[Transparency 1]

A WORLD OF DISAPPEARING BOUNDARIES: TRADITIONAL ORGANIZATION OF INFORMATION IN AN ELECTRONIC ENVIRONMENT

by
Arlene G. Taylor
© 1997

[Transparency 2]

Information technology is in the process of eliminating boundaries of all kinds. Among the many changing boundaries libraries face are what is to be included in a catalog, what is to be included in a catalog record, international coding of records for machine manipulation, new concepts in authority control as international needs are taken into consideration, challenges of multiple-language subject approaches (as well as the need to use multiple thesauri), and the possibility of the use of classification schemes as switching "languages." I will address each of these issues in turn.

What is Included in Catalogs

[Transparency 3]

Catalogs have traditionally been defined as indexes that tell what is owned in a particular local library. That is, a user in the past could enter a library, look in the catalog under author, title, or subject, and learn whether the library owned a particular item or whether the library had items on a particular subject. Whatever the user found in the catalog could potentially be retrieved in that location, if it were not lost or already charged out to another patron.

The concept of "union catalogs" stretched this definition a bit. Items found in a union catalog were owned by one or more of several cooperating libraries. Union catalogs could be made for

© 1997 Arlene G. Taylor

libraries in a geographic area, such as a state, or for branches of a library system, such as a public library system for a city, or a university library system with subject specific branches. Such catalogs were rare, however, in the later days of card catalogs. The filing in large catalogs was too great a burden. As technology made it possible to collect records so that book catalogs could be produced easily, union catalogs became more common. And, today with online catalogs, virtually all university libraries and public libraries with branches have union catalogs. There are also now many consortia consisting of independent library systems in a given area that have union catalogs for their combined holdings. However, union catalogs still have not, until recently, contained records for anything not physically owned by at least one of the libraries involved.

Now, though, the boundaries of the catalog as mechanism to show what is physically owned are rapidly disappearing. In preparation for this presentation I posted some questions to AUTOCAT, an electronic discussion group for catalogers, and asked for voluntary responses.

The first question had to do with whether the catalog includes records for things not physically owned by the library. A few respondents commented on the inclusion of records for items owned somewhere at the institution but not under the control of the library. All the archival respondents indicated they have records for items not owned (e.g., all the historical records repositories in a state). Several academic library respondents indicated that they are cataloging Internet documents. In these cases the documents to be cataloged have been chosen according to collection development policies of the libraries. That is, if a document is something that a bibliographer would select for inclusion in the collection if it were a physical document, it is believed that it should be cataloged so that patrons can gain access to it even though it is not "in" the library.

© 1997 Arlene G. Taylor

Some libraries subscribe to electronic journals, which they then catalog – the idea being that they have purchased the right to make these serials available to their patrons and so must catalog them so that patrons will know about their existence. Some libraries have begun attaching information about related things that are not owned to catalog records for things that are owned (e.g., the URL for the home page for a museum is added to the catalog record for a guidebook for that museum). A few people indicated that they only catalog web sites that are originally mounted and maintained by the library or by the institution in which the library falls. Related to this was the answer indicating the need to catalog the electronic journals that are "pointed to" on the library's home pages.

[Transparency 4]

The loss of boundaries concerning what is included in the catalog is not surprising considering that the boundaries of catalog use itself have been disappearing for quite some time. As soon as telephone and modem access became available, libraries began providing off-site access to online catalogs. It was no longer necessary to go physically to the library to look in the catalog. Soon it became possible to "telnet" to libraries miles away, first nationally, and then internationally. A later innovation, the Z39.50 protocol, makes it possible to search many catalogs without first having to learn the search commands of each different catalog system.

Finally, Web interfaces are removing the boundaries that have existed between the set of bibliographic records that describe information resources and the resources themselves. That is, one can now enter a catalog through a web interface, find a record containing a URL, click on the URL, and immediately be taken to the information at that site.

© 1997 Arlene G. Taylor

What is Included in Catalog Records

[Transparency 5]

Another change being dealt with is what is acceptable metadata to be included in a catalog record. (Metadata is the word being used to mean a short encoded description of a document – it gives a description of and other data about a particular information resource.) As the number of records pulled from a variety of resource files increases, new decisions must be made about the information in the catalog record itself. For many years in the United States the standard in most library catalogs was to use records created at the Library of Congress (LC). Catalogers often made local changes to those records, and for everything that did not have "LC Copy" available they created records in-house, adhering to standards set by the Library of Congress. Technology began to make possible the creation of catalog cards through photographic processes. United States libraries could contribute original cataloging to the *National Union Catalog* (NUC), and they could copy records from the NUC and photocopy them onto card stock.

Later, technology made possible the sharing of records through online systems. The Library of Congress began accepting "shared cataloging" from other countries in the 1960s. That policy entailed accepting the description provided by the other countries' catalogers, but putting all the access points into the form necessary to be under LC's authority control. As time has gone on, we have been accepting more resource files as the basis for our cataloging. OCLC and RLIN include records contributed by the National Library of Canada, The British Library, and members from all over the world. Libraries are accepting more and more of these records without change to the descriptive parts of the records. Some are also accepting access points without change. The latter policy, though,

© 1997 Arlene G. Taylor

abdicates responsibility to provide users with catalogs that make sense, because our boundaries in authority control have not yet been eliminated. I'll come back to this point shortly.

After the publication of the *Anglo-American Cataloging Rules, Second Edition* (AACR2), that set of rules began to become an international standard, as more and more countries adopted it. Elimination of boundaries concerning description in catalog records has led both to enhancing records with more information than called for by AACR2 and to the creation of records with less information than is called for by AACR2. Technology has been the key to allowing both of these outcomes. Technology allows the scanning in of tables of contents, for example. Abstracts may be pulled in from other online sources. On the other hand, when a library is faced with a backlog and limited resources, it is possible to create records with less information than required by AACR2 without fear of causing undue hardship for other libraries. A standard for Core-level Records has been set by the Program for Cooperative Cataloging (PCC) coordinated by LC. It has specific minimum requirements, and all access points must be under authority control; technology makes it fairly easy to enhance records that were originally put into the system at a Core Record level. In October 1997 there was an International Conference on AACR that has resulted in changes that will allow even greater use of it as a standard for description of all kinds of information.

In March 1995 a group met at OCLC to address the problem of providing metadata for network-accessible materials. The result was the definition of a data element set that could be used by information providers to describe their own resources. This set has come to be known as the Dublin Core (having been first crafted in Dublin, Ohio, location of OCLC's headquarters). Much more work has been done to define the elements in order to make them useful across communities. If this in fact becomes a standard, it would encourage authors to provide metadata simultaneously with their work. A

© 1997 Arlene G. Taylor

Dublin Core record might become the catalog record for the work, or it might be enhanced by a cataloger before being entered into a catalog.

How Catalog Records are Coded

[Transparency 6]

The third area of elimination of boundaries that I will address is the international coding of records for machine manipulation. Libraries in the U.S. are not yet having to deal with this issue. Virtually all U.S. libraries' records are coded with USMARC. However, national libraries are dealing with this issue. One current project is the ongoing effort by representatives of the United States, the United Kingdom, and Canada to reconcile the differences among USMARC, UKMARC, and CAN/MARC. Once this has been implemented, libraries in each country will need to deal with the recoding and relearning that will be necessary.

We must also consider UNIMARC. This is the international code used to switch from one country's national MARC to another's. For example, in order for U.S. records to be sent to some other countries, the coding of the records is changed from USMARC to UNIMARC. Then the receiving national library translates the UNIMARC coding into the version for that country. Some countries that have come to the coding of data later than others have not created their own national version of MARC but instead have started with UNIMARC in the first place. They are spared the translation phase on both ends of the process. I believe this boundary, too, will eventually vanish.

Another kind of coding of data that librarians are investigating is the coding of original information sources with Standard Generalized Mark-up Language (SGML) and/or Hyper-Text Mark-up Language (HTML). This coding has become an international standard in a very short time. Archives

© 1997 Arlene G. Taylor

already are using SGML/HTML for coding finding aids. It is possible that such coding will become the means for removing boundaries that now exist in the coding of metadata (i.e., bibliographic [catalog] records). LC has created a Document Type Definition (DTD) for translating MARC records to SGML.

Authority Control

[Transparency 7]

The fourth area where boundaries are disappearing is that of authority control. Here is another area with which academic libraries in the United States are not yet really dealing. Virtually all libraries in the U.S. use the Library of Congress authority file for forms of names in access points in their records. They also, for the most part, have no catalog records with notes or subject headings in languages other than English.

As with international coding for machine readability, the national libraries are the ones dealing with the issue of disappearing boundaries in authority control. A major conference on Research Issues for Authority Control was held at OCLC in April of 1996. At this conference Barbara Tillett, from the Library of Congress, talked about how using a single form of name conflicts with providing access internationally. It would be possible to agree on an entity (e.g., person, corporate body, family, jurisdiction, title of a work, etc.), if we did not have to agree on one form of name. If we could have an authority record that lists the different language names having to do with a particular entity, we could then select a preferred form of name to display in any particular catalog; but we would not have to say the English form is "correct" or, no, the Portuguese form is "correct" or whatever. This new concept is being called "access control" rather than "authority control." ***[Transparency 8]*** Here, for example, is

© 1997 Arlene G. Taylor

an edited form of the LC Authority record for Confucius. (I had to edit out some name forms in order to get this all on one slide.) It is clear here that this person is known internationally by many names. Why should the "authorized" form be the English one? If we had access control instead of authority control, a particular library could choose for the default display whichever of these forms is most appropriate for its users. However, a user's search for any one of these forms would retrieve records related to the person that English-speaking people know as Confucius. ***[Transparency 9]*** Another example is Aristotle. (Again, I had to edit out a few name forms to get this on one slide.) On this example, line 20 identifies the form in line 13 as being a Japanese form. This, of course, is a transliterated version. There is no reason why an access control record should not include the form in Japanese script.

[back to Transparency 7]

A first step toward international shared resource records for access control is being called the Anglo-American Authority File (AAAF). This is a file maintained at the British Library that is made up of the United States Name Authority File (USNAF) combined with the British Library's Name Authority List (BLNAL).

In order for international access control to be successful, there must be standard sets of data elements. The International Federation of Library Associations (IFLA) has a Working Group on Transnational Exchange of Authority Data, chaired by Barbara Tillett, that is working on this. One part of their work is toward a possible ISADN (International Standard Authority Data Number), although further development of such a number has been postponed for now in favor of working out ways of sharing national authority files internationally. Either system would allow identification of entity records internationally, and these could be managed locally by specifying a preferred form for display. The

© 1997 Arlene G. Taylor

IFLA Working Group is also working on guidelines for authority and reference entities and on a UNIMARC for Authorities.

Multiple Language Subject Control

[Transparency 10]

Access control for multi-lingual *subject* vocabularies will be a particularly difficult task to undertake. Most American libraries are currently ignoring the problem, partly because the major language for the Internet has become English, and English is learned as a second language in school in many parts of the world. English speakers easily become complacent and egotistical and begin to assume that anything worth knowing will be in English. This, of course, is not true; but even if it were true, there would still be a need to get non-English speakers to the information they need by identifying concepts in various languages. This is what has been recognized by some United States public libraries that have begun to provide Spanish language subject headings for at least the Spanish language materials they have acquired for their libraries.

Few academic libraries have an online catalog with an interface in a language other than English. A few have Spanish interfaces, and a very few have French interfaces. These libraries have not dealt with logical display, however. Spanish and French headings are displayed alongside Library of Congress headings in English, and cognates are listed twice -- once for the English and once for the Spanish or French. Some have *print-based* assistance for users who speak other languages: e.g., Spanish, Vietnamese, and Chinese.

Other countries have more experience with multiple language subject control than does the United States. Canada, for example, has systems with French and English subject heading control and

© 1997 Arlene G. Taylor

with French and English interfaces. The United States and the United Kingdom have had to deal with word differences and spelling differences in their attempt to combine headings in British English with those in American English.

Multiple Thesaurus Subject Control

[Transparency 11]

Access control for multi-thesaural vocabularies has had somewhat more attention in United States libraries. For many years the Library of Congress and the National Library of Medicine cooperated in using both LC Subject Headings (LCSH) and Medical Subject Headings (MeSH). At first libraries that acquired these records all had card catalogs; so they chose to make subject cards for only one set of headings or the other. With the advent of online catalogs, however, it became easy to display both sets of headings. Displaying them is one thing; presenting them logically and clearly to users is another. In some cases the displays are not clearly delineated. For example, in some libraries the subject list used is given in abbreviated form after each heading – i.e., (LC) following LCSH headings, (MeSH) following those headings, and (LCC) following Library of Congress Children's headings. The abbreviations are meaningless to most users. In many others headings from other lists are displayed in an identical fashion to LCSH and the headings are indistinguishable. One problem is that capitalization affects displays, and while many LCSH and MeSH subjects are the same, the capitalization is not; thus users may have to browse through two lists of the same subject. I believe that having two lists may prevent some users from realizing all the different titles we own if they only look under one subject heading when there is another subject heading which means the same thing, but is phrased differently. ***[Transparency 12]*** These points are illustrated by these examples: MeSH: Nursing--dictionaries

© 1997 Arlene G. Taylor

LCSH: Nursing--Dictionaries ; MeSH: Education, Nursing LCSH: Nursing--Study and teaching ; MeSH: Schools, Nursing LCSH: Nursing schools

A few projects have attempted to pull together terminology used in several different thesauri for a subject area. The most far-reaching and successful of these is the Unified Medical Language System (UMLS). Begun in 1986 by the National Library of Medicine (NLM), the UMLS project aims to aid the development of systems to help health professionals and researchers in the retrieval of biomedical information from a variety of sources. Relevant information is scattered across many databases that use very different controlled vocabularies and/or classifications. The UMLS project brings together in a uniform format about fifty different vocabularies and links different terminology used for the same concepts. It does not attempt to make judgments about the "best" terminology to use – the usages in the source vocabularies are preserved; but it does establish new relationships among terms from different vocabularies.

Classification

[Transparency 13]

The final area of disappearing boundaries that I will discuss has to do with the use of classification schemes. At the OCLC Authority Control Conference that I mentioned earlier, Joan Mitchell spoke about the emerging role of the Dewey Decimal Classification (DDC) as a multi-lingual, multi-thesaural authority control system. The abridged edition of DDC has been linked with the *Sears List of Subject Headings* in the U.S. for many years. In most editions of *Sears*, the preferred terms for concepts have been accompanied by one or more abridged DDC numbers. The full edition of DDC is now linked to LCSH headings in the electronic forms of the DDC. When a cataloger is considering a

© 1997 Arlene G. Taylor

particular DDC number, he or she may view sample LC subject headings that have been used with that number in catalog records found in OCLC. It is also possible to search for numbers by searching LC subject terms in the online system index.

DDC has the potential to be a multi-lingual authority control system because translations into a number of languages either already exist or are being worked on: e.g., Arabic, French, Greek, Hebrew, Italian, Persian, Russian, Spanish. In at least one public library system in the United States the Spanish edition of DDC has been placed in branches where many patrons are Spanish-speaking. The index to the Spanish DDC serves as entry vocabulary to the system so that a number found that is appropriate to a user's sought subject can be searched to find subject headings or works that will fill the user's information need.

Unfortunately for most *academic* libraries in the United States, DDC as a multi-lingual, multi-thesaural authority control system will not be very helpful. Only about 25 percent of academic libraries in the U.S. use DDC. Some of the remaining 75 percent have older materials still classified with DDC, but most have completely reclassified to the LC Classification Scheme (LCC). LCC has not yet moved in a multi-lingual, multi-thesaural direction. About 90% of *public* libraries in the U.S., however, use DDC and will be able to use its multi-lingual advantages.

Another classification system that is multi-lingual is the Universal Decimal Classification (UDC). There are full editions in English, French, German, Japanese, Russian, and 8 other languages. Abridged editions exist in 17 languages and 5 alphabets. Proponents of UDC claim that it is the most universal bibliographic scheme in existence. (This may be true even though it still shows the Western bias of its Dewey Decimal Classification roots.) It is widely used for classification of documents on the Internet

© 1997 Arlene G. Taylor

(as is DDC). As for the impact of the UDC in the United States, however, it is used only by a few specialist subject bibliographies. I know of no libraries in the U.S. that use it.

Conclusion

[Transparency 14]

In conclusion it seems clear that disappearing boundaries will continue to change the way we view, create, and interact with library catalogs. Changes in the different areas of catalogs and cataloging, though, are coming at different paces. Catalogs are rapidly changing to contain records for things not owned but to which a library wishes to give its patrons access. Catalog *records* will continue to vary greatly as to the amount of information contained in each. Ideally there needs to be an internationally agreed-upon standard for a minimum record. The PCC is now emphasizing its Core Record standard, which is based upon AACR2.

The coding of records will change more slowly. Again, ideally, we will have an internationally agreed-upon coding scheme. It might be a version of MARC, or we might move to something like SGML, or we might just need uniform translation of coding into a standard display where the coding is not apparent to users; but we do need to be able to exchange records and use each others records easily.

Authority control for names and subjects must begin to take various languages and various traditions into account. We must find ways to display names and subjects that do not imply that one country's preferred form is better than another's. We have had enough experience to know that for subjects we cannot get specialists to agree on one set of terms for all the myriad concepts in the world – not even in English, let alone in multiple languages! But we must work with vendors to create systems

© 1997 Arlene G. Taylor

that will display terms from various thesauri in ways that will make sense. Finally, the potential to use classification systems in place of verbal systems or as switching devices among verbal systems should be explored.

Disappearing boundaries are and always have been a challenge. Too often we've dealt with them in the past by building fences to delineate the perceived positions of the boundaries. In libraries and cataloging it is time to take down or at least to leap over the fences.

© 1997 Arlene G. Taylor