Title
“Why Metadata? Why Me? Why Now?”

Author
Schottlaender, Brian E. C.

Publication Date
2002-03-08

"Why Metadata? Why Me? Why Now?"

ALCTS Metadata Institute
San Diego, California
8 March 2002

Brian E. C. Schottlaender
University Librarian
University of California, San Diego
Email: becs@ucsd.edu

Transcribed by
Larry Heiman
Head, Copy-Cataloging Team
University of California, Irvine
Email: "Larry W. HEIMAN" <LHEIMAN@uci.edu>

Revised by
Brian E. C. Schottlaender

Abstract:

This introductory overview will consider why metadata issues are central to discussions about the evolution of library services—particularly digital library services—and why the cataloging community is, and should be, front and center in those discussions.

There is a young woman who works in my office who is responsible for taking minutes at various meetings that take place within the UCSD Libraries, including those of the library management group we simply call Cabinet. She had been doing that beautifully for many, many months when finally she approached my deputy one day and asked, "What is this meetadata [sic] I keep hearing about?" So I thought about calling my talk this morning, "So what's up with metadata?," but decided instead to pose the three questions, "Why metadata? Why me? Why now?" and to do that as a means of providing an introductory overview and framework for the presentations to follow. While I intend, at the end of my presentation, to briefly answer the three questions, I plan also to use them as a way of framing what I hope will be a useful overview of both the information environment and the professional environment in which we finds ourselves.

Why metadata?

If anything characterizes the information universe in which we find ourselves today, it is its fluidity. There are a proliferating number of information resources, a number of which have highly specialized needs, and many of which are complex and packaged in such a way that their individual components may actually require management. It is these characteristics of the information universe itself that inform the characteristics of the metadata environment as well.

As we all know, that there are probably as many definitions of metadata as there are people whom one asks. In fact, when the ALCTS CC:DA Task Force on Metadata summarized in an appendix to its *Final Report*[1] the various definitions that they came across in the course of their work, they included well over 25 such definitions. One of the best definitions they included was Clifford Lynch's, "a cloud of collateral information around a data object."[2] What I like about the definition is its use of the words "cloud" and "collateral." "Cloud" evokes for me that character in *Peanuts* who always had that cloud of dirt all around him, while "collateral" makes clear the inherent relationship between data and their metadata. The ALCTS CC:DA Task Force on Metadata used the definitions that they found in the environment to craft their own definition. In many ways, of the formal definitions of metadata that exist I consider theirs still the best, with Lynch's standing as the best informal definition. The Task Force definition is this: metadata are "structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities."[3] Whether one considers metadata to be structured data that describe the characteristics of a resource or clouds of collateral information around data, there is an inherent relationship between metadata and the information objects they describe.

Having said earlier that one of the characteristics of the information environment in which we find ourselves is a proliferating number of resources, the corollary to that is: metadata, metadata everywhere. Also, having said earlier that the resources in the digital information environment in which we find ourselves are increasingly specialized, increasingly fluid, and increasingly complex, then the implication on the metadata side is that metadata are having to do more and more things. A word one hears a lot in discussions of metadata is the word "schema" (kind of a fancy word for scheme). There are a lot of definitions of that word as well. The most useful that I've come across thus far is Murtha Baca's in her *Introduction to Metadata* that she edited for the Getty Research Institute. She defines schema, I think quite usefully, as "a set of rules for encoding information that supports specific communities of users."[4] There are three kinds of schema I want to talk about and those are encoding schema, metadata schema, and architectural schema.

Encoding schema are many; the four I want to mention are MARC, SGML, HTML, and XML. MARC, with which we are arguably the most familiar, is the standard structure for encoding Machine Readable Cataloging data, most often bibliographic and authority data. SGML (Standard Generalized Markup Language) is an international standard, ISO 8879, that prescribes a format for embedding descriptive markup within a document and then goes on to specify a method for describing the structure of that document. SGML basically has three claims to fame: first is its extensibility, second its structuring capabilities and, third its validation capabilities. SGML is crafted in such a way that it allows one to deal with the complex package resources mentioned earlier. It allows one to do that in a highly structured fashion and it allows one to do that in such a way that one can validate the structure as the markup is taking place. Those are its upsides; it has downsides as well. It is extremely complex and applying it demands, frankly, a rigor that many of us don't care to invest.

Consequently, the markup language with which most people are more familiar and more comfortable is HTML (Hypertext Markup Language). There is a widespread perception, and frankly a misperception, that HTML is actually "dumbed-down" SGML. In fact, while HTML is based on SGML, it is not a subset of SGML. The people who created it definitely had SGML in mind when they did so. It is intended for marking up hypertext, multimedia, and reasonably small and simple documents.

The most recent markup language talked about in the community is XML (Extensible Markup Language), conceived basically as a happy medium, if you will, between HTML and SGML. In fact, XML is a simplified subset of SGML intended for Web applications. It retains SGML's extensibility as is implied in its name, along with SGML's structure and validation capabilities, but it's much simpler to apply.

The number of metadata types is proliferating as the information metadata are intended to manage proliferates as well. Whereas once upon a time long ago—say, five years—when people used to talk about metadata and how complex the world of metadata was, they were usually talking about four basic types—descriptive, administrative, technical, and rights—there are now a good many besides, including:

- security
- personal information (e.g., Vcard)
- commercial management (cost, etc.)
- content rating
- preservation

I shall focus now primarily on descriptive metadata, including reviewing several of the schema being used in the descriptive metadata community.

In the cataloging community, ISBD (the International Standard Bibliographic Description), is the widely adopted schema for describing many types of library materials. AACR2 (the second edition of the *Anglo-American Cataloguing Rules*), meanwhile, is a content standard for bibliographic data relating to library materials and for formulating access points for authors, titles, related works, etc. The PCC (Program for Cooperative Cataloging) core record standards—of which there are now almost a dozen, for everything from books to multiple character sets—are all MARC-based descriptive metadata schema.

The non-cataloging community descriptive metadata schema familiar to most is Dublin Core (known officially as the "Dublin Metadata Core Element Set"). It is, as its name implies, a core set of elements, a simple set of data elements, meant to be used to describe and to facilitate discovery of document-like objects in a networked environment. That phrase "document-like objects" is important because the digital information universe is full of digital objects that are not "document-like" and Dublin Core (DC) may or may not be useful in describing those objects.

In the governments documents community metadata schema in use include AGLS (the Australian Government Locator Service), a DC-based descriptive schema. FGDC, meanwhile, is a metadata standard developed by the Federal Geographic Data Committee for digital geospatial metadata. It, in turn, is based on CSDGM (the Content Standard for Digital Geospatial Metadata), an XML-based metadata standard.

In the art community, the REACH (Record Export for Art and Cultural Heritage) Element Set is a metadata standard elements set developed by RLG, the Research Libraries Group. The core elements in the set are drawn from various standards in the cultural heritage and art environments, including the CDWA (Categories for the Descriptions of Works of Art), the Data Dictionary created by MESL (the Museum Educational Site Licensing project), and the Access Points crafted by CIMI (the Consortium for the Computer Interchange of Museum Information). The REACH Element Set is interesting because it functions as a meta-metadata set to the extent it incorporates a variety of data elements from other metadata standards. The VRA Core Categories, meanwhile, is a metadata elements set created by the Visual Resources Association Data Standards Committee to describe visual resources and the images that describe them.

Two other important descriptive schema are the TEI and EAD Headers. Both TEI (Text Encoding Initiative) and EAD (Encoded Archival Description) are document types within SGML. Their headers, in turn, are those sets of elements within their respective document-type definitions designed to contain identifying data about instances of each. Both function, in effect, as metadata supersets.

There are two non-descriptive metadata schema I want to mention just briefly, PICS and A'Core. PICS (Platform for Internet Content Selectivity) was one of the first and remains one of the most robust content rating initiatives. PICS metadata are instrumental in a commercial search engine's ability to allow one to restrict access to certain kinds of Internet content. A'Core (Admin Core), meanwhile, is a metadata standard for metadata. Systems use A'Core metadata "to determine the currency and integrity of content metadata, and provide details on how to contact entities involved with the management of content metadata."[5] In the increasingly complex systems within which we

are working, there are multiple metadata standards or schema at work managing different kinds of objects. It is the Admin Core that helps systems keep track of all these metadata.

A word now about identifiers, a highly concentrated kind of descriptive metadata. ISBN and ISSN—the International Standard Book and Serial Numbers—are probably the most familiar; ISAN, the International Standard Audiovisual Number, perhaps less so. Beyond these, there are variety of URIs, or Uniform Resource Identifiers, the most familiar of which is the URL—the Uniform Resource Locator that characteristically begins "http://". Other types of URIs include URNs and URCs, or Uniform Resource Names and Uniform Resource Characteristics. Both are intended to function much the same way as URLs do, but are intended to obviate some of the problems that URLs have including, most notably, the propensity that URLs have to change regularly. It is hoped by those who are working on URN and URC development that names and characteristics are more constant than are locations. Whether that is so remains to be seen.

With some exceptions, the descriptive metadata schema I have just described generally focus on syntax. Semantics has remained the domain of library cataloging, notably as manifest in AACR2. In other words, many metadata schema focus on statements like, "When you describe an object, you should be sure to include in that description its title, creator, etc.." The schema rarely go on to tell one how to do that. Instead, it is the library cataloging community, and its AACR2, that have focused on deciding whether an object has a title or a creator and, if so, how to formulate them.

Having talked a little bit about a variety of descriptive schema, let me now talk about three architectural schema. The first, INDECS (Interoperability of Data in E-Commerce Systems), is important in part because of who created it. It was established to integrate a variety of standards

developed by communities that concern themselves with copyright, including the copyright societies' CIS (Common Information System) plan, the record industry's ISRC and MUSE projects, the audiovisual community's ISAN initiative, the publishing industry's ISBN and ISSN initiatives, and the DOI initiative.[6] In our environment the two architectural schema that you are more likely to have heard of are RDF and the Warwick Framework. RDF—the Resource Description Framework—is an infrastructure for "encoding, exchange, and reuse of structured metadata."[7] The Warwick Framework sort of extends that concept, and is actually "a container architecture for diverse sets of metadata."[8] In other words, it's a comprehensive infrastructure for network resource description. At UC San Diego's Supercomputer Center, computer scientists have developed something called the "Storage Resource Broker," or SRB, which is predicated on the Warwick Framework. It is a software suite that allows one to pull a variety of digital objects into a container architecture that can handle basically any kind of metadata. The container architecture really doesn't care what metadata schema you have used to encode your data because it manages everything at a kind of super ordinary architectural level.

Why me?

Let me talk a little bit now about the second question, "Why me?" Metadata is about access. Cataloging is about access. Cataloging describes content and content relationships. Kevin Butterfield in his paper "Catalogers and the Creation of Metadata Systems" talks about cataloging as an "invisible process of order-making."[9] At the risk of stating the obvious, the Internet could use some order-making. Cataloging, Butterfield points out and we all know intuitively, is not about rules; nor is it about the records those rules result in. It's about standards, it's about vocabulary development, and it's about the development of systems for description and classification. Those are processes the cataloging community has been involved in for a long, long time—decades, if not

centuries—and, we have a lot of experience and expertise to offer other communities who think that these issues are only now being considered for the first time. The ALCTS Committee on Cataloging: Description and Access, Task Force on Metadata and the Cataloging Rules in its *Final Report* dated August 1998 noted that "Catalogers need to be involved with emerging metadata standards. Our bibliographic and cataloging expertise is invariably useful and often welcome in defining data elements and preparing usage guidelines."[10] I'm not sure that "often welcome" was actuated in 1998. Now, four years later, I think it is much more accurate.

In that report, one of the co-authors, John Attig of Pennsylvania State University, observed that the intersection between cataloging and metadata is, or should be, the common user tasks that they support. IFLA's *Functional Requirements for Bibliographic Records*[11] (FRBR) describes these common user tasks as fourfold—find, select, identify, and obtain. I suspect that when Attig uses the word "common" he is using it in two ways. First of all, the tasks themselves are common; second, they are common to a very broad spectrum of users. Cataloging and metadata, I would suggest, intersect in the degree to which they support these tasks.

Why now?

Earlier, I noted that most metadata schema have focused on syntax rather than semantics. Putting it another way, most, if not all, content standards are library-based—AACR, AAT (Art & Architecture Thesaurus), LCSH, APPM (Archives, Personal Papers, and Manuscripts). They all came out of the library community. Very few, if any, content standards exist outside the library community. It is my perception, apropos of the observation in 1998 that cataloger's expertise is "often welcome," that the initial non-library ambivalence, if not outright hostility, to content standard development is beginning to evaporate, and that there is a growing recognition in the non-library community of the

utility and desirability of content standards. In fact, I am seeing an increasing confluence between the cataloging and metadata communities. So much so, that the two communities are becoming harder and harder to distinguish, which is exactly as it should be. Consider, for instance, the following from Stuart Weibel:

> The 15 Dublin Core elements might be more coherently expressed if they are related to an underlying logical model such as that expressed in the Functional Requirements for Bibliographic Records (FRBR) of the International Federation of Library Associations.[12]

Coming from the same person who had earlier put the Dublin Core forward as an alternative to cataloging, the statement approaches revelation. The Dublin Core "qualifiers" that you will have heard reference to are basically an attempt now to enrich the DC Element Set by referencing a variety of content standards—subject thesauri, authority control systems, classification systems.

So, on the one hand, there is a growing recognition in the metadata community of the relevance of the work that we in the library cataloging community have been doing. On the other hand, commercial and legal interests in rights management may be what bring the communities even more closely together if for no other reason than the fact that rights management requires a degree of descriptive specificity that is characteristically practiced, thus far at least, by the cataloging community and not by the metadata community. Put in FRBR terms, whereas the focus of the metadata community has been on "find" and "obtain," the "identify" task, which is essential to effective rights management, has been the focus of much greater effort within the cataloging community. The *Final Report* of the ALCTS Task Force on Metadata and the Cataloging Rules stated: "Our catalogs have become one tool among many, but those many are not separate or isolated from one another. The catalog is one tool in a network of tools."[13]

I want to close by talking a little about the challenges and opportunities that are in the environment now for us in the metadata and cataloging communities. A significant challenge is that of the degree of "fixity," or lack thereof, of e-documents. "Fixity" is a concept most recently articulated by David Levy of the University of Washington's library school faculty to refer to how fixed in time and space documents are or are not. Levy observes that in a print environment documents are much more fixed in time and space than they are in a digital environment.[14] I would like to suggest that the degree to which electronic documents are not fixed in time and space has serious implications for the necessarily dynamic nature of the metadata associated with those documents.

Another challenge is that of content standards. While it is true that there is a growing confluence between the cataloging and metadata communities, there is still an incredible amount of work to be done on content standards and—related to that—on controlled vocabulary sets. Development of the latter development gets particularly complex when it needs to take place across various sectors of the content community: the vocabulary set that the art and architecture community is comfortable with is going to be quite different from that the social sciences community is comfortable with. Consistent deployment of metadata across variant content communities is going to be a challenge, as is harmonization (a word I prefer to "compatibility") of metadata sets.

Perhaps more so than anything, interoperability issues are going to be challenging. Interoperability is defined as "the ability of two or more systems or components to exchange information and use the exchanged information without special effort on either system."[15] When the communities talk about interoperability, they invariably mean syntactic, semantic, and structural interoperability. These are challenging technical issues to be sure. As challenging, however, are the cultural interoperability issues that persist between the communities themselves.

To return, in closing, to my three questions: "Why metadata? Why me? Why now?" I'll respond by saying because it is inescapable and seemingly more evident everyday; because it is what we're about; and, finally, because not only do we need it as another in our network of tools to do what we do, but it needs us to help fully realize its potential.

REFERENCES

[1] Association for Library Collections & Technical Services. Committee on Cataloging: Description and Access. Task Force on Metadata. *Final Report* (June 2000).
<http://www.ala.org/alcts/organization/ccs/ccda/tf-meta6.html>

[2] *Ibid.*

[3] *Ibid.*

[4] Baca**,** Murtha, ed. *Introduction to Metadata.* Los Angeles: Getty Research Institute, 1998.
<http://getty.edu/research/institute/standards/intrometadata/>

[5] Iannella, Renato and Debbie Campbell. "The A-Core: Metadata about Content Metadata." (Internet-Draft, 30 June 1999).
<http://metadata.net/admin/draft-iannella-admin-01.txt>

[6] Bearman, David, Eric Miller, Godfrey Rust, Jennifer Trant, and Stuart Weibel. "A Common Model to Support Interoperable Metadata: Progress Report on Reconciling Metadata Requirements from the Dublin Core and INDECS/DOI Communities." *D-Lib Magazine* (January 1999).
<http://www.dlib.org/dlib/january99/bearman/01bearman.html>

[7] Miller, Eric. "An Introduction to the Resource Description Framework." *D-Lib Magazine* (May 1998).
<http://www.dlib.org/dlib/may98/miller/05miller.html>

[8] Lagoze, Carl. "The Warwick Framework: A Container Architecture for Diverse Sets of Metadata." *D-Lib Magazine* (July/August 1996).
<http://www.dlib.org/dlib/july96/lagoze/07lagoze.html>

[9] Butterfield, Kevin. "Catalogers and the Creation of Metadata Systems."
<http://www.oclc.org/oclc/man/colloq/butter.htm>

[10] Association for Library Collections & Technical Services. Committee on Cataloging: Description and Access. Task Force on Metadata and the Cataloging Rules. *Final Report.* (August 1998).
<http://www.ala.org/alcts/organization/ccs/ccda/tf-tei2.html>

[11] IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records: Final Report.* München : K.G. Saur, 1998.

[12] Weibel, Stuart. "The State of the Dublin Core Metadata Initiative." *D-Lib Magazine* (April 1999).
<http://www.dlib.org/dlib/april99/04weibel.html>

[13] Association for Library Collections & Technical Services. Committee on Cataloging: Description and Access. Task Force on Metadata. *Final Report* (June 2000).
<http://www.ala.org/alcts/organization/ccs/ccda/tf-meta6.html>

[14] Levy, David. "Fixed or Fluid? Document Stability and New Media" in *Proceedings of the 1994 European Conference on Hypermedia Technology.* ACM Press, 1994.

[15] Association for Library Collections & Technical Services. Committee on Cataloging: Description and Access. Task Force on Metadata. *Final Report* (June 2000).
<http://www.ala.org/alcts/organization/ccs/ccda/tf-meta6.html>