Next Generation Data Formats: Whats and Hows

MARC is a data structure -- a container designed to hold information, specifically bibliographic information. Designed in 1965, its original purpose was to provide a means for transmitting the data necessary to print library catalog cards.

The development of MARC begs questions regarding the definition of library catalogs. What is the purpose of the catalog? Through the use of authority lists, controlled vocabularies, and very specific filing procedures, the purpose of a library catalog is to list the inventory of a specific library as well as to create relationships between works ultimately building an intellectual cosmos out of an apparent chaos.

Fast forward to the present day. Computers are on every desk. They have more memory, more disk space, and more computing power than was ever imagined in 1965. Yet, meanwhile, back at the ranch, MARC still forms the core of our "integrated library systems". MARC records are copied from OCLC, parsed, and stuffed into databases, but because these databases are poorly normalized, it is a major task to maintain the content. On the other hand, the rest of the information world is using indexing technologies to provide search. No need to specify fields. Instead, free text queries and relevancy ranking derived from statistical analysis and harnessing the power of Web is the norm.

Not all gloom and doom

The purpose of this essay is to: 1) illustrate how the core elements of librarianship are relevant in today's environment, and 2) outline a method for reframing these elements. It is not the "what" of librarianship that needs to change but rather the "how". The profession's core mission -- to collect, organize, archive, and disseminate data, information, and knowledge for our respective communities -- is certainly valued by society and demonstrated by the exploding numbers of institutions who are doing the same things. On the other hand, the techniques of librarianship -- the "how" -- do not take advantage of the environment. Here's a way to get where we need to go.

Refine the definition of librarianship

Refine the definition of librarianship. Considering the current environment -- a milieu of electronic information and licensed content -- library materials may or may not be physically brought into our physical libraries. By harnessing the power of the 'Net, cataloging processes can be supplemented by tagging and reviews. Preservation brings on new challenges, but the principles remain the same. Make many copies of things and some of them will survive. When it comes to dissemination, consider the computer as the primary interface between librarians and patrons.

Constantly ask yourself, "What is my library's goal? Where do we want to be in one year? Three years? Five years? Ten years?" This is a never-ending process. The specific answers will change over time, but the framework will probably remain the same.

Reduce dependence

Reduce your dependence on third-party, "closed source" vendors to provide you with content and software solutions. As computers appeared, many libraries had their own "homegrown" library catalogs. These systems gave way to commercial systems. Through this process the profession shielded itself from technology. We computerized our work environment by mimicking our paper process -- automation. This created another silo since our systems did not work with other systems.

In the current environment we are increasingly licensing our content instead of owning it. Can we really depend on publishers to maintain content for decades? Is preservation best served by hosting content in a single location or purchasing insurance against its loss? Are access-only collections a viable long-term solution?

By supporting open access content and open source software to a greater degree, the library profession can ensure increased viability in the future. Open access content, by definition, can be copied without restriction. This addresses preservation issues and the ownership of content. Open source software makes our computing environment more flexible and transparent. It provides a way for libraries to have more control over their technical infrastructure.

Exploit technology

Exploit and combine the use of relational databases, indexing technologies, XML, and the Internet. A relational database is really a set of one or more lists of discrete data sets "joined" together by a set of common elements called "keys". Since data/information in databases is associated with these keys -- pointers -- and not necessarily specific values it is possible to do global find/replace operations throughout the database by editing only a single field. Ironically, databases are not

very good when it comes to search because users must know the structure of the database in order to search. If you want to provide search, then you want to employ an index. It functions exactly like back-of-the-book indexes. Feed an indexer a document. The indexer parses the document into individual words. It then saves the words, their position in the document, and a document identifier to disk. Given a word a search engine will loop through the word list and return document identifiers matching the query. Indexer/search engines excel at search for two reasons. First, a knowledge of the data's underlying structure is not necessary. Second, statistical analysis can be employed to calculate the relevance of a given document.

Learn to how two read and write XML files. Like MARC, XML is a data structure. Despite all of the debates for or against MARC and XML, the best reason for using XML is everybody else is using it. A large part of what it means to do librarianship is dissemination. If we want to share our content, then we need to share it in a language everybody else can understand. That language is XML.

Finally, learn how to exploit the Internet. With the advent of our global network it is easier to communicate and get the input from a much larger number of people. The era of centralized authority is waning. While centralized authority will not vanish completely, its importance is diminishing. Libraries need to use this to their advantage.

Work collaboratively

Work with sets of peers and stakeholders inside and outside your library to design and implement solutions to shared problems. When it comes to digitizing content in an academic library, consider digitizing content that is of interest to your local students, instructors, and scholars. While there is significant evidence that special collections content is of wide interest to the outside community, digitizing the content needed by your local constituents will pay off quicker. Increasingly data sets are becoming a part of the scholarly landscape. No longer is it satisfactory to do a set of experiments or conduct a number of surveys and then write about the results. It is becoming expected to make the data used to come to the scholarly conclusions available as well. Who is going to collection, organize, preserve, and disseminate this data? Figure out ways to make it easy for the institution to collect locally developed materials. They might include things beyond formally published articles. They could also include much of the gray literature: pre-prints, conference presentations, student research, etc.

"Next generation" library catalogs

Considering today's environment, people have no problem finding content. It abounds. It is so plentiful we still feel like we are "drinking from the 'proverbial' firehose." With all this information, whether we find it on Google or through a library system, the big question really is, "What are people going to do with the information?" Our catalogs need to reflect the changing expectations or we need to face the consequences of not keeping up with the times. To these ends, it is almost a trivial computing task to combine the technology of the day to create a "next generation" library catalog: 1) define a collection development policy, 2) build the collection, 3) describe and/or manifest the collection using XML, 4) manage the XML using relational databases, 5) make the XML searchable by indexing it, and 6) provide access to the index.

In my opinion, the real opportunity for "next generation" library catalog systems is not search, but services. The profession needs to figure out ways to enable patrons to use the data they find. Libraries, as opposed to Google, are in a unique position in this regard because libraries are more able to place content into the context of the user. Based on who the user is is possible to taylor search results and provide additional services against the content. Here is a list of possible services: add to my collection, annotate, cite, compare & contrast, create different version of, create flip book, create tag cloud from, delete from my collection, do concordance against, do rudimentary morphology, find opposite, find similar, hilight, incorporate into syllabus, map to controlled vocabulary term, plot on a map, print, rate, review, save, search, search my collection, share, summarize, tag, trace author, trace citation, translate.

Summary/conclusion

Libraries have an enormous set of opportunities available to them. It is fashionable to do library work. It is just that the work does not necessarily manifest through books, but rather the content of books and journals and images and data sets, etc. Moreover, it is not about tried-and-true library techniques. Instead it is about using the methods and technologies of the times. It is not the "what". It is about the "how".

Eric Lease Morgan
University of Notre Dame

May 1, 2008

http://infomotions.com/musings/ngc4mla/