

CHAPTER 7

CATALOGING AND METADATA CREATION IN DIGITAL INFORMATION ORGANIZATION: OLD CONCEPTS, NEW CHALLENGES

Beth Davis-Brown

INTRODUCTION

The promise of digital libraries implies the possibility of disseminating materials and information far beyond what has ever been imagined. Early digital library efforts, such as the Library of Congress' National Digital Library Program and the projects sponsored by the digital library I and II initiatives in the United States, showcase digital facsimiles of unique documents and artifacts previously available only to curators and scholars. In Japan, the National Diet Library, Kyoto University Library, the University of Tsukuba, and the University of Library and Information Science are actively planning to publish digital content on the World Wide Web (see site reports, Appendix C). One could view "digital information organization" as having two facets:

1. The creation of cataloging information to enable searching, discovery, and retrieval of information in digital format.
2. Accomplishing this task with methods that scale to effectively handle quantities of data exponentially larger than libraries have ever done. A key issue impacting the wide dissemination of digital information is the scalability of providing information (metadata) to structure and enable searching, navigation, and presentation of online documents. This paper will address some of the issues involved in creating cataloging and metadata, and discuss attributes of print documents and other analog formats that must be replicated in the electronic environment.

This author's participation in the WTEC study tour stemmed from experience in cataloging and classification and as manager of a team that digitizes historical legal materials for the Law Library of Congress and the National Digital Library Program. As the only "librarian" on the study tour, the author paid special attention to problems and issues concerning metadata creation and scalability of cataloging systems. These issues are just being articulated in both Japan and the United States, and call for thought and discussion. The goal of this chapter is to provide an introduction to factors that impact the growth of digital library technology and content from a practitioner's perspective.

On the surface, provision of metadata to accompany digital objects does not seem difficult. Roughly speaking, many people think that all that must be done is to take existing cataloging information, convert it to the appropriate format, and link it to the digital images. The process is not that simple due to several factors. First of all, the conversion of a physical artifact implies not just putting information into a new format but the concomitant goal to display the information in a logical way. To do that, information in addition to the content must be produced or extracted to enable the structure and display of the data. If existing schemes for classification and indexing are used, human intellectual capital is necessary at some point in the process to apply thesaurus terms and enable other access points (catalog).

TRADITIONAL LIBRARY PRACTICES

Many libraries in the United States and in Japan have not yet cataloged all of their holdings, and may not have all of these records in machine-readable format. A shortage of staff, sometimes rocky transitions from manual to automated cataloging work flows, and the “information explosion,” have created arrearages or backlogs in cataloging departments that most institutions do not publicize. The discipline of cataloging has devised methods and policies to describe physical artifacts such as books, periodicals, microforms, sound recordings, and maps. These descriptions are largely based on the physical “container” in which the information resides, and thus are considered format-based description. (For more on bibliographic description, see AACR2.) Intellectual description, that is, data about the subject of the information in the “container” and a classification number reflecting subject analysis, is created by catalogers. The Library of Congress develops and maintains a very large thesaurus of controlled subject headings with documentation of references and related terms (see LC Subject Headings). The Library of Congress also maintains the *LC Classification Schedule* and the *Dewey Decimal Classification Schedule*. Variants of these classification schemes are being used by libraries in Japan. The library community's cooperation in development of these systems has made possible interoperability and interchange of bibliographic information on a global basis. Cooperatively built databases of bibliographic information such as OCLC and RLIN in the United States, and NACSIS in Japan, provide economies of scale as libraries collectively create and share the world's bibliography.

A discussion of “granularity,” or the level at which an item is described, is a conceptual key for understanding digital information organization. “Item level” cataloging is probably most familiar to readers as they use online library catalogs to find monographs and multimedia materials. That is, one cataloging record is made for one work. Archives and special collections often catalog at the “collection level,” insofar as it is not feasible to individually describe every letter in a huge archive or assign meaningful classification numbers to millions of photographs. With indexed journal articles, the “item” to be cataloged might be the title of the journal along with an accounting of the individual issues, or holdings. Article-level indexing information gives further description of the intellectual content of “pieces” of each issue of the journal. Conversely, one catalog entry might exist only at the title level of the serial publication without the more in-depth indexing information. Clearly, the article-level indexing provides greater access and description; it is also more expensive and labor-intensive to create and maintain. The topic of granularity of description is important because the creation of cataloging data is one of the more expensive aspects of traditional library methods of providing access to materials.

The digital world requires this same cataloging data as well as information necessary to structure and present electronic documents. The library community is cooperating with professionals from the computer science, text encoding, and museum communities to develop the Dublin Core metadata standard, a fifteen-element set to describe digital resources (see Dublin Core Metadata). Generally, metadata as discussed in this context falls into three major categories.

Physical/Structural Metadata

Physical/structural metadata is information about the digital object and its relationship to other digital objects in a repository. Structural metadata might include file location on a server or in a repository; file format; file size; relationship to other files; sequence, or date of creation. For example, a sequence of 35 mm photographic negatives may have been imaged. To present the negatives in the order in which the photographer created them, information is needed to structure the images in the original sequence in addition to the format and size of the file, date of creation, and internal numbering scheme. To extend the analogy to books on library shelves would be data about the shelf number (physical location); number of pages (size); the fact that the pages are numbered (sequencing); and binding (defining the item). This indication is a logical way to structure these materials as well as a means to indicate provenance of the materials.

Intellectual Metadata

The term intellectual metadata refers to information that provides access to the subject or content of a digital object. Intellectual metadata can be thesaurus terms associated with a file or item; indexing achieved by full text search and retrieval as described in Dr. Croft's Chapter 5; classification according to standard schema;

and associations with related sources of intellectual data such as bibliographies, archival finding aids, or cataloging records. Again using the example of 35 mm negatives from a roll of film, intellectual metadata would consist of who created the images (photographer/author), controlled or uncontrolled vocabulary terms describing the images, and perhaps a classification number. The related data might be the photographer's captions or references to a work in which the images were published.

Rights and Permissions/Access Management Metadata

Rights and permission/access management metadata functionally describe the goal of the encoding of rights and permissions information at the computer file level for digital objects. For example, an archival collection of photographs may have been made available to researchers, one at a time, in a special collections reading room for many years. However, the literary trustee of that collection may nonetheless object to widespread dissemination of these images on the World Wide Web. Thus, rights and permissions data must be associated with each image to indicate its status for distribution.

As mentioned previously, the creation of metadata has traditionally been one of the most expensive aspects of making library materials available. At the Library of Congress, one of the key decision factors when selecting collections for digitization is whether or not cataloging information already exists for a collection, especially in terms of intellectual metadata. The costs of scanning and even having text keyed and proofed are minimal in comparison with paying subject experts and professional catalogers to describe materials according to standardized methodology. Strategies for minimal level cataloging and reducing cataloging access points abound, but the activity remains very expensive.

There are aspects of current library practices in the United States and Japan that impact the potential reality of a global digital library at many levels. The costs and complexity of creating appropriate data needed to present materials reformatted digitally have been discussed. Additionally, if this information is not created accurately and with future presentation needs in mind, the digital materials can be unusable. For example, if a book is mislabeled or misshelved in library stacks, it is still available by inspection by a deck attendant or users. Conversely, if a digital file is not linked correctly to its related bibliographic record, finding aid, or previous pages in a sequence, it is essentially irretrievable. Thus, data must be created and checked for quality at a high level to ensure usability (see LC RFP 96-18).

HANDLING AND PRESERVATION

One of the principal goals of digitally converting historical and rare materials is to preserve the knowledge contained in them long after the lifetime of the physical container. Ironically, the very act of scanning these materials can cause damage to the physical artifacts if care is not taken in handling and treatment. In the library community, there has always been talk of the tension between preservation and access. To preserve treasures, they must be safeguarded, kept away from light and stress, and used only under restricted circumstances. Digital libraries *seem* to provide a solution to this problem—the possibility of creating facsimile digital images and distributing them widely while sheltering the original artifact from prolonged abuse. Institutions such as Keio University's HUMI Project and the Tsukuba University Library exhibit admirable leadership to the library community by submitting their treasures to the scanning process. Tsukuba University Library and the National Diet Library have stated they plan to rescan materials repeatedly as greater storage space, high speed networks, and higher quality display technology allow for superior images. In the United States this idea has not been expressed due to preservation, cost, and labor considerations. To retain knowledge in materials published on paper and other unstable media, handling and preservation concerns are significant factors to face when considering the possibility of a global digital library.

DIGITAL BINDING

Dr. Rama Chellappa and this author coined the concept of “digital binding” during discussions and site visits in Japan. For a while, thought has been given as to how to expand the definition of metadata backwards, if you will, to physical artifacts, and aspects of their physicality that give us information about use. The fact

that a book is bound and that in Western languages should be read from left to right are implicit pieces of metadata. To present that same book in digital format, each file representing every page image and the beginning and end of the book must be coded in a way to allow for coherent display to the user. It is nice to allow the user to “turn” pages of a document, which requires encoded information indicating digital file sequence and document boundaries as they relate to the original artifact. As this illustration suggests, activities that are taken for granted with artifacts, such as knowing in which order the pages should be read and where the boundaries of the document lie, must be recreated and made explicit for digital presentation.

Other thoughts about the “digital binding” concept came in a session with President Makoto Nagao of Kyoto University. Professor Nagao developed Ariadne, a multimedia digital library system that was demonstrated publicly in October 1994 (Nagao n.d.). Nagao discussed the difference between traditional book publishing versus publishing on the Internet, stating that, “There are so many information creators besides professional publishers on the network, and some parts of information created by these creators are so important that the collection of these digital information content(s) is urgent for libraries (Nagao n.d.)” This led to the thought of other attributes of published materials that might be emulated in the “digital binding” arena to allow for the study of information that is naturally inherent in published materials and that indicates authorship, provides version control, and defines the document. For example, in libraries or bookstores, electronic or physical stamps on the artifact indicate ownership; the date of creation and printing is fixed and is usually expressed on the verso of the title page; the content is physically immutable because the item is bound or packaged together; and the status and reputation of the publisher provide verification and to some extent authentication of the information. One relies more on material printed by Oxford University Press than vanity publications from typewriters and photocopy machines. For “digital binding” to occur, would not these same authentication and verification features need to be replicated in an electronic environment? How then, might one recreate these aspects of “boundness” or “publishedness” in the electronic environment, assuming that the information is critically important? Both legal and technical aspects of this question are interesting to explore and potentially prototypical in this context.

Publishers function to collect fees from consumers of information, either through sales or database access fees. The publisher takes a risk and may be rewarded or penalized economically for the gamble of publishing an author’s work. In the online climate, President Nagao suggests that “a digital library cannot exist without a charging mechanism to users” in order to charge and collect licensing or usage fees for digital books. Dr. Chellappa and this author hope to explore some of these issues further by prototyping technical means through which to provide a “digital binding” scheme from the perspective of multimedia and text materials.

INTEGRATION OF DIGITAL INFORMATION

How will digital information be integrated with materials represented in traditional resources, such as library catalogs, citations, and abstracts? As mentioned earlier in this article, many libraries have not yet cataloged all of their holdings at the title level. With a new world of digital information on the Internet, libraries are struggling with policies to determine how they will integrate these materials with older formats. For example, how does a user of a public library in Kyoto or in Washington, D.C. know that the Library of Congress offers THOMAS, an Internet resource of contemporary bills and acts of the U.S. Congress? The Internet cognoscenti claim, “people that use the WWW know how to find what they need with search engines. Cataloging is over.” However, many “average” information seekers have only vague ideas of the resources on the WWW and how to go about finding them. It is for these users that the integration of descriptive information about digital materials into traditional means of information discovery is essential. This topic was addressed as early as 1994 at the Seminar on Cataloging Digital Documents held at University of Virginia and the Library of Congress (UVA and LC 1994). Follow-up meetings have continued to discuss these issues in the library community and have given rise to organizations such as the Digital Library Federation (LC 1995).

CONCLUSION

The practical implementation of processes to describe and access information can be complex. While the issues concerning description and access of information are not difficult to understand in a strictly intellectual

sense, coordinating the creation of cataloging and metadata in a production environment can be. Much of the data necessary to successfully retrieve and present digital information online is inherent in the conversion/creation process of that information. If files are not named properly or essential metadata are not captured at the time of image scanning or record creation, the material can be unusable without highly expensive manual intervention or reprogramming. The process of quality review, worthy of an article in itself, is an activity that is crucial at both ends of the digital content production cycle to verify that the digitally converted information is legible, clear, and properly labeled.

A final conclusion from what was seen in Japan and from what this author knows to be true in the United States is that digital content management on a large scale is a huge question impacting digital libraries in Japan and the United States. Traditional methods of description and access are not practical, affordable, or appropriate for large amounts of digital material. The library community has led admirably in terms of standardizing data formats and standards for description and access that make bibliographic records interoperable. In Rama Chellappa's Chapter 6 concerning image retrieval in Japan, and Bruce Croft's impressions regarding text (Chapter 5), there are some research agendas with the goal of achieving scalable solutions to content conversion and management problems. But until that time, digital information organization must continue to be studied, prototyped through projects such as the DLIB II initiative, and considered carefully by professionals in many disciplines.

REFERENCES

- Anglo-American cataloguing rules*, second edition (AACR2) Revisions 1983. Joint Steering Committee for Revision of AACR.
- American Library Association, 1984. Chicago.
- Digital Library Federation homepage. <http://lcweb.loc.gov/loc/ndlf/ndlfhome.html>.
- Dublin Core Metadata*. http://purl.org/metadata/dublin_core.
- Library of Congress. 1984-. *Subject cataloging manual. Subject headings*. Subject Cataloging Division, Processing Services.
- Library of Congress. 1995. Organizing the Global Digital Library Conference. Library of Congress Digital Library Visitors' Center. December 11, 1995. <gopher://marvel.loc.gov/00/loc/conf.meet/gdl>.
- Library of Congress. 1996. Organizing the Global Digital Library II (OGDL II) and Naming Conventions. Library of Congress, Washington, DC. May 21-22, 1996. <http://lcweb.loc.gov/catdir/ogdl2/>.
- Library of Congress and University of Virginia. 1994. *Proceedings of the Seminar on Cataloging Digital Documents*. University of Virginia Library, Charlottesville and the Library of Congress. October 12-14, 1994. <http://lcweb.loc.gov/catdir/semdigdocs/seminar.html>.
- National Digital Library Program. *The Library of Congress Requests Proposals for Digital Images from Original Documents, Text Conversion and SGML-Encoding*. <http://memory.loc.gov/ammem/prpsal/coverpag.html>.
- Nagao, Makoto. Copyright in the age of digital libraries.
- Nagao, Makoto. Multimedia digital library: ARIADNE.

