expensive and the benefits may not be obvious to the institutions or individuals most likely to bear the cost.

Metadata for cross-domain discovery

At the other end of the spectrum, some digital library activities are focusing on allowing users to find resources across an information universe that spans communities, nations, types of information, and types of institution. The Dublin Core Metadata Initiative has been building consensus through a series of workshops and working group activities since March 1995. From the start this has been an international effort to develop a common core of semantics for resource description. The path has not been easy. Active debates have highlighted the differing priorities and expectations of different communities. The Dublin Core Element Set (Simple Dublin Core) was submitted to NISO as a draft standard (Z39.85-xxx) this year.(25) The set has 15 elements (listed in the left-hand column in Table 1); all elements are optional and all repeatable. The elements themselves are unlikely to be sufficient as an internal metadata schema for any particular project or application. Some proponents see the elements as the basis for a schema that can be extended by adding or refining elements; others prefer to see the set as a view of a richer, more complex description.(26) This view can be used as a framework for mapping different element sets into a common set for indexing and searching. The DCMI has also developed a model for extending the simple element set while maintaining the objective of interoperability. Elements can be refined. For example, Date.Created and Date.Issued are refinements of Date. Element refinements are guided by the "dumb-down" principle. If the refinement term is not recognized, it should be reasonable to treat the value of the

qualified element as if it had no qualifier. This is an extremely important principle for supporting broad interoperability. The second type of qualification for Dublin Core elements is to specify an encoding scheme or controlled vocabulary. In July 2000, a set of exemplary qualifiers was published as a result of proposals made by working groups at the DC-7 workshop in October 1999.(27) Tom Baker describes Dublin Core as a pidgin language in a discussion of simple and qualified Dublin Core.(28) The existence of the 15-element set has provided an important focus for other interoperability initiatives.

Simple Dublin Core was recently adopted as the core metadata record format for the experimental Open Archives initiative. This initiative is testing whether a simple mechanism that allows service providers to harvest metadata records from content repositories will facilitate and stimulate the development of valuable services that draw on content from many repositories. An initial impetus was the belief that access services (such as reference-linking, portals, and selective dissemination services) could benefit from harvesting records for e-prints and other "grey" literature. Records will usually have links back to the host repository through persistent identifiers for the full content. The concept also allows the development of comprehensive search services that include significant web-accessible resources currently hidden from the "spiders" that crawl the web on behalf of search engines. The so-called "deep web" includes the content of most digital libraries, such as American Memory, whose web presence is largely ephemeral, with records retrieved from a database in response to each search and displays generated dynamically. The Open Archives harvesting framework includes the ability to harvest records in other metadata formats (e.g. MARC or the rfc1807 bibliographic format used for the Networked Computer Science Technical Reports Library). The Open Archives initiative is based on the premise that simple records (almost certainly derived dynamically from a more complex schema used internally) provide the first step to cross-domain discovery. Service-builders can choose to harvest records in a richer schema when available. Specifications for records marked up in the Extensible Markup Language (XML) must be made accessible for any schema used.

A third activity aimed at cross-domain discovery is the development of a very basic profile for the Z39.50 information retrieval protocol for cross-domain discovery. This specification also supports Simple Dublin Core marked up in XML as a transfer format for records. In many ways, the stimulus for this activity is the success of the World Wide Web. People clearly find value in web search engines, whatever their shortcomings. An enormous mass of information accessible from a single search box has clear appeal; many people prefer to try several queries and skim through pages of hits than to use Boolean queries to increase precision. The computational and linguistic tools built into commercial search engines are enhanced frequently. The architecture of cyberspace has changed the relationship between bibliographic control and access. Today, ironically, resources under good bibliographic control are likely to be less widely accessible than those simply mounted on the web. The motivation behind these simple-minded interoperability efforts is to encourage broad access to resources of value.

Types, formats, and genres of digital content

The fund of knowledge is represented by a much richer set of resources than static pages on paper, and resources beyond those traditionally found in libraries, even multi-faceted libraries like the Library of Congress. Knowledge has always been conveyed through buildings (and the archaeological sites they

become), works of art, physical specimens of flora and fauna, artifacts of different cultures and lifestyles, and human memories. Photography, sound recordings, and motion pictures have added to the fund of knowledge both in their own right as means of expression and as richer surrogates than words on paper. The digital era has added not only reproductions and analogs of older forms of information, but also new digital resources of enormous variety. Some fall into obvious categories. The broad category of datasets includes census and other survey results, gene sequences, images and other sensor data from space, geospatial information that allows maps to be generated dynamically, decades of financial statements for publicly traded corporations, directories of people and places, and structured lexical resources, such as dictionaries and thesauri. More complex digital resources include mathematical and chemical knowledge expressed in structural forms that permit dynamic manipulation (and the software that performs or lets users perform the manipulation), interactive software for education and entertainment, and collections of re-usable "open source" software code. Digital libraries are being built to manage, serve, and support discovery of all these categories of resource. Some of these digital libraries are extensions of traditional libraries; many have developed from other well-established activities in organizing information (for example, for collections of social science datasets). Dynamic information resources, such as the web site that delivers up-to-the minute details of event and results at the Olympic Games challenge all traditional practices for organizing and recording for posterity. However, this too could be represented as a series of snapshots of bit-patterns, a digital resource, a set of computer files. The metadata required to support discovery and use of digital resources must clearly represent the intellectual nature and genre of the content; in a digital library, the fact that a resource can be represented by 0s and 1s is an assumption, not a useful categorization.

Attempts to develop general hierarchical categorizations for genres or types of information have usually failed. Even within American Memory, the content does not fit into a neat hierarchy. Are maps a subclass of images? How do you relate page-images of sheet music, song transcriptions, and recordings of performances? The Type working group of the Dublin Core Metadata Initiative developed, with much debate and without unanimity even in a small group, a high-level list of types (the DCMI Type Vocabulary: DCT1): Collection, Dataset, Event, Image, Interactive Resource, Service, Software, Sound, Text.

MARC records can hold information about the type of a resource in several ways. Svenonius notes that there are seven different places where a "document type" can be indicated.(29) Each element or indicator has a different set of possible values. Given the different guidelines for use and the different functions these seven elements serve, type information in MARC records has proved impossible to use uniformly within American Memory. Type indicators in the header are excellent as triggers for systems based on MARC records. They indicate what guidelines have been applied to content fields and therefore can be used to configure appropriate displays or procedures. However, the coded values can not be incorporated into a general keyword index, and are therefore unavailable to a user as a search term as they expect. Elsewhere, Svenonius remarks, "the use of one device to serve multiple functions, ..., while favored by the principle of parsimony, nevertheless introduces a lack of flexibility that can be an obstacle as technology changes."(30) The principle of parsimony results in type information for some classes of material appearing only within a complex physical description not designed for machine parsing.

The DCT1 list by itself would not prove sufficient for item-level genre distinctions in American Memory. Several of the categories don't apply to a body of converted analog materials and those that do not support the *selection* objective adequately for the typical American Memory user. Svenonius argues, "For document types, as for general-format types, it is not possible to construct a classification that is both natural and whose categories are mutually exclusive." American Memory experience supports her argument. The current feeling of the Dublin Core Type working group is that some communities will develop controlled lists of terms, but agreement across communities on the important categories or even common definitions for the same terms is unlikely. Lacking initial agreement on an acceptable "standard" typology, American Memory does not have explicit type values in all metadata records. This shortcoming means that searches limited to images may retrieve other categories of content, since the limit is actually by collection and some collections included will have text or sound as well as images. Colleagues agree that adding high-level type information consistently across the metadata records would provide more benefit for American Memory users than any other change. Interoperability will certainly be well served if descriptive records shared or exchanged always include type information, even if all content within the repository or collection providing the data is of the same type. Based on experience with American Memory, users might be best served by the inclusion of any applicable terms from the DCT1 list **and** additional terms at finer levels of specificity.

Metadata for search and metadata for display

The user's objectives are supported not by raw metadata, but through the tools and systems that can take advantage of the metadata. In both library catalogs and digital library services, the functionality for discovery, finding, and collocating is determined not by the metadata but also by the indexes constructed for that metadata. Different systems provide different options for configuring indexes. Public interfaces to library catalogs usually combine different metadata elements (e.g., MARC fields and subfields) into a relatively small set of indexes. For example, a keyword search by subject may find the term in any of the MARC subject fields (e.g. personal names, terms from authorized vocabularies, uncontrolled terms, genre terms, etc.). Once a record has been retrieved, elements can be labeled more specifically.

Some digital libraries, including American Memory, take the same approach. Individual elements that may be usefully distinguished (and labeled) to support the act of selection are lumped together to support discovery, finding or collocation. The University of Washington Libraries have used collection-specific element sets for their collections of digital reproductions; each set is mapped to the fifteen elements of the Dublin Core Metadata Element Set for cross-collection retrieval.(31) Retrieved records show the specific labels. For example, a collection of pictures of plants has a variety of fields relating to preferred soil quality and climate, whether the plant is native to the state, and other botanical details. For search purposes, these fields are all included in the Description index; on the display they are individually labeled. American Memory uses a similar approach. All descriptive notes are lumped into an overall text index; on display, a summary or abstract is usefully presented first and labeled as such. Many collections in American Memory call for unusual metadata elements, such as musical features for folk songs and descriptions of the key mechanisms in a collection of flutes. For searching, these are all treated as notes and included in the general keyword index. For display, however, the metadata format includes tags that are ignored by the indexer, but provide labels for use in the record display.

The indexing approach currently used in American Memory was developed empirically and iteratively, based on data elements usually available across heterogeneous sources and expectations of content custodians and users; it has seven primary indexes that support searching of metadata (Title, Alternative Title, Creator, Contributor, Subject, Any Text, Number). The developers of the Alexandria Digital Earth Prototype at the University of California, Santa Barbara have described the framework they use for querying metadata from distributed sources. Based on experience over several years of working with geo-referenced information, they chose eight search buckets (Geographic Location, Type, Format, Topical Text, Assigned Terms, Originator, Date Range, Identifier). For the most basic cross-domain discovery, the developers of the Bath Profile for Z39.50 identified Author, Title, Subject, Any. The Bath Profile effort is not strictly a digital library project, but has as an aim, interoperability between library catalog systems and "other electronic resource discovery services." Table 1 provides an informal tabular comparison of the clusters for indexing of these three projects and the alignment with Dublin Core.

Table 1: Comparison of search buckets for metadata for digital library projects

Dublin Core Metadata Element Set	**American Memory (local search buckets in parentheses)**	**Alexandria Digital Library (search buckets in bold)**	**Bath Profile (Z39.50) for Cross-Domain Discovery**
	Digital library of reproductions of historical sources (in text, image, sound, video)	Distributed digital library for geographically-referenced information	
Elements that support discovery			
Title	**Title** (TITL) **Alternative Title** (ALTTITL), usually searched with TITL.	(Topical text)	**Title**
Creator	**Creator** (AUTHOR)	(Originator)	**Author**
Contributor	**Contributor** (OTHER), usually searched with AUTHOR	(Originator)	
Publisher	(TEXT)	(Originator)	

Date	Display date (TEXT) Sort date (used only to sort search results within collections for which dates are known well enough to normalize)	(Date range)	
Subject	**Subject** (SUBJ)	**Assigned terms**	**Subject**
Coverage (spatial and temporal)	Geographic subject. (hierarchical placename, indexed as SUBJ, also used to support browsing of placenames and map-based selection by state)	**Geographic location** (footprint in geographic coordinates) **Temporal coverage** (indexed as Date range)	
Description	Summary (TEXT). Other notes (TEXT)	(Topical text)	
Type (genre)	(SUBJ)	**Type**	
Language (of resource)	Language (TEXT)		
Format (digital)		**Format** (for delivery, online or offline)	
	Any text elements, including textual fields in other indexes. (TEXT)	**Topical text** (includes title, description and any other text, including assigned terms)	**Any**
		Originator (includes creator, contributor, publisher)	
		Date range (includes date and temporal coverage)	
Elements that primarily support identification and navigation and use			
Identifier	**Identifier** (NUMBER)	**Identifier**	
Relation	Related items		

Source			
Rights			
	Repository	Reproduction number (NUMBER)	

How do users search in digital libraries?

The columns in Table 1 are compromises, reflecting a balance between what users ask for and what the metadata can support. Reading between the lines, I see a much stronger similarity in desired functionality for American Memory and the Alexandria Digital Library than the table would imply. The only significant difference is that American Memory users do search for titles, for example, for books and songs. With that exception, the search buckets used for the Alexandria Digital Library (ADL)(32) would suit American Memory users well -- if the metadata were more consistent. [The separate index for alternative titles in American Memory is for efficiency; it allows the search engine to generate hit lists based on titles without retrieving the full records.]

Searching by topic (assigned terms and almost any text)

The common text bucket proves invaluable when dealing with heterogeneous data and it is useful to include subject terms in this bucket. Indexing engines designed for full text will find word variants automatically, relieving users from knowing when formal subject terms use the plural form. The distinction between assigned subject terms and textual description is, nevertheless, valuable. In American Memory, it allows us to generate browse lists, which are actually static (but easily and automatically regenerated) HTML pages with "canned" searches. The subject index also permits navigation from subject terms on record displays to other records assigned the same terms.

Searching by originator

Although American Memory indexes the primary Creator separately from the other Contributors, searching and browsing for creators and contributors is usually done together. Combining the indexes has been considered. In American Memory, the addition of roles to Creators and Contributors proves valuable, particularly for non-text materials (e.g. to distinguish composer from lyricist or illustrator of sheet music). Authorized forms for names are extremely valuable in American Memory. However, names within text are also an important access point.

Searching by date range

The Alexandria Digital Library is designed for powerful searching by geographic location and date range. Strict machine-parsable content standards are used for those metadata elements. American Memory users would love to be able to search more effectively by date and place. For much content in American

Memory, unfortunately, dates of creation for the original item are uncertain and date-ranges recorded are often so broad as to be useless for discovery. For certain collections where chronology is important, normalized dates have been generated and can be used to sort search results. It is interesting that the Alexandria Digital Library has chosen to index date of creation/publication in the same bucket as date of coverage. In American Memory, there are many instances in which the dates are essentially equivalent (for example, the date a photograph was taken, a letter written, or the proceedings of the Congress recorded). For published books and maps, the distinction is often important, but users may be interested in either or both. The Alexandria Digital Library uses special overlap and containment queries for date ranges and location. Such queries would not be efficient with a full-text engine, but American Memory users would like the capability.

Searching by place

Geographical location is an area in which traditional descriptive practices aimed at human-readable displays do not transfer well to digital libraries (or even support finding and collocation in traditional library catalogs). In American Memory's metadata from heterogeneous sources, location may be expressed in many ways: as an informal place-name, as a subdivision in a topical subject heading, as a traditional subject heading (e,g., Brooklyn (New York, N.Y.)) or a hierarchical place name (e.g., country -- state -- county --city). Of these, by far the most useful for manipulating automatically with simple text-based tools has been the hierarchical place name. We look forward to being able to use gazetteers to convert place-names to bounding boxes (coordinates) of the type used by the Alexandria Digital Library.

Searching or limiting searches by type

Users often know that they are looking for an image or for a map and would like to exclude other types of information from the start. Searching by more specific genre terms is also useful, for example for posters or cartoons. As pointed out earlier, type categories expressed in terms that users recognize should be available for limiting or searching. Controlled vocabularies are useful, but are likely to be domain-specific. Convenient distributed access to vocabulary or authority registries is a part of Licklider's vision that has not yet been achieved.

Improved tools to support access to resources in digital libraries

Access will be enhanced through better tools for generating and transforming metadata, better tools for sharing and exchanging metadata, better tools for search and retrieval, and better tools for post-processing search results. The emergence of XML (Extended Markup Language) and its widespread support as a syntax for exchanging metadata and content, particularly for e-commerce transactions and services, is stimulating the development of better tools for transforming and sharing metadata. This, in turn is leading to support for XML from vendors of database software and text-indexing engines.

It appears clear that XML will provide the syntax for metadata exchange among digital libraries in the coming years. A few examples of its adoption to support interoperability include: the Bath Profile (as a

record syntax option for search and retrieval using Z39.50); MPEG-7 (for the MPEG-7 Description Definition Language); the Open Archives Initiative (to allow information service providers to harvest metadata from data providers); and the Federal Geographic Data Committee(33) (as the syntax for the Content Standard for Digital Geospatial Metadata). XML is also being used as a syntax to represent content objects, such as the proposed Open eBook standard. American Memory has relied heavily on common indexing of heterogeneous metadata, with different element sets and in two different digital formats (MARC communications format and an XML-like syntax for simple (Dublin Core-like) records. Migration to an XML syntax with a formal DTD or schema is anticipated.

Transformations from one XML metadata schema to another (especially from a rich one to a simpler one) can be facilitated using Extensible Stylesheet Language Transformations (XSL/T). This is just one example of the powerful XML-based tools emerging. XSL/T is already in common use for transforming finding aids marked up to the EAD standard (in XML) to HTML for display on the web. XML also includes the ability to mix and match metadata element definitions (semantics and syntax) from other schemas (using the "namespace" feature). As the flurry of XML-related activity on the World Wide Web Consortium web site in late 2000 shows, the general acceptance of XML signals the beginning of a further period of experimentation and development. Among the unanswered questions relating to the use of XML as a syntax for metadata is how soon (or whether) there will be widespread adoption of the Resource Description Framework, an elegant modeling framework for descriptive schemas.(34) RDF is layered on top of XML, using a particular XML-based syntax for metadata. RDF-specific tools will be needed to take full advantage of its potential for scalable interoperability. Whether the mix-and-match potential of XML namespaces will be widely exploited also remains to be seen.

In the past, text-indexing engines, relational database management systems, and SGML-based storage and retrieval systems all offered different functionality to support the finding, collocation, and discovery objectives for digital libraries. A text-indexing engine can handle heterogeneous metadata and full text in a single system; the capabilities for matching word variants (stemming) have been invaluable for American Memory. Features standard in relational database systems, however, such as sorting by date, have been implemented by additional programming. SGML-aware systems have advantages for substantial bodies of highly structured textual content, such as books and periodicals. Recently, products in one category have found ways to integrate the capabilities of another. ORACLE now has a full text search module, CONTEXT. Some text-indexing engines can now index text stored within relational databases. All such products are announcing "support" for XML.

The other area where tools are emerging is in automated integration of thesauri and other knowledge bases to support more intelligent retrieval. Such tools can compensate for less complete metadata. These knowledge bases could also be more widely used as a resource when creating metadata. Digital libraries will benefit from network-accessible thesauri and authority files that can be queried dynamically from systems that are used to generate metadata (whether automatically or by human catalogers).

Looking ahead

The vision that Licklider and his colleagues expressed in 1965 of libraries that allowed richer "interaction

with the fund of knowledge" is still a goal to strive for. I make no attempt to look ahead another thirty-five years. From the trenches, my view toward the horizon includes rich metadata schemas for content that warrants it, simpler schemas that encourage broader access to organized knowledge through interoperability, and ongoing popularity of simple-minded searches supported by intelligent tools in the background. As communities develop rich metadata schemas, I hope they take advantage of the existing fund of knowledge on organizing knowledge. Elaine Svenonius looks back on the history of cataloging and discusses principles, practices, and most refreshingly, problems with practices. Picking up on one of the problems she mentions, the shortsightedness of using one "device" to serve multiple purposes, I offer a few snippets of advice to those designing or applying metadata schemas. I present the advice in my own words, but am confident that many of my colleagues in the National Digital Library Program share the sentiment, because it reflects frustrations they have expressed.

In metadata schemas, draw clear distinctions between elements that serve different purposes. Some examples from the American Memory experience include:

Content type (genre, mode of expression)	Types used to identify a set of guidelines used for cataloging or description.
	Types used to trigger behavior in a particular system or application.
	Types as terms for users to use in queries.
Dates and periods	Dates or date ranges intended for automatic manipulation, such as sorting or access through a timeline slider.
	Dates, date ranges, or periods intended to be readable by humans.
Geographic locations	Coordinate-based locations, that can be used (a) in a map-based query interface that retrieves items ranked by distance from a query point or (b) to respond to queries that look for inclusion within or overlap with geospatial footprints.
	Hierarchically structured names that can be used for simple map-based querying and for conversion to geospatial footprints using gazetteers.
	Place names intended to be read by human users.

Finally, I would like to express my thanks to the organizers of the conference on *Bibliographic Control for the New Millennium* for asking me to contribute a discussion paper. Without this stimulus, I might

never have read *The Intellectual Foundation of Information Organization* by Elaine Svenonius from cover to cover. Her deep experience and shrewd analysis shed light on our struggles with heterogeneous metadata in building American Memory and provide articulate confirmation for some of the lessons we have learned from experience. This is what "interacting with the fund of knowledge" is all about.

References

Baker, Thomas. (2000). "A Grammar of Dublin Core." D-Lib Magazine, October 2000. [http://www.dlib.org/dlib/october00/baker/10baker.html]

Frew, James, Michael Freeston, Linda Hill, Greg Jane, Mary Larsgaard, Qi Zheng. (1999). "Generic Query Metadata for Geospatial Digital Libraries." In Proceedings of the Third IEEE Meta-data Conference, April 6-7, 1999 [http://computer.org/proceedings/meta/1999/papers/55/jfrew.htm]

Lagoze, Carl. (2000). *Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience* http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR2000-1801

Lessig, Lawrence. (1999). *Code and other Laws of Cyberspace.* New York: Basic Books.

Licklider, J.C.R. (1965). *Libraries of the Future.* Cambridge: MIT Press.

EU-NSF Working Group on Metadata. (1999?). *Metadata for Digital Libraries: a Research Agenda.* [http://wwwlis.iei.pi.cnr.it/DELOS/REPORTS/metadata.html]

International Organization for Standardization (ISO). (2000) *Overview of the MPEG-7 Standard* (version 3.0, May/June 2000) ISO/IEC JTC1/SC29/WG11 N3445 [http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm]

Svenonius, Elaine. (2000). *The Intellectual Foundation of Information Organization.* Cambridge: MIT Press.

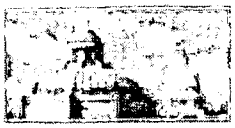
Notes:

1. Licklider (1965, e.g., p. 39)
2. Although the prediction was accurate in its timing, Licklider envisioned systems that drew much more extensively on the concepts of artificial intelligence being explored in the 1960s than has been the case. Today's information system components for search and retrieval, such as Internet search engines and tools for matching gene sequences and documents, rely heavily on brute force methods made possible by the development of ever faster processors and networks, and ever denser media for computer memory and data storage.
3. Some URLs to try: http://www.badgers.org.uk/; http://www.uclan.ac.uk/library/musrail.htm;

http://skipjack.net/le_shore/worcestr/birding/birding.html

4. In keeping with the usage adopted by the World Wide Web Consortium and the Dublin Core Metadata Initiative, I use metadata as a singular collective noun and the anglicized plural for schema.
5. Svenonius (2000, chapter 1, p. 2)
6. Lessig (1999, chapter 7, p. 87). One of Lessig's main points is that regulation of the architecture of cyberspace is as necessary to society as the regulation of physical space (through building codes, establishment of parks, environmental controls, etc.). The consitution and most existing laws, however, were framed in a world constrained by physical space and demonstrate "latent ambiguities" when applied to cyberspace.
7. EU-NSF Working Group on Metadata. (1999?)
8. Svenonius (2000, chapter 2, p. 17)
9. Svenonius (2000, chapter 2, p. 20)
10. Dublin Core Metadata Initiative. http://www.purl.org/dc/
11. Open eBook Forum. http://www.openebook.org/
12. JPEG2000. http://www.jpeg.org/JPEG2000.htm
13. MPEG-7. http://www.cselt.it/mpeg/standards/mpeg-7/mpeg-7.htm
14. IMS Meta-data Specification. http://www.imsproject.org/metadata/
15. Federal Geographic Data Committee. http://www.fgdc.gov/
16. VRA Core Categories, Version 3.0. http://www.gsd.harvard.edu/~staffaw3/vra/vracore3.htm
17. Art Museum Image Consortium (AMICO). http://www.amico.org/
18. Visual Information Access (VIA), Harvard University. http://hul.harvard.edu/ldi/html/via.html
19. Encoded Archival Description.http://lcweb.loc.gov/ead/
20. California Digital Heritage Image Finding Aids, Online Archive of California, California Digital Library. http://www.oac.cdlib.org/dynaweb/ead/calher
21. Rare Book, Manuscript, and Special Collections Library, Duke University. http://scriptorium.lib.duke.edu/
22. International Committee for Documentation of the International Council of Museums (ICOM-CIDOC). http://www.cidoc.icom.org/
23. The Perseus Project. http://www.perseus.tufts.edu/
24. Licklider (1965, p. 5)
25. Draft Standard Z39.85-200X, The Dublin Core Metadata Element Set. http://www.perseus.tufts.edu/
26. Lagoze (2000)
27. Dublin Core Qualifiers. http://purl.org/dc/documents/rec/dcmes-qualifiers-20000711.htm
28. Baker (2000)
29. Svenonius (2000, Chapter 7, endnote 12, p. 214)
30. Svenonius (2000, Chapter 6, p. 93)
31. Dublin Core Data Dictionaries, University of Washington Libraries. http://www.lib.washington.edu/msd/mig/datadicts/
32. Frew (1999)

33. Federal Geographic Data Committee. http://www.fgdc.gov/
34. Resource Description Frameowrk. http://www.w3c.org/RDF/



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

NAS study and 2 articles from the LC staff *Gazette*

Conference program

Speakers, commentators, and papers

Conference sponsors

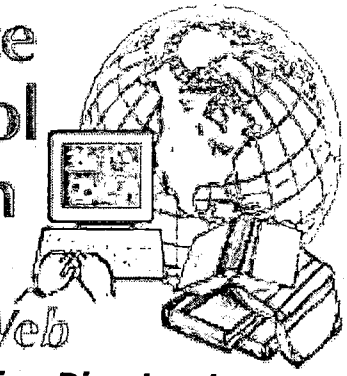
Conference discussion list

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

Caroline Arms

Information Technology Services
Library of Congress
Washington, D.C. 20540

Some Observations on Metadata and Digital Libraries

About the presenter:

Caroline Arms has been at the Library of Congress since 1995, as a technical program coordinator for the National Digital Library Program based in the Information Technology Services division of the Library. In particular, she has been the technical advisor for the *Library of Congress / Ameritech National Digital Library Competition*. Between 1997 and 1999 this competition made awards for twenty-three projects to digitize primary source materials to complement and enrich the Library's American Memory resource. By October 2000, twelve have been integrated into American Memory. Prior to joining the Library, Arms worked at the Falk Library of the Health Sciences at the University of Pittsburgh, as the first Director of Computing at the Amos Tuck School of Business Administration at Dartmouth College, and providing computing support to researchers at the University of Sussex and the Open University (in the United Kingdom). She has a B.A. in Mathematics from Oxford University and an M.B.A. from Dartmouth College. In the late 1980s, Arms edited two volumes for EDUCOM, *Campus Networking Strategies* and *Campus Strategies for Libraries and Electronic Information*, both published by Digital Press.

Full text of paper is available

Cataloging Directorate Home Page

Library of Congress Home Page

Summary:

The Internet has stimulated the development and deployment of collections of digital content managed and made available over the network for particular communities or purposes. These digital libraries, with their associated services, have varied ancestry. Some, like American Memory have been built by libraries or other archival institutions. Others have emerged from user communities to provide shared management and networked access for important digital resources, such as survey data for social scientists, sensor data from satellites or telescopes for astrophysicists and other scientists, or instructional resources for faculty and teachers.

The metadata elements needed to allow specialist users to find, identify, select, and obtain the resources they need and to navigate the web of relationships among them do not necessarily match the elements and rules for bibliographic cataloging of materials traditionally held by libraries. The potential for coordinated access to resources of different types from different sources, however, calls for a level of commonality among metadata schemes. Simple and rapid access to full content may reduce the need for some cataloging details, since the user may be able to use the full content or an automatically created summary, such as a thumbnail of an image or outline derived from marked-up text, to aid selection. On the other hand, although archival collections in paper form are often described as a whole or at the level of a series or physical container, item-level identification is essential in a digital library, increasing the cataloging cost. However, content in digital form can be a source for automatically generated metadata; such metadata will be less costly but flaws that would be easily corrected or avoided by a human cataloger may go undetected. In digital libraries, not all relationships between items have to be recorded in catalog records. Relationships between digital works can be embedded when the work is created or derived automatically by analysis of the full content. Citations can link to the works referenced, providing navigation capabilities far richer than those possible through catalog records.

This paper will draw on experience gathering together metadata from heterogeneous sources for American Memory, particularly for the collections digitized and cataloged at other institutions through the LC/Ameritech competition. It will also reflect on several initiatives to develop rich structured metadata schemes for specific domains and others to find simple approaches to support resource discovery across domains. Trends and commonalities will be identified and influences among metadata schemes highlighted.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").