Designing data for use: from alphabetic order to linked data

This article describes the evolution of library metadata from the card to the internet. A functional analysis shows that library metadata has changed as the technology that carries and uses that metadata has changed. Library data also carries with it legacy practices, like the emphasis on headings designed for alphabetical discovery which has however been replaced by keyword access. Linked Data is an emerging technology that will allow data to directly use the World Wide Web as a storage and access platform.To prepare library data for that platform new concepts, described here, must be integrated into library data practices, and some legacy practices must be re-evaluated.

KAREN COYLE
Consultant

Data that you produce, whether catalog records or acquisitions system information, has to be appropriate to the functions that will be applied to it and that it is expected to support. If you want to present your data in alphabetical order then you have to have a string that represents a left-to-right alphanumeric ordering of the data in question. You may have to deal with initial articles if you do not want your data sorting under the initial 'The' or 'A'. If you want to present your users with facets that will allow them to choose slices out of your data store, then you will need bits of information that are unitary: simple, defined information units that can be applied to all records for which they are appropriate.

This article looks at some of the developments that library metadata has undergone, and how these developments are related to changes in technology.

Data designed for alphabetical order

Library data was originally designed for physical catalogs, either in book or card format, that relied on alphabetic order for discovery. In this environment, entire heading strings are devised that have meaning when encountered in a sorted order:

Times (London)
Times (New York)
Times (Zurich)
Hamlet. French. 1932
Hamlet. French. 1948
Hamlet. German.

These headings have at least three different functions; they

- provide a human-readable short description of where the user is in the overall alphabetical order
- provide access points the user can follow to retrieve the rest of the information about the record the heading is part of
- identify some 'thing' in the bibliographic universe. The 'thing' can be a title, a person, a subject, a corporate body, etc.

Functionality in a catalog of this nature is limited to services you can provide around ordered strings. Users must know the beginning of the string in order to begin browsing through the available headings, so it is not good for people who come to the catalog without a clear idea of what they are seeking. The alphabetical catalog is most effective in an environment where users are likely to know what they are looking for (e.g. scholars) and where personalized help is available when a user's self-guided approach to the catalog fails.

Data designed for keyword access

The confluence of library data in a machine-readable format and increased power of computer

systems led to a generation in which keyword searching was applied to library data. Keyword searching is primarily used for textual information and can be used to retrieve any text, from telephone books to poetry. It has a rather indiscriminate virtue in that it does not matter if the data is good or correct, or even what language it is in, as long as the string used in the query matches a string in the data.

Users take to keyword searching because it requires very little effort or prior knowledge. It has the unfortunate characteristic of masking its failures, since users do not see what they have not retrieved. Unlike a heading browse, the keyword search produces a set of retrievals and usually does not make suggestions of related or nearby items that the user may have missed.

There is not much you can do to facilitate keyword searching, other than provide text on which it can operate. It is a hammer-like tool, not a fine instrument. The use of keyword searching against library data undoubtedly provides some improvements in retrieval, but it totally ignores the purposes for which library data was designed, and thus makes little use of some of the best efforts of library catalogers.

Because keyword searching makes it possible to search on any text, it is the search method of choice for full-text resources. Regardless of the quality of the metadata for any given full text, keyword searching is useful to reveal minor topics within the text. In particular, it can be used to discover proper nouns (place names, person names, names of companies or events) in the text that would not be brought out in the catalog entry. Although keyword searching has many faults, it will probably continue to be used on texts.

There is, of course, nothing as simple as keyword searching for non-textual resources, although image and sound pattern recognition continue to be the object of research. Until such searching is readily available for those resources, retrieval will continue to rely on provided metadata, even though that metadata itself may be the object of keyword searching.

Designing data for faceting

Another way to facilitate discovery in metadata stores, especially in large databases, is to provide facets that can winnow down general results into more relevant and more manageable sets. Facets are not generally initial search functions but are refinements that can be applied to a retrieved result set. To be most effective, facets need to be concise values that appear once per record. Ideally, the facet should divide up the retrieved set into a small number of smaller sets. Think of a pie chart: a pie chart with thousands of different slices is not very useful; one with a handful of slices that cover the majority of the items retrieved gives a user some obvious choices. Facets suggest aspects of the retrieved set that users may not otherwise be aware of, informing them that the set can be reduced by a choice of language, of author, of topic, or publication date range.

Although some library catalogs have made effective use of facets, it is not easy to find data in library records that lends itself well to faceting. Because data for faceting has to be precise (the same information must always be expressed in the same way), the best source of facets is in the fixed fields where the data is taken from controlled lists. Unfortunately, some of that fixed field data does not always yield useful facets.

For example, while year of publication is a coded value, there are literally hundreds of dates in any library catalog, and reducing a large retrieved set by an individual year is probably not meaningful. On the other hand, language of text can be useful, even though there are potentially hundreds of different languages. This is because in many libraries a few languages will predominate, thus making a pie with a few large slices and perhaps many smaller ones. Language of text, however, may be a value that could be applied to an entire session or to a user identifier, since many users are only interested in a single language. Making this a facet choice that has to be selected for each search may not be the best solution.

Titles are rarely useful as facets, since most titles are unique strings, and authors can be useful as facets in certain circumstances. For example, if you have done a keyword search on 'twain huck finn', an author facet could divide the set between those that are by Mark Twain and those that are about Twain's *Huck Finn* and by other authors. That's a kind of roundabout way to distinguish between by and about.

What functions as facets in the library catalog is often an accident of the catalog data rather than a conscious choice on the part of the library community. Most catalogs can facet a retrieved set on

resource format (book, music recording, serial, movie) because that is coded in the Leader of the MARC record. Not all can facet on genre (mystery, romance, science fiction) because those are often not discrete values in the record.

Faceting by subject area would seem to be of obvious utility. However, any given LC subject heading is unlikely to isolate a wide swath of records from a retrieved set. The Library of Congress Subject Headings themselves consist of facets, but it takes some work, as demonstrated by OCLC's FAST project, to turn these into usable units of meaning. Clever programs can create subject facets from classification numbers or shelf locations, using upper-level subject areas like the two-letter LC Classification subject areas, or the top ten categories of the Dewey Decimal system.

All of these are interesting ways to extract facets from our current data, but a better way to produce faceted catalogs and discovery systems would be to study user needs, decide what facets would serve the users best to fulfill those needs, and then create data that facilitates the use of those facets in our systems. While obviously easier said than done, this is the normal progression from requirements analysis to data modeling that is advised for systems design. If we think faceting helps our users in their information research, we should find it worthwhile to make the effort to provide a good faceted system.

Designing data for linking

Each of the design issues above requires either some additions to the library record or a method of managing library catalog record data in a database. Designing data for linking goes beyond additions to the catalog record: it requires that we adopt a significantly different metadata methodology. This methodology is based on technologies that permit sharing data over the web and making connections between disparate data stores based on data elements that they have in common. 'Linked Data' must be designed to allow the greatest degree of data re-use, and that means data re-use by members of the heterogeneous web community.

There are three primary standards for Linked Data: the first is that Linked Data uses a simple structure of statements for all expressions of data and metadata; the second is that, wherever possible, Linked Data uses controlled lists for its content; and the third is that every 'thing' in the data should be named with an identifier rather than a language term.

For libraries, moving to Linked Data would be a multi-step process. The first step requires us to analyze our catalog records to discover the data that they contain. The catalog record today is primarily text, and even though much of that text is highly controlled, it is not easily usable as data. The difference between text and data is that one can compute on data; one can perform counts or comparisons, find equivalence or difference using algorithms.

What is data and what is text?

This is text:

> 23 cm.

It looks very data-like and precise, and for the human reader (who understands its meaning in the context of the catalog record) it is meaningful. To a computer program, however, it is not meaningful. To explain this to a computer, we would need to say, in a machine-coded format:

– the size of the item
– the height
– measured in centimeters
– value=**23**

The fact that some information in our machine-readable catalog records is entered twice, once in a text field (such as the date subfield of the publication statement) and again in a fixed field, is proof that the text-versus-data concept is part of our data model, but text continues to be the dominant information carrier in our records.

Another aspect of data is being very clear on what the particular datum is describing. In a catalog record for a book, we might assume that all of the information in the record is about the book, but that assumption would not be true. In this record, some of the data modifies related data elements, like the author's birth and death dates, and the additional information about the edition.

100 1_ $a Alcott, Louisa May, $d 1832-1888.
245 10 $a Little Women.
250 __ $a Modern abridged ed. $b Illustrated by David K Stone.
260 __ $a Racine, Wis., $b Whitman Pub. Co. $c 1965
650 _0 $a Sisters $z New England $v Drama.

If we were to break this record apart into its separate statements about the resource we would have something like:

this book: has author: Alcott, Louisa May
this author: has dates: 1832-1888
this book: has title: Little Women.
this book: has edition: Modern abridged ed.
this edition: has additional information: Illustrated by David K Stone
this book: has place of publication: Racine, Wis.
this book: has publisher: Whitman Pub. Co.
this book: has publication date: 1965
this book: has subject: (topic) Sisters (place) New England (genre) Drama

Such an analysis already moves us in the direction of Linked Data because it clearly connects the information in the record with what it is describing.

Once we have an analysis of the data elements, we need to take another step away from text: we need to create or use identifiers for every element that can be clearly and uniquely identified. Although we may not have thought of it in this way, the use of authoritative name, title and subject entries is an act of identification. The defect in our method, however, is that our display form and our identifier use the same alphanumeric string. Therefore, when we decide to change a display from 'Cookery' to 'Cooking' we change the identifier as well as the display. That violates the primary rule of identification, which is that the relationship of the identifier and the thing identified must be unique and constant. The way to mitigate this is to use immutable identifiers to identify the name or term, and allow the display to change. Thus if our author is first entered into the authority file as

display: Alcott, Louisa May
identifier: lcna: n 79117152
and is later amended to
display: Alcott, Louisa May, 1832-1888
identifier: lcna: n 79117152

The constancy of the identifier means that even as a display may change, we know that we have the same name. In addition, we have defined the name as one that has been controlled by a particular community and in a specific manner: in this case, the Library of Congress Name Authority file.

Use of identifiers means that data can be language-neutral. The same identity can have display forms in any or all languages and scripts. If we wish to share our data on the global web, language neutrality is essential.

The transition from today's data records to the statements of Linked Data is probably the most difficult one for us to conceptualize. In the record example I gave above, the components of a bibliographic record were broken into a series of three-part statements such as:

this book : has title : Little Women

This structure is called a *triple* because it has three parts:

subject : predicate : object

The subject is the thing that is the focus of your statement. The predicate states what attribute you are describing, and the object is what you are saying. It mimics a simple sentence, as you can see. A bibliographic description is realized as a set of statements that have the same subject. These statements can also be part of a web of data, illustrated below as a graph with nodes and arcs (See Figure 1). The nodes can be both subjects of

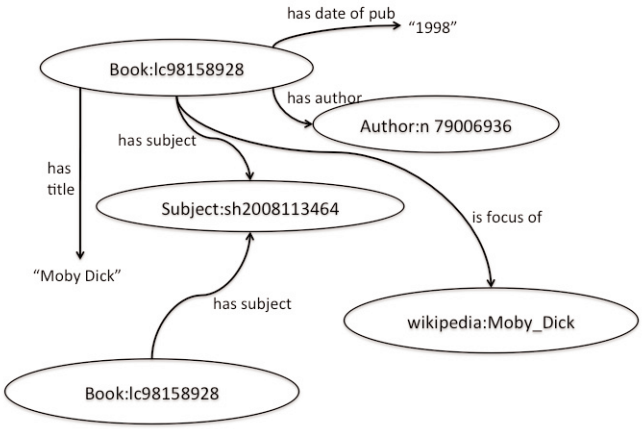

Figure 1. Bibliographic data in a graph format

one or more statements as well as objects of others, and it is this that allows a web of data to form from Linked Data contributed by different communities.

Note that connections can be made between things that are represented by identifiers but that objects that are textual in nature are always dead ends because they do not have enough identity to be in the subject position.

The possibility of Linked Data from libraries on the web is of interest outside of the library community. The World Wide Web Consortium, for example, has commissioned a group, known as the W3C Library Linked Data Incubator Group, to help increase global interoperability of library data on the web. For research communities active in the Linked Data space online, bibliographic data is an important part of their knowledge creation activities. Yet before undertaking an effort to transform library data to the Linked Data format, it is also essential that it be justified in terms of increased service to library users.

The library catalog today serves at least two purposes: it is an inventory database that describes what resources the library holds; and it is the primary means by which users can discover library resources. What it is not, however, is an information discovery system. This may seem counterintuitive, but not to library users who do their information seeking on the web and turn to the library catalog only when they want to know what the library has. The user comes to the library catalog with something in mind; an author, a title, a specific subject. This is exactly what the catalog has been designed to do, and this design has been relatively consistent since the time of Anthony Panizzi and his 91 rules for the British Museum catalog entries.

At one time, the library catalog was the only technology that provided searching on information sources. Once the scholar had found a book or journal, the library catalog's job was done. Any connections that were to be made between one resource and another were left to the individual or to the scholarly community.

Our catalogs still follow this model: the result of any search is one or more described resources. A search on an author's name gives essentially the same results as a search on a subject: a set of catalog records. Within any retrieved set there may be a great variety of items: books by the person, books with the person as subject, books with the person as an added entry. The retrieved items may span centuries of printing and reprinting. Users facing a large retrieved set surely must be overwhelmed. Because the data is presented to them as catalog records, it is nearly impossible for them to develop an understanding of the contents of the records retrieved: what years are covered? What languages? Is this everything this author wrote?

The Linked Data format would make it easier for us to mine that retrieved set and present the results to the user in a way that summarized the set for the user. Not only that, we could make correlations between data in the records: the subjects most related to this person; the places most often related to this subject; persons most frequently listed together as subjects; publication times. Linking to data from other communities could facilitate interesting enhancements like locating library resources on maps or linking to research projects that cite library resources. Linked Data could also create paths from web resources to libraries and vice versa.

Conclusion

There is not a single 'right way' to create data; each desired function we want to perform has its requirements. Throughout modern library history, libraries have modified their bibliographic data to take advantage of new technologies and to provide better services to users. There are new technologies available to us today and they may provide us with greater visibility for libraries on the web and the ability to take the library into the information space preferred by information seekers.

Suggested readings and resources

Berners-Lee, T, and Lassila, O, 'The Semantic Web: Scientific American', *Scientific American* (2001).

Coyle, K, 'Understanding Metadata and Its Purposes', *The Journal of Academic Librarianship*, 31 (2), 160-163. Available at http://kcoyle.net/jal-31-2.html (accessed 28 April 2011).

Coyle, K, 'Understanding the Semantic Web: Bibliographic Data and Metadata', *Library Technology Reports*, January, 2010, 6 (1).

World Wide Web Consortium. Library Linked Data Incubator Group. http://www.w3.org/2005/Incubator/lld/wiki/Main_Page (accessed 28 April 2011).

■ **Karen Coyle**
Consultant
E-mail: kcoyle@kcoyle.net
Web: http://kcoyle.net

The DOI for this article is 10.1629/24154. Click here to access via DOI:
http://dx.doi.org/10.1629/24154
For a link to the full table of contents for the issue of *Serials* in which this article first appeared, click here:
http://serials.uksg.org/openurl.asp?genre=issue&issn=0953-0460&volume=24&issue=2