Chapter 1

Metadata Developments in Libraries and Other Cultural Heritage Institutions

Abstract

This issue of Library Technology Reports *(vol. 49, no. 5) "Library Linked Data: Research and Adoption" focuses on research and practice related to library metadata. In order to more fully understand this world, we also need to consider the work being done in the archival and museum communities. In chapter 1, we lay the foundation for our exploration in libraries, archives, and museums (LAM) and consider the role and impact of this work in the broader world of the Semantic Web, linked data, and data-rich web services. This chapter starts by introducing a model for understanding the component parts of metadata systems and concludes by outlining the process for creating and publishing linked data.*

Introduction

This issue of *Library Technology Reports* (*LTR*) builds on previous work in this series, including the *LTR* issue by Coyle on the Semantic Web as well as Witt's issue on the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) standard, and it touches on topics referenced in Breeding's 2009 issue on web services and SOA as well as Nagy's 2011 case study analysis of library uses of next-generation discovery platforms.[1] In fact, this issue saw its first iteration in 2007, when Eden discussed metadata and information organization issues in libraries.[2] In that issue, Eden explored current metadata issues and asked what information organization might look like in the coming years. At the time, Library 2.0 and web services were newly emerging terms in library and information science literature, and while there was a vision for what library metadata and information systems might become, there were not many examples in the field.

Much has happened between these *LTR* issues, not the least of which has been the funding and creation of new national and international organizations whose goal is to bring together and publish the collections of cultural heritage and memory institutions. Since 2007, LAM (libraries, archives, and museums) communities have developed new cataloging and archival processing frameworks (e.g., RDA, DACS, and CCO) and are keenly interested in exploring the impact of new information systems on user needs. In the library world, this discussion has led to the BIBFRAME initiative and a focused effort to implement the Resource Description and Access (RDA) specification. In museum communities, the International Council of Museums has updated the CIDOC Conceptual Reference Model (CIDOC-CRM) as well as the Cataloging Cultural Objects (CCO) specification. In archives, standards like Encoded Archival Description (EAD), Describing Archives: A Content Standard (DACS), and Encoded Archival Context (EAC) have taken hold, and new information systems like ArchivesSpace are emerging to serve archival metadata and object management needs.

ArchivesSpace
www.archivesspace.org

The definition of new standards and development of systems across LAM and publishing communities are focused on defining techniques, standards, systems, and services that meet the changing information sources and needs of LAM patrons. The patron's information requirements are grounded both in a need for physical and digital information artifacts and also in a

community of practice in which making connections among information sources is as important as discovering information resources. While part of the LAM profession focuses on literacy and information engagement issues, the metadata community typically focuses on exploring how information systems and structures support these needs and enable cross-community and cross-repository access and aggregation.

As community needs change and metadata standards and systems are developed to meet these needs, LAM institutions are revisiting common questions of metadata exchange, migration, interoperability, scale, sustainability, and value. This issue of *LTR* focuses on these questions as we explore three continuing efforts in the LAM metadata community: the Library of Congress BIBFRAME initiative, the Europeana digital library, and the Digital Public Library of America (DPLA). The metadata systems of these efforts were chosen for exploration because they represent major initiatives in LAM metadata and because each community is interested in developing and delivering metadata-rich solutions that have a national and international impact.

In order to better explore the general direction of metadata research and practice in LAM communities, this issue begins by examining the broad direction of metadata development and research in chapter 1, continues with a deep dive into the underlying standards and building blocks of metadata systems in chapter 2, and broadens back out with case study explorations in chapter 3. In chapter 4, we explore the broad questions of metadata research and development within the context of our case study analyses.

A Brief History of LAM Metadata

The MARC standard that has served libraries well for the last forty years includes a robust metadata schema, an efficient exchange standard, and a detailed encoding and data storage system.[3] While we often talk about MARC in bibliographic terms, it is a much wider standard that includes bibliographic description, storage of holdings data (MARC21 Format for Holdings Data—MFHD), storage of authorities and classification data, and storage of community data.[4] These MARC formats have provided libraries with a detailed and interconnected ecosystem within which they could create, share, and validate records and authorities.

In addition to developing MARC to store bibliographic and other library-specific data, LAM institutions also explored the use of MARC in archival and museum settings. These uses included MARC Format for Archival and Manuscripts Control (MARC AMC); MARC Archives, Personal Papers, and Manuscripts (APPM); and MARC for Visual Materials (MARC VM). While MARC flourished in library settings, these additional implementations found mixed success in archival and museum contexts, and in the last decade other cataloging models and dissemination methods have superseded these efforts.[5] The museum community, for example, has found value in a specialized cataloging protocol, Cataloging Cultural Objects (CCO),[6] and the archival community has found value in Encoded Archival Description (EAD), Metadata Encoding and Transmission Standard (METS), and the cataloging specification Describing Archives, A Content Standard (DACS).

While these new standards emerged based on a need to accommodate new types of resources, one issue that has persistently confounded metadata work in the digital age is the complications associated with working with digital surrogates of physical resources.[7] This is not for lack of appropriate data models and cataloging approaches but rather due to the natural tension between creating highly accurate metadata and creating metadata efficiently. In the Dublin Core (DC) community, this issue is discussed as the "one-to-one principle." For example, in cataloging a digital surrogate in DC, it is easy to blur the line between a physical object and its digital surrogate by using the date of a historic photo in the `dc:date` field while providing the URL to the digital surrogate in the `dc:identifer` field. While the Qualified Dublin Core (QDC) standard introduces new properties that help differentiate metadata about physical objects from metadata about their digital surrogates, the issues of specificity, cost, and value in relation to metadata work means that adopters may find themselves, with the best interests of their institutions at heart, making choices out of sync with the standard.

A second important area of research and practice in metadata is that surrounding metadata provenance and version control. As metadata is increasingly shared, aggregated, repurposed, and reused, understanding where the metadata came from and how it was intended to be used is important in ensuring accurate disambiguation, deduplication, and rights tracking. As we will see in our case study exploration, being able to track metadata sources at a very detailed level is important for resolving discrepancies in description as well.

The Motivation for a New Approach to Metadata

As libraries have moved through iterations of discovery platforms (faceted discovery, federated search, web-scale discovery), issues of metadata schema harmonization, system scale and performance, and metadata quality guided the developmental direction of systems. Nagy's *LTR* issue used a case study approach to explore how libraries had addressed these issues.[8]

In addition to the user interactivity foundation of next-generation catalogs, one of the key features that differentiated faceted catalogs from more traditional integrated library system catalogs was the use of a faceted search platform. In recent years, open-source tools like VuFind, DSpace, and Blacklight have capitalized on the speed and flexibility of the Apache Solr indexing platform to deliver lightweight and fast discovery services.

VuFind
http://vufind.org

DSpace
www.dspace.org

Blacklight
http://projectblacklight.org

Apache Solr
http://lucene.apache.org/solr

Two key technologies facilitated the development of these tools. The first was the use of industry-standard ways of recording data. Also known as "serialization," the shift away from MARC 21 storage formats to XML-based formats made it much easier to bring together bibliographic data and data from archives and museums, which was more likely to be in an XML-based format. MARC-to-XML converters like SolrMarc were the building blocks that made this transition possible. The second key technology that enabled widespread experimentation with new discovery platforms was the adoption of application programming interface (API) techniques in system design. APIs are an important design feature of Solr that allow modular development of metadata management and discovery systems.

SolrMarc
http://code.google.com/p/solrmarc

In the last few years, these discovery services have begun shifting from localized implementations of catalogs to cloud-based services that include not only local holdings but also journal database and full-text indexing. The deployment of discovery environments in the cloud requires new techniques for metadata management as well as new data structures and database platforms that perform well at this higher scale. Some examples of these environments include open-source platforms like Open Library and the Kuali Open Library Environment, community services like LibraryThing, and enterprise-level web-scale library service environments.

Open Library
http://openlibrary.org

Kuali Open Library Environment
www.kuali.org/ole

LibraryThing
http://librarything.com

While libraries are forging ahead with these new platforms, there is still a pressing need for a continued fundamental restructuring of the metadata models and records that serve as the foundation for these systems. While the MARC metadata standard is generally perceived as a standard whose time has passed,[9] it has also been credited as a standard that formed the foundation of computerized resource description and fostered widespread standardization of bibliographic description.[10] In rethinking approaches to bibliographic metadata, Miller and his colleagues point to three core functions of MARC that need to be replicated in new standards and systems: the capture of data about the "intellectual essence of a work," the capture of data related to the "actual instance of the work," and the capture of data to uniquely identify (e.g., LCCN, ISSN, ISBN) and situate (e.g., related works, series statements) a work in a larger body of knowledge.[11]

This interest in capitalizing on rich and complex metadata while also developing new models that are naturally interoperable with other standards is a common theme in metadata discussion. For example, although there is a history of rich descriptive practice in archival institutions, the practice has also been shown to have difficulty scaling to meet the addition of new archival materials, making it more difficult for archives to serve patron needs.[12]

Greene and Meissner's focus on the user is another common theme in metadata research. This focus was expanded from the notions in the Paris Principles during the definition of the Functional Requirements for Bibliographic Records (FRBR) and *IFLA Statement of International Cataloguing Principles* models.[13] While the bibliographic community largely supported FRBR, it also has found implementing the model difficult, in part because of metadata quality issues, but also because of the inherent difficulty associated with shifting from one model for representing resources to another.[14] In fact, fifteen years after its definition, the model has yet to be fully adopted, and as new specifications like BIBFRAME are emerging, the library community may be signaling an interest in moving on without ever having realized this goal. Finding this balance between what

is achievable (e.g., "more product") and the interest in preserving the granularity and specificity of metadata has been at the center of discussions of cost, value, and relevance.

The RDA community, for example, has spent considerable effort in building out a new series of cataloging rules geared toward accommodating an increasingly complex world of information objects. This complexity has been the source of debate in the library community[15] and continues to be a popular topic on cataloging discussion lists. In April 2013, for example, a poster to the AUTOCAT Listserv suggested a "full stop" in the implementation of RDA given the "piecemeal" approaches to implementation that had been discussed previously.[16] In ten days, this post garnered seventy-two responses that debated the core issues of RDA adoption. The interest in these topics has also been seen in the enthusiastic reception of the DPLA release and the detailed discussions of the still-developing BIBFRAME specification.

In this section, we have touched on metadata-related issues of web integration, adherence to industry standards, metadata management at web scale, the need to meet user needs, and the need to explore new solutions and develop new tools. At the same time, age-old issues of metadata quality, richness, granularity, specificity, cost, value, and sustainability are part of the critical framework with which we evaluate each new schema and technology.

The broad direction of current research in the metadata world involves work in the design and application of linked open data (LOD) and linked open vocabularies (LOV). New communities have been formed to discuss these new structures, and the LAM world has reached out to existing communities like the World Wide Web Consortium (W3C) to collaborate in the development of the underlying specifications of LOD and LOV. In the remainder of this chapter and throughout the rest of this issue, we explore LOD solutions and consider how both the building blocks of LOD and the specifications deployed on top of them answer the technology, design, application, use, value, and sustainability questions that are central in all metadata research.

A General Framework for Discussing Metadata

In chapter 2, we will explore the building blocks of LOD systems, including the general rules for LOD, a general data model for LOD (the Resource Description Framework), and the design and application of vocabularies and ontologies. In preparation for that discussion, we need to have an understanding of the building blocks of metadata itself. Metadata is not just a definition of fields into which content is placed or a way of encoding (i.e., serializing) that information. Metadata is comprised of a data model, rules for how content is formatted, rules for how content is represented, rules for how content is stored, and rules for how content is exchanged. Regardless of the technology or philosophical foundation of a metadata platform, these building blocks play key roles in defining how a metadata system works and what it is capable of doing. These constructs are discussed in depth by Elings and Waibel as data fields and structure, data content and values, data formats, and data exchange.[17]

Table 1.1, adapted from Elings and Waibel's work, shows the connection between four building blocks of metadata systems and some example standards for each component. The third column suggests alternative terminology that is commonly used in literature to discuss these types of systems. A fifth building block, labeled *data model*, is also listed in this table. While Elings and Waibel connected data model with data structure, data models in an LOD context deserve separate consideration. The data model that is the focus of LOD is the Resource Description Framework (RDF).

To enable us to start on common ground, table 1.2 unifies the content from table 1.1 and defines each of the concepts in this aggregated model. While a detailed exploration of these concepts is beyond the scope of this issue, the interdependencies and relationships among them should become apparent in chapters 2 and 3.

There are other ways of parsing metadata schemas, including the use of types of roles (e.g., descriptive, administrative, technical)[18] or the use of metadata schema features.[19] In this issue, we rely on these building blocks so that we can understand the technical components of metadata rather than the specific functions and roles of the various schemas and vocabulary.

Conclusion

This issue of *LTR* features a number of figures and tables based on data harvested from the linked data services discussed in chapters 2 and 3. Because of the complexity inherent in RDF-based data, you will find that these figures often contain a small snapshot of data from a series of RDF statements and may be difficult to read. In order to make it possible for readers to replicate these figures as needed and access entire example records from linked data services, all of the appendixes, figures, and tables and the files used to generate them have been checked into a GitHub repository. In addition, this issue mentions data dissemination services including application programming interfaces (APIs), data visualization tools including Gephi, and data querying tools (SPARQL). Rather than documenting the process for accessing and using these tools in this issue, tutorials have been created and are

Table 1.1
Metadata concepts and structures (adapted from Mary W. Elings and Günter Waibel, "Metadata for All: Descriptive Standards and Metadata Sharing across Libraries, Archives and Museums," *First Monday* 12, no. 3 (March 2007): 7–14, http://firstmonday.org/ojs/index.php/fm/issue/view/225)

Elings & Waibel's terminology	Schema examples	Other popular terminology
	RDF, Entity-Relationship, Key-value, Graph database	Data model
Data content	CCO, AACR2, RDA, DACS	Content rules, cataloging principles
Data structure	CDWA, MARC, EAD, ONIX, OWL, SKOS	Metadata schema
Data format	XML, XML ISO2709, JSON, JSON-LD, RDFa	Encoding, Serialization
Data exchange	OAI, Z39.50, SRU, SPARQL	

Table 1.2
Definitions of the five building blocks of metadata structures

Metadata building block	Definition
Data model	The way in which relationships between resources and their metadata and among resources are documented. In essence the data model is the foundation on top of which the other components are built.
Content rules	Content rules govern how information is extracted or generated from resources and used to create a representation. Examples of content rules include RDA, CCO, DACS, and the IFLA statement of cataloging principles.
Metadata schema / vocabularies	Data structures and schema govern how information extracted from the resource is described and stored in a metadata object.
Data serialization	Data formats and serializations refer to the standards that are used to record the generated metadata and typically refer to some digital form of encoding.
Data exchange	Data exchange standards refer to a standard that governs the sharing of metadata between systems.

accessible in the GitHub repository. More information about how to access and make use of this data is available on the GitHub site.

GitHub LTR *repository*
https://github.com/mitcheet/ltr

Gephi
http://gephi.org

Throughout the remainder of this issue we will use these five building blocks of metadata—data model, content rules, metadata schema, data serialization, and data exchange—as a framework with which we can explore LOD as well as our case studies. In chapter 2, we focus on LOD and take a fairly technical look at how linked data works. In chapter 3, we explore the metadata aspects of the BIBFRAME initiative, the Europeana digital library, and the Digital Public Library of America using this framework and consider the similarities and differences among these services. In chapter 4, we take a broader view of the metadata questions and cross-domain issues we discussed in this chapter and explore where the LAM community is with these issues today. In doing so, we may find it difficult to maintain firm boundaries among these five components of metadata or between key issues and opportunities.

Notes

1. Karen Coyle, "Linked Data Tools: Connecting on the Web," *Library Technology Reports* 48, no. 4 (May/June 2012); Michael Witt, "Object Reuse and Exchange (OAI-ORE)," *Library Technology Reports* 46, no. 4 (May/June 2010); Marshall Breeding, "Opening Up Library Systems through Web Services and SOA: Hype, or Reality?" *Library Technology Reports* 45, no. 8 (November/December 2009); Andrew Nagy, "Analyzing the Next-Generation Catalog," *Library Technology Reports* 47, no. 7 (October 2011): 5–7.
2. Brad Eden, "Information Organization Future for Libraries," *Library Technology Reports* 43, no. 6 (November/December 2007).
3. Michele Seikel and Thomas Steele, "How MARC Has Changed: The History of the Format and Its Forthcoming Relationship to RDA," *Technical Services Quarterly* 28, no. 3 (2011): 322–334, doi:10.1080/07317131.2011.574519.

4. Library of Congress, "MARC Format Overview," MARC Standards website, Network Development and MARC Standards Office, retrieved April 29, 2013, www.loc.gov/marc/status.html.
5. Esther Green Bierbaum, "MARC in Museums: Applicability of the Revised Visual Materials Format," *Information Technology and Libraries* 9, no. 4 (December 1990): 291–299; Beth M. Russell and Robin L. Brandt Hutchison, "Official Publications at Texas A&M University: A Case Study in Cataloging Archival Material," *American Archivist* 63, no. 1 (Spring/Summer 2000): 175–184; Richard P. Smiraglia, ed., *Describing Archival Materials: The Use of the MARC AMC Format* (Binghamton, NY: Haworth Press, 1990).
6. Erin Coburn, Elisa Lanzi, Elizabeth O'Keefe, Regine Stein, and Ann Whiteside, "The Cataloging Cultural Objects Experience: Codifying Practice for the Cultural Heritage Community," *IFLA Journal* 36, no. 1 (March 2010): 16–29.
7. Steven J. Miller, "The One-to-One Principle : Challenges in Current Practice," in *DC-2010 International Conference on Dublin Core and Metadata Applications: Making Metadata Work Harder: Celebrating 15 Years of Dublin Core, ed. Diane Ileana Hillman and Michael Lauruhn* (Dublin, OH: Dublin Core Metadata Initiative, 2010), 150–164.
8. Nagy, "Analyzing the Next-Generation Catalog."
9. Roy Tennant, "MARC Must Die," *Library Journal* 127, no. 17 (October 15, 2002): 26–28.
10. Seikel and Steele, "How MARC Has Changed."
11. Eric Miller, Uche Ogbuji, Victoria Mueller, and Kathy MacDougall, *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services* (Washington, DC: Library of Congress, November 21, 2012), 7.
12. Mark A. Greene and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," *American Archivist* 68, no. 2 (Fall/Winter 2005): 208–263; Dennis Meissner and Mark A. Greene, "More Application While Less Appreciation: The Adopters and Antagonists of MPLP," *Journal of Archival Organization* 8, no. 3/4 (2010): 174–226.
13. IFLA, Statement of International Cataloguing Principles (London: IFLA, 2009), www.ifla.org/files/assets/cataloguing/icp/icp_2009-en.pdf.
14. Erik Mitchell and Carolyn McCallum, "Old Data, New Scheme: An Exploration of Metadata Migration Using Expert-Guided Computational Techniques," *Proceedings of the American Society for Information Science and Technology* 49, no. 1 (2012), doi:10.1002/meet.14504901091.
15. Karen Coyle and Diane Hillmann "Resource Description and Access (RDA): Cataloging Rules for the 20th Century," *D-Lib Magazine* 13, no. 1/2 (January/February 2007), doi:10.1045/january2007-coyle.
16. Michael Mitchell to AUTOCAT mailing list, "RDA and the Plague," April 10, 2013, http://listserv.syr.edu/archives/autocat.html.
17. Mary W. Elings and Günter Waibel, "Metadata for All: Descriptive Standards and Metadata Sharing across Libraries, Archives and Museums," *First Monday* 12, no. 3 (March 2007): 7–14, http://firstmonday.org/ojs/index.php/fm/issue/view/225.
18. Murtha Baca, ed., *Introduction to Metadata*, ver. 3.0 (Los Angeles: Getty Research Institute, 2008), www.getty.edu/research/conducting_research/standards/intrometadata.
19. Craig Willis, Jane Greenberg, and Hollie White, "Analysis and Synthesis of Metadata Goals for Scientific Data," *Journal of the American Society for Information Science and Technology* 63, no. 8 (August 2012): 1505–1520, doi:10.1002/asi.22683.