

How Metadata and XML Relate to Existing and Future Cataloging Practices within the Library Field

April Younglove
ayounglo@emporia.edu
School of Library and Information Management
Emporia University
July 2006

Metadata: an overview

In her article, “Understanding Metadata and its Purpose,” well-known cataloger Karen Coyle (2005) wryly repeats the quip that “Metadata is cataloging done by men” (§ 1). Roy Tennant (2006a), author of *Managing the Digital Library*, makes the more politically correct joke that metadata is “cataloging done by those paid better than librarians” (§ 1). These jokes speak a fear that exists within the library profession. As more physical libraries adopt and incorporate technology from digital libraries, librarians are increasingly finding themselves in a field that is being described in technological terms. According to L. D. Matson and D. J. Bonski (1997), “a digital library is a library that maintains all, or a substantial part, of its collection in computer-processible form as an alternative, supplement, or complement to the conventional printed and microfilm materials that currently dominate library collections” (§ 6).

In the article “Do Digital Libraries Need Librarians,” author Lisa Dallape Matson (1997), librarian for the U.S. Department of Justice, explains that metadata is nothing more than the ‘library-centric’ concept of “controlled vocabulary” recast into “engineering terms” (§ 2). In other words, although librarians and systems engineers are using different terminology, they are “pushing the same boulder up the same hill” (Matson & Bonski, 1997, § 2). Some cataloging librarians fear though that by defining data management in computerized terms instead of in library terms, they, and the years of skill and insight that they bring to information architecture, will be displaced by technologists who value speed and quantity over grace and quality. Tennant (2006), however, proposes a more optimistic future where, “the modern cataloger will one day be a software-enabled specialist who can gather, subset, normalize, and enrich piles of records for a specific audience or purpose” (§ 7).

In the broadest sense, metadata includes any data that is about data, however, contemporary use of the term typically refers to “machine understandable information about web resources and other things” (Berners-Lee, as cited in Allard, 2002, p. 237). Tennant (2002a) more precisely describes metadata as “a structured description of an object or collection of objects” (§ 2), while Coyle (2005) prefers to describe it as “constructed information . . . not found in nature . . . made by people for a purpose or a function.” (§ 2-3). Coyle (2005) compares the way librarians help searchers by metatagging digital information in a catalog to the way maps aid readers by highlighting

the specific information needed for navigation. In other words, a hiking map will most prominently display walking trails and natural landmarks while a road map of that same area might point out detours and highway information instead. “Just as there is no single kind of map that serves all needs,” she explains, “there is no one kind of metadata for documents or other information objects . . . it is not the object itself that determines the metadata but the needs and purposes of the people who create it and those who it will serve” (Coyle ¶ 3).

Although the exact definition of metadata may be elusive, how it is used and why are clear. Metadata is used to search through vast amount of material quickly and accurately. The article “the Keys to the Kingdom”(2006), in the journal, *Broadcast Engineering*, illustrates the capabilities of metadata by asking the reader to imagine being a film archivist who needs every presidential speech in the last decade that mentions unemployment. The speeches must have been delivered at a luncheon and must not be State of the Union speeches. Looking through a list of titles or going through the physically archived tapes and loading each one to check its contents is next to impossible (¶ 1-4). Imagine how much easier that search would be if each tape had been cataloged onto a computer and if individual tape records had been tagged with searchable keywords, or metadata, like, “state of the union” and “luncheon.” This type of metatagging can be entered by a librarian, or tagged records might be released by the creators of the videos as part of their product. In an online environment, webmasters and automated software programs can embed information about posted documents “such as headline, abstract, byline, first paragraph, second paragraph and so on to modestly improve search results” (Crosman, 2006, ¶ 7). This “semi-structured information” can help search engines “present the headline, abstract, graphics and the first and last paragraphs of an article, [so that] the user gets a good idea of what it's about-much better than the typical document "snippet" that's often no use at all” (¶ 8).

Metadata can also be gathered from informal sources like social tagging or mined from unstructured information by automated means. Even members of the general public can participate in metatagging through a process called Folksonomy. At websites such as Flickr (<http://www.flickr.com/>), an online photo management site, and Del.icio.us (<http://del.icio.us/>), a social bookmarking site, users can create metadata by uploading digital photos and sharing website links and then publicly tagging them with any and as many search terms as they desire. While this process is joyfully non-elitist, it is also so non-standardized that it can become un-navigable and, at times, meaningless. Another way to generate metadata is to use data mining tools to automatically search through untagged documents (unstructured data) for key words and repeated phrases. Recently there have been attempts to apply this technology to organize, or “cluster” search results by logical human connections instead of by simplified machine matching (Crosman, 2006, ¶ 9). This process, called sentiment extraction, uses “concept search tools [to] put results in context, as in Paris the city versus Paris the person or Apple the company versus Apple the fruit” (¶ 18).

MARC: Good old-fashioned metadata

Although they may not have always used the term ‘metadata,’ librarians have long been experts in it. In the 1960s, librarians from the Library of Congress computerized their entire book catalog by working with systems engineers to create the Machine Readable Catalog (MARC) (Broadcast Engineering, 2006, ¶ 8). In the 70s and 80s MARC expanded to also include formats for objects like manuscripts and visual materials (Barta Norton, 2004, ¶2). Since then, “MARC formats have become a nearly ubiquitous medium for the exchange of bibliographic information among libraries, facilitating the growth of bibliographic utilities such as OCLC [the nonprofit Online Computer Library Center] and a new era of cooperative cataloging” (¶ 2). MARC has had the power, flexibility, and complexity to remain the industry standard metadata container for over thirty years (Tennant, 2002a, ¶ 3). In fact, MARC was so well-designed initially that until recently there hasn’t been any library interest in creating or using any alternative metadata containers.

In our rapidly-evolving digital environment though, the challenges to MARC are many. Can MARC accommodate object description beyond books and movies? What about art objects, slides, historical photos, digitized images, or even so-called “born digital” information like emails or full-text digital novels? Can MARC make these available to a wider audience outside of the library catalog? And how well can it network the results? (Barta-Norton, 2004, ¶ 1). Users of MARC are beginning to notice that the book trade uses XML to create records that “provide cover art, pull quotes from reviews, descriptive text, author biographies, and other useful material that MARC records typically lack” (Tennant, 2006, ¶ 2). In order to provide similar features, one issue MARC must address is how to provide increased granularity. Granularity refers to how finely data is chopped and stored (Guenther & McCallum, 2002, ¶ 12). In order to include a book review in a MARC record, one is “reduced to such questionable tactics as smashing it into a note field. As time goes on . . . we may begin to find that MARC isn't quite as extensible or granular as it will need to be” (Tennant, 2002b, ¶ 6).

Perhaps the biggest obstacle MARC faces is demonstrating a robust interoperability between metadata schemas (Barta Norton, 2004, ¶ 16). It seems like MARC would be an ideal format for this task, however, the standard structure used for sharing metadata online is XML, and XML is not MARC compatible (Coyle, 2005, ¶ 14). Until quite recently, there have not been any library-specific metadata formats complex enough to accommodate and transfer a MARC record without data loss. For MARC to be exchangeable with simpler, more sharable, non-library formats, it has to sacrifice its granularity. Because MARC records have a valuable high-quality granularity, Tennant advises that “it [granularity] should not be permanently surrendered through internal compliance with an external standard, unless the benefits clearly outweigh the drawbacks and no alternatives are possible” (Tennat, 2002b, ¶ 14).

Library 2.0 & the XML Push

In their controversial and much discussed white paper, “Do Libraries Matter? The rise of Library 2.0” (2006) authors Ken Chad and Paul Miller propose a radical re-conceptualization of libraries that would mimic recent Internet trends popularly dubbed

Web 2.0. These trends are largely built on XML and include “mash-up” technologies like blogging and RSS feeds (p.5). Chad and Miller’s Library 2.0 would operate “according to the expectations of today’s library users. In this vision, the library makes information available wherever and whenever the user requires it” (Chad & Miller, 2005, p.3). Their vision also includes a new global, multi-perspective, free catalog, built on a single recognized format. This proposed catalog will be made possible, they say, by “using the kind of salable distributed technology that Google, Amazon and others deploy” (p. 9) -- in other words, by eliminating the use of tools like MARC and ushering in XML instead (p.6).

XML, short for eXtensible Markup Language, was released by the World Wide Consortium in December 1997 as a public domain metalanguage. It is a general data format; it contains metadata but is not itself metadata. XML is like MARC because it is a human readable data container, but it is not like MARC because it is inherently neutral and “can be used for any number of applications. In particular, XML is often used as a document format, and is the broader format from which HTML is derived” (Coyle, 2005, ¶ 5). XML is also more flexible than MARC. A typical MARC entry would have to record the title of the book, *Anne of Green Gables*, like this: 245\$aAnne of Green Gables. In XML, however, it might look like <245>Anne of Green Gables</245>, <ti>Anne of Green Gables</ti>, or even <title>Anne of Green Gables</title>. It also has an unlimited hierarchical structure, unlike MARC which only allows for fields and subfields. This flexibility allows XML to maintain the structural integrity of more complex formats while making them exchangeable across platforms via the web. According to Suzie Allard (2002), Assistant Professor at the School of Information Sciences at the University of Tennessee, XML’s metadata and metalanguages “are regarded as two of the most vital for the future because they can answer many of the challenges facing digital library designers including interoperability, digital object description, user interfaces, architecture for collection organization and scaling” (p.237). Many publishing businesses have also adopted XML, and database provider giant LEXIS-NEXIS “is currently implementing XML into its databases, and has introduced its plans to information professionals” (239).

Not everyone is ready to throw out the baby with the bathwater yet though. Nancy Barton-Norton (2004), in her article, “MARC applications for description of visual materials,” states that because “millions of bibliographic records have been created using MARC,” a total overhaul of the MARC format would be “logistically, economically (and perhaps politically) unfeasible” (p. 55). In addition to raising questions about who should author XML metadata and ensure that it maintains its quality and granularity, some librarians also wonder how XML’s lack of universal cataloging standards will play out. Currently XML does not have a usage guideline to direct catalogers in preventing errors and misinterpretations like “the *Anglo-American Cataloging Rules (AACR)*, which defined how and with what information a cataloger was to fill the MARC metadata container” (Tenant, 2002a, ¶ 12-13). Coyle (2005) explains why these types of traditional cataloging standards are important for the benefit librarians as well as users: conformity allows cooperative cataloging and record exchanges. This collaboration is enabled by “library systems vendors,” who depend upon standardization, “to create a product that can be

used in any library, just as the standard sized catalog card could fit into any card catalog drawer” (§ 17). Although XML holds many exciting promises for librarians, most have decided that it is too early to declare MARC dead (Barta-Norton, 2004, p. 55).

MARC and XML: Beyond All or Nothing

In discussing the strengths and weaknesses of MARC and XML, it is important to point out that librarians have created and are creating a number of new data exchange formats that do not require users to make an either/or choice between the two. Dublin Core, a document description metadata, requires MARC users to sacrifice a great deal of granularity when translating records for sharing over the web. However, other formats, like MODS, can translate MARC almost directly into XML (Coyle, 2005, § 15). There is even a new standard, METS, that can either hold its own internal descriptive metadata, or point within its records to external data created in other standards like Dublin Core, MODS and MARC (Tenant, 2002a, § 8). As Tenant explains, “More standards are not necessarily a problem, as long as they can fruitfully interoperate and exchange data when required” (§ 10).

Although its records are not granular enough to support or replace MARC, Dublin Core (DC) has attributes that might make it a useful supplement. Because it is simple to use and has no cataloging rules (like AACR), Dublin Core is helpful for describing quirky data and materials that might not otherwise be cataloged. Dublin Core was created by an independent group in Dublin, Ohio and supported by OCLC¹, with the intent that it would encourage the cataloging of Internet materials at the most basic level (author, title, date etc.). The format contains only 15 core elements, although none of the elements are required to complete a record, and is so effortless that anybody can use it. DC fields are very broad, but can be clarified with subfields. For example, the core field “creator = Mark Twain” can be narrowed by adding additional information, such as, “Mark Twain = author” (Coyle, 2005, § 11). An archivist cataloging 1,000 photographs of a city with no other information other than, “Main Street, circa 1910” penciled on the back of several of the photos would not be able to organize them in MARC. With DC, however, she might scan them and then include: “date = circa 1910,” and “description = Main Street”. Although this record cannot be included within a typical library catalog, it can be posted online for the benefit of webcrawlers and search engines (§ 18-19).

“Interested experts,” (www.loc.gov/standards/mods/) in conjunction with the Library of Congress' Network Development and the MARC Standards Office created MODS, or Metadata Object Description Schema, in order to “complement other metadata formats and to provide an alternative between a simple metadata format with a minimum of fields and little or no substructure such as Dublin Core and a very detailed format with many data elements having various structural complexities such as MARC 21” (Guenther & McCallum, 2002, § 3-5). For the most part, MODS is a subset of MARC, translating its AACR records into XML friendly format by substituting human-understandable tags for numbers -- so that the MARC 245 field, in MODS, becomes “title” (Coyle, 2005, § 15). Unfortunately MODS does not actually add any additional capabilities to MARC beyond

¹ Dublin, Ohio, incidentally, being the home of OCLC

enabling XML file sharing, and, in fact, collapses a few of its data fields to make the records more exchangeable (Guenther & McCallum, 2002, ¶ 5).

METS, the Metadata Encoding and Transmission Standard, on the other hand, does not sacrifice any exchanged data because it not a true metadata container. Instead, METS is a wrapper that uses place holders to refer to foreign material. METS consists of six elements that contain or point to the digital object's metadata: a header, descriptive metadata, administrative metadata, file section, structural map and behavior section. With the exception of the header and structural map, all elements are optional and can be stored externally in any format. For internally contained data, METS prefers XML. Coyle describes METS as being like a digital binding and cover for the computerized files that make up a virtual book. Within this binding, METS has a sort of copyright page that provides the technical data about "file formats, the technology used in scanning if the item began its life on paper, and the digital transformations and compression that have been used on the files" (Coyle, 2005, ¶ 16). Although the LC serves as its maintenance agency, sponsoring its websites and listservs, METS is a non-proprietary open standard. It was conceptualized in 1990s as an alternative wrapper for the maintenance, storage and retrieval of digital objects in digital repositories and its form was finalized in 2001 by the Digital Library Federation (www.loc.gov/standards/mets). Guenther and McCallum (2002) hail METS as a resource that "can be used to collect digital resource metadata for submission to the repository, serve as the place for the metadata within the repository and be the supplier of information to the tools that provide the resources to the patrons" (¶ 33).

Although no existing metadata container is good enough to solve all of MARC's problems or replace it, the efforts to enhance MARC are still nascent and are valuable to those who are exploring metadata storage solutions. Tenant encourages readers to play with new standards as they emerge and see how they could interoperate. A digital novel, for instance, could be saved in MODS, but use METS to organize all of its files (2002a ¶ 9). He declares, "We're in a brave new world, in which MARC must make room for METS and MODS and whatever else becomes necessary to do our work better" (¶18). It may be, as he suggests, that as librarians use and investigate new formats, the best features of each will triumph and general bibliographic standards will logically emerge. Coyle (2005) is hopeful, though, that systems engineers will also realize that "it is the content of the metadata records, not their record structure, that makes the difference between a single-system solution and a coherent bibliographic universe" and predicts that "we may see that when metadata grows up, it becomes cataloging" (¶ 20).

References

- Allard, S. (Fall 2002). Digital libraries: A frontier for LIS education. *Journal of education for library and information science*, 43 (4), 233-248. Retrieved July 20, 2006 from WAW library Online course reserves.
- Barta-Norton, N. (2004). MARC applications for description of visual materials. *Journal of Educational Media & Library Sciences*, 42(1), 21-36. Retrieved July 7, 2006 from Wilson Web Database: Library and Information Science.
- Broadcast Engineering. (2006, June 7). *Metadata: The keys to the kingdom*. Retrieved July 7, 2006 from Academic Lexis-Nexis.
- Chad, K. & Miller P. (2005). *Do libraries matter? The rise of Library 2.0* [A white paper]. Retrieved July 9, 2006 from http://www.talis.com/downloads/white_papers/DoLibrariesMatter.pdf
- Coyle, Karen. (2005). Understanding Metadata and its purpose [preprint]. *Journal of Academic Librarianship*, 31(2), 160-163. Preprint. Retrieved from http://www.kcoyle.net/jal2_Metadata.html
- Crosman, P. (2006). The Perfect Search. *Intelligent Enterprise*, 32. Retrieved July 7, 2006 from Academic Lexis-Nexis.
- Guenther, R. & McCallum S. (2002). New Metadata standards for digital resources: MODS and METS. *Bulletin of the American Society for Information Science and Technology*, 29(2), 12-14. Retrieved July 9, 2006 from http://www.findarticles.com/p/articles/mi_qu3991/is_200212/ai_n9150534
- Matson, L.D. & Bonski, D. J. (1997). Do digital libraries need librarians? An experimental dialog. *Online*, 21(6), 87-92. Retrieved from WAW library Online course reserves.
- Tennant, R. (2002a, April 15). Digital libraries – Metadata as if libraries depended on it. *Library Journal*, np. Retrieved July 13, 2006 from <http://www.libraryjournal.com/article/CA206408.html>
- Tennant, R. (2002b). The importance of being granular. *Library Journal*, 127(9), 32-4. Retrieved July 7, 2006 from Wilson Web Database: Library and Information Science.
- Tennant, R. (2006, April 15). The new cataloger. *Library Journal*, 32. Retrieved July 7, 2006 from InfoTrac.