Chapter 4

Issues, Opportunities, and Trends in Metadata

Abstract

Chapter 4 of Library Technology Reports *(vol. 49, no. 5) "Library Linked Data: Research and Adoption" takes a broad view of the concepts explored in chapters 1–3 in considering current research and practice in the library metadata community, particularly in relation to the development of new LOD systems and the incorporation of existing metadata into those systems. In order to ground the exploration, this chapter uses as its data source the notes and transcripts from an April 2013 NISO meeting for the Bibliographic Roadmap Project. This data was analyzed using a qualitative content analysis approach with the goal of identifying metadata trends and themes as well as community attitudes and perspectives.*

Introduction

In the first three chapters of this issue, we explored the conceptual (chapter 1), technical (chapter 2), and practical (chapter 3) aspects of metadata in library and other cultural heritage and memory metadata. In chapter 4, we conclude our exploration of metadata trends by using another source of data to help inform our understanding of the issues, opportunities, and trends in this domain and, in doing so, seek to contextualize the frameworks, tools, and services we have focused on in this issue of *LTR*.

Our data source for this discussion includes eleven documents that resulted from an in-person and virtual meeting held in Baltimore, Maryland, in April 2013. The NISO Bibliographic Roadmap project is an Andrew W. Mellon–funded project that seeks to build a community-developed road map for a "bibliographic information exchange ecosystem."[1] The meeting in April focused on broad community value and direction topics, including questions about interoperability, content rules, user needs, provenance, scalability, and business models.

> *NISO Bibliographic Roadmap Development Project*
> www.niso.org/topics/tl/BibliographicRoadmap

In preparation for the meeting participants completed a survey to help generate topics and themes for discussion. These themes included ideas like user needs, business models, general goals, metadata interoperability, openness and sharing, rules, system prototyping, and metadata provenance and authority. During the meeting, breakout group discussions focused on these themes and recorded artifacts of their discussion that were subsequently published for review and editing.

> *NISO Bibliographic Roadmap Meeting: Documents*
> https://sites.google.com/site/nisobibrm

Data Analysis

The NISO Bibliographic Roadmap Project has a number of information products already available. These include detailed project proposals, recorded sessions and presentations from the April meeting, transcripts, and notes. In addition, there are early publications

Table 4.1
Document names and length of documents used for data analysis

Document name	Length (words)
Areas for discussion	2,296
Breakout Group - Business Models	570
Breakout Group - Goals	403
Breakout Group - Interoperability	1,220
Breakout Group - Open/share	383
Breakout Group - Provenance / Authority	223
Breakout Group - Prototyping	428
Breakout Group - Rules	334
Breakout Group - Users	93
Day 2 meeting discussion	4,421
Discussion from input survey	922
Other spare notes	530

reflecting on the process and outcome of the session.[2]

For the purpose of this chapter, only the transcripts and breakout group notes, available as text documents online, were analyzed. The content of these documents indicates that they were created by analysis and summarization and may not capture the nuanced activities and discussions in the breakout groups. Although these resources are secondary data resulting from the discussion groups and breakout sessions, they represent a useful dataset in that they include brief and highly distilled concepts and ideas.

The length and scope of the files varied considerably. The longest transcript ("Day 2 discussion") was over 4,420 words, while the shortest document, "Breakout Group—Users," was 93 words. The user theme, however, was well represented in other documents, indicating a difference in note-taking approaches rather than a lack of interest or discussion output. Table 4.1, sorted alphabetically by document title, shows the names and lengths of documents used from the NISO Google sites.

Given this understanding of this data source, quantitative analysis techniques are not likely to yield useful information but there is much value in using qualitative methods to scan, identify, and categorize the themes and trends discussed.

The documents were analyzed using a semiformal content analysis approach in which metadata evaluation models were used to suggest predefined codes, but new codes and relationships were also created throughout the coding process. The two predefined coding frameworks including the metadata building blocks model (table 1.2) and a taxonomy of metadata schema functions defined by Willis, Greenberg, and White.[3]

Table 4.2
Taxonomy of themes and discussion topics generated from content analysis

Issues	Adoption barriers Business models Cost Implementation Institutional responsibility Literacy issues Migration Open and contractual licensing Organizational issues Original cataloging in LOD Personal privacy Standards compliance Sustainability Technical issues Timeline for implementation Training (see literacy) User needs evaluation
Opportunities	Community collaboration Demonstration of value of libraries Innovation LAM collaboration New research methods Open data publishing Patron engagement
Impact	Adoption Community collaboration Community vision Organizational work Staffing
Metadata evaluation	Compatibility Consistency Data integrity / trust Data-centric evaluation Efficiency Metadata value Metrics-based evaluation Provenance / responsibility Quality assessment Sustainability Use cases User-centric evaluation
Metadata functions	Aggregation Computation Data publishing De-duplication Discovery Interoperability Mapping Metadata Lifecycle

We have sufficiently reviewed the metadata building blocks model to understand its constituent parts. The taxonomy of metadata schema objectives were derived by Willis, Greenberg, and White using a content analysis approach on nine scientific metadata schemas and consist of twenty-two codes covering a wide range of metadata features and functions, including

data validation, provenance, data archiving, and data documentation. These codes largely focus on technical operations such as data validation, scheme simplicity, scheme stability, data publication, and data interchange but also include more conceptual elements such as data life cycle, sufficiency, and abstraction. This taxonomy was designed to describe both the focus of a scheme (a scheme is designed to support data interchange) and to describe features within the scheme (the scheme features the ability to record provenance metadata).

Using these two taxonomies as sources for categories, codes were applied, created, and mapped to represent document content. The fact that the focus of the documents was largely on discussing issues at a high coding level resulted in five broad discussion-focused themes. The resulting metadata discussion framework included five top-level themes: Issues, Opportunities, Impact, Metadata Functions, and Metadata Evaluation. Table 4.2 shows these five broad themes and their child categories.

The emphasis in table 4.2 on the Issues and Metadata Evaluation categories indicates where areas of concentration were occurring in this set of breakout groups. This is a reasonable fit with the purpose of the meeting and may not be representative of the discussion occurring across library metadata communities, but the topics are certainly in sync with the issues identified in our review of literature in chapter 1.

Discussion

The initial review of the documents in this dataset and the generation of the coding framework shown in table 4.2 are only a first attempt at understanding the complex metadata issues. While further work is needed in developing and refining this framework, the parent and child topics identified in this initial review show both remarkable similarity to the issues discussed in chapter 1 and interesting questions regarding specialized but very important areas of interest. With this in mind, in the following five sections we take a more detailed look at some of these cross-discussion themes and consider both the issues mentioned in the NISO documents and how our case study services addressed these issues. These sections are based on cross-discussion themes that appear to pervade the analyzed NISO documents. The five themes are (1) metadata quality, (2) open data and business models, (3) literacy, training, and cross-community engagement, (4) aggregation, provenance, and trust, and (5) implementation, interoperability, and scale.

Metadata Quality

Metadata functionality questions and technical issues were a pervasive and threaded theme across the analyzed meeting minutes. Metadata functionality and quality questions centered on issues of conversion, operations, and use as well as questions of value and impact. There was a strong emphasis on library-related standards in this discussion, including BIBFRAME, RDA, ISBD, MARC, UNIMARC, and other standards, but there was also discussion of interoperability with museum and archival standards. The child themes of the Metadata Function category included historically important tasks such as discovery and interoperability as well as new tasks such as computation, aggregation, and mapping. In fact, throughout the breakout sessions and the Day 2 afternoon discussion documents, the theme of "mapping, not migration" recurred. This theme focused on questions of how libraries could move from current standards to a new bibliographic standard, and proponents of the "mapping" approach emphasized the idea that in a linked data context, metadata is not migrated between standards but rather is designed to naturally interoperate with other schemas using shared and compatible vocabularies.

Open Data and Business Models

Within the coded excerpts from our dataset, there were complex discussions occurring on open data, business model, and metadata quality themes. For example, the presence of "open data publishing" as both a concern and an opportunity reflects a general concern expressed in discussions in the Business Model breakout session and Areas for Discussion document regarding how libraries, consortia, information system vendors, and publishers would respond to calls for open data and metadata publishing.

In addition, the open data theme often coincided with questions about what new metadata systems might look like and whether they would be open source or commercial. For example, a key issue in this area was how newly designed specifications will accommodate inventory control as well as discovery and resource-sharing services. In our case study exploration, there seemed to be an emphasis on discovery and digital object management functionally over physical object management.

LAM institutions are addressing open data and the impact on their core business model in different ways. Harvard, for example, has released its bibliographic data as open data, and many libraries are publishing some collections and resources using Creative Commons licenses. At the same time, system vendors and cooperative organizations are seeking models in which the technology systems and metadata services that are an important part of their business model will change in LOD environments. It appeared from the NISO meeting that we are still too early in the exploration of LOD to fully understand the implications on current business models.

Harvard Library Open Metadata
http://openmetadata.lib.harvard.edu

Literacy, Training, and Cross-community Engagement

One of the common threads in the NISO discussion documents focused on issues of literacy and training, both from an individual and institutional perspective and from a community and partnership perspective. Training and standard literacy are valid concerns for librarians who are just now grappling with RDA implementation issues and are facing the need to learn even more detailed standards.

While the case studies we explored did not address these issues explicitly, the use of simple REST-based APIs and accessible metadata schemas and data serialization formats shows that there is real interest in making the data accessible to experts and nonexperts inside and outside of the library community. This focus is seeing developments in other projects as well, including LiAM, an Institute for Museum and Library Services (IMLS)–funded project at Tufts University that seeks to better understand the role and application of linked data in archival settings. One of the key products of project LiAM is a guidebook focused on the topic of Linked Data use in archives that includes a technical overview, service impact, organizational structuring, and next steps. According to the project documentation, the guidebook idea was inspired by other open publications, such as the Getty Research Institute's *Introduction to Metadata*.[4]

Tufts University—LiAM: Linked Archival Metadata
http://sites.tufts.edu/liam

In addition to literacy issues, cross-domain engagement was an important recurring theme. The interest in cross-domain work was evident in all three of our case studies, either by extensive inclusion of metadata or by incorporation of external standards and vocabularies. While our exploration focused on the library side of this work, museum and archival communities are engaged in similar standard definition and community outreach efforts. The CIDOC Conceptual Reference Model (CRM), for example, is a metadata specification turned towards museums that has been translated to support RDF data modeling but has interest in the library community. This community has a wealth of work that includes alignment with common library standards like FRBR, which has a CRM mapping called $FRBR_{OO}$,[5] and OAI-PMH via LIDO (Lightweight Information Describing Objects).

The CIDOC Conceptual Reference Model
www.cidoc-crm.org

Aggregation, Provenance, and Trust

In none of the three services that we explored did we look deeply at the process for metadata ingest, conversion, and synchronization. It was clear that Europeana had given this process considerable thought, however, and the complexity of its data model showed the importance of metadata tracking when working with multiple sources of data. This issue has not been extensively addressed in bibliographic specifications to date, and it was clear in the NISO documents reviewed that while provenance and trust are key issues, there is little consensus on how to ask, much less how to answer, these questions.

Issues of trust and provenance have become particularly important in the bibliographic community as libraries ponder moving from a "master record" to a "distributed linked data" approach to metadata creation and use. The Europeana Data Model (EDM) addresses this issue by employing the proxy feature of the ORE data model but also observes that the emerging named graphs (aka quads) would help address the provenance issue by providing a fourth resource pointer to include in a triple so that a statement could identify a subject, predicate, object, and associated graph. Quads are currently being discussed in a W3C working group. While similar structures were seen in the DPLA model, it would appear that this is an area of potential development for bibliographic-centric metadata specifications.

N-Quads: W3C Working Group Note
www.w3.org/TR/2013/NOTE-n-quads-20130409

Implementation, Interoperability, and Scale

Each of the services discussed in chapter 3 was at a different point in its community-building activities, and it was difficult to understand if those phases represented steps on a continuum or paths leading in different directions. Across the breakout sessions there were instances where concern was expressed regarding the time required to develop, implement, and adopt these new systems. Indeed, both the DPLA and the BIBFRAME initiatives have moved quickly in the last few years, and this speed seems to be both a motivating and a concerning factor for libraries looking at these new specifications.

In conjunction with issues of implementation and interoperability, there was some mention of

computational techniques, cloud computing, and web-scale services in general, but no concrete ideas for how these tools might help shorten the timeline required to implement a new specification or system once it has reached a deployable state. It is likely, however, that the lessons learned by LOD projects will prove useful here while also contributing to the critical mass of linked open metadata that appears to be the natural outcome of this process. At the same time, while libraries, archives, and museums share similar interests in this space, how these institutions have shared and aggregated data has differed historically and as such poses challenges for how we share data in the future.

For example, in addition to API availability in our reviewed services, there is also a growing number of SPARQL endpoints for LAM metadata available on the Web. As this list grows and as researcher literacy with the SPARQL format develops, new opportunities exist to help researchers discover new information by constructing new graphs of information through search of multiple linked data repositories together. This work requires both technical and information literacy efforts, but at the same time it provides an avenue for interested LAM institutions to join an active research community without having to transform all of their underlying systems. While running a SPARQL endpoint and converting and loading data into that endpoint is not a trivial process, open-source tools like Apache Jena and Virtuoso are maturing and becoming easier to implement and use.

W3C: SparqlEndpoints
www.w3.org/wiki/SparqlEndpoints

Apache Jena
http://jena.apache.org

Virtuoso
http://virtuoso.openlinksw.com

Conclusion

It is likely that we are still too early in our community's work with LOD and LOV to understand the potential impact of these specifications on our metadata systems as well as our organizations. In recent years, libraries have focused on cloud applications, and as a community we have discussed the merits and problems associated with moving our technology to the cloud. Linked data poses a new "metadata web" model in which LAM institutions no longer maintain central repositories of metadata but rather work to ensure that their local metadata is properly linked and connected with others'. This is a new type of outsourcing but shares many of the same outcomes, efficiencies, and economies of scale for nondifferentiating services and the preserved ability to create and publish metadata services that highlight the distinguishing collections and services of an institution.

This approach to metadata publishing and management may mean that LAM institutions need to further separate inventory control work and resource description work and reinvent the systems they use. While as a community it appears that we are not yet sure what these new systems will look like, it is clear that if LOD is to take off, libraries and cultural heritage institutions need to find ways to publish their data as LOD in environments that support SPARQL querying. These systems are at the foundation of research needs for users seeking to discover resources across institutional silos and discover new knowledge through computational reasoning.

These changes in technology, in metadata modeling, and metadata serialization, may mean that in addition to being in a "post-MARC" era, LAM institutions are entering an era where XML and traditional record-based formats are being supplanted with other serialization methods and data models. The use of SPARQL query structures, N3 and Turtle RDF serializations, JSON serializations, and other RDF-inspired data models shows that XML files and certainly record-based and flat-text file data models are no longer the cutting edge of metadata technology. This shift is important because it enables libraries to better represent their metadata but also because it helps patrons leverage computers to make more efficient and detailed use of published data. For example, a key advantage of JSON serializations is that the data can be used programmatically without ingest and transformation work and RDF modeled data can be directly imported into a new breed of Semantic Web data analysis tools.

While these issues of literacy, cost, value, openness, user needs, and technology continue to be important questions to ask in the metadata world, there appears to be growing consensus about what the next generation of metadata will be, not only for libraries but for a wide range of cultural heritage and memory institutions. One of the enduring values of the Web that made it central to how people engage with information is the notion that information in the digital world is not bounded because of scale, authority, or cost because the efficiencies, communities, and economics of the Web changed how people engaged with and valued information. As LAM institutions endeavor to make the metadata they have created and curated a fully functional part of this web of information, we are likely to see similar shifts in how we think about these issues as well.

In this issue of *LTR* we explored current research and practice with metadata in library and other

information settings. In chapter 1, we synthesized a five-element framework to identify the building blocks of metadata and to help us understand where the focus of research and practice is in metadata. In chapter 2, we used this framework to explore the concept of linked data by looking at its components (e.g., RDF, RDFS, OWL, RDF/XML, SPARQL). In chapter 3, this understanding of linked data technologies and models helped us use a case study approach to consider three cutting-edge metadata systems. This analysis yielded interesting similarities and differences among the systems and shed light on a potential development path for LOD/LOV systems in libraries, archives, and museums. This included a technical development path ([1] Define an LOD model; [2] aggregate data; [3] publish data for user and computational access; [4] enhance LOD endpoint integration; and [5] disseminate via SPARQL endpoints.), as well as common community-building and data-licensing issues that need to be addressed throughout the process. With an understanding of LOD project activities, we turned our attention to the communities surrounding them as well as the broad research and implementation goals of these communities. In chapter 4, we explored community goals and engagement by performing qualitative analysis of the minutes of a recent NISO meeting on bibliographic metadata. Our exploration touched on a number of data-modeling and technical design issues but also revealed some enduring themes such as metadata quality, interoperability, and use. The analysis of these themes yielded five main research areas: (1) Metadata quality, (2) Open data and business models, (3) Literacy, training, and cross-community engagement, (4) Aggregation, provenance, and trust, and (5) Implementation, interoperability, and scale.

These research areas cut across many of the development and implementation activities identified in chapters 2 and 3. For example, the emergence of graph-based data models and the increasing use of non-XML serialization technologies are posing new challenges in technical and data-modeling activities that impact metadata quality, open data publishing, and data aggregation areas. In addition, as new systems and tools are developed, it is important to think critically about their interoperability, implementation, and scale.

The transition from our current metadata models and records to new models, for example, will require computational techniques that will enable large-scale and highly reliable transformation to these new models. The challenges in this space are organizational as well as technical, but an important first step is in generating LOD and LOV endpoints at scale while also finding ways for individual libraries to implement their own LOD stores and map their current records to these endpoints accurately and efficiently.

Libraries are fortunate to have the fields of computer science and computational linguistics as sources of methods and algorithms that will support large-scale text analysis and reconciliation problems. At the same time, however, LAM institutions as a whole are finding that the licensing issues associated with publishing metadata and digital collections under open licenses are more challenging. This is clearly a manageable issue, as our exploration of LOD environments found a number of systems that had addressed these problems, but challenges still remain in making the full text of resources available for research and in making all of the records in the bibliographic universe available.

The threaded question "How will we accomplish this?" was found throughout the NISO documents and is a common theme in discussions within the LAM community. In order to answer this question, user needs analysis, training, and community outreach are important research areas for metadata work that need to be explored alongside the technical, organizational, and licensing questions. These issues are not of interest just to information professionals but also to our communities of users. The transformation of bibliographic and digital collection metadata to LOD and LOV environments, for example, opens up new opportunities for researchers to work with collections and metadata using computational and cross-repository techniques. Engaging in and supporting this type of research will require understanding how to support complex data querying, collection, and visualization and will be built on a distributed infrastructure of LOD repositories.

Advances in data modeling, technology and information design, user interaction, visualization, literacy, community engagement, data licensing, and open publishing are the mechanisms by which high-level questions about metadata value, community information needs, and institutional impact are being considered. These high-level questions are grounded in our professional values of equity, information freedom, and public service. The questions are also grounded in our regard for the foundational role information institutions play in society. Given this relationship, how the library, archives, and museum communities address the detailed technical and conceptual questions of linked open data will shape the evolution of our profession and knowledge institutions.

Tools and Data Used in This Issue

This issue of *LTR* mentioned a number of metadata tools and application programming interfaces (APIs), as well as a number of figures that were generated using graph visualization software. Rather than documenting each of these tools and including code in appendixes, information about these resources, as well

as sample RDF/XML files used in generating the figures, are available in a GitHub repository. More information about how to access and use these resources can be found on the site.

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

Notes

1. "NISO Bibliographic Roadmap Development Project," accessed May 16, 2013, www.niso.org/topics/tl/BibliographicRoadmap.
2. Todd Carpenter, "Bibliographic Roadmap Proposal," January 13, 2013, www.niso.org/apps/group_public/document.php?document_id=9975&wg_abbrev=ccm; NISO Bibliographic Roadmap Meeting, outline for meeting, April 15–16, 2013, Baltimore, MD, https://sites.google.com/site/nisobibrm/; Roy Tennant, "The Post-MARC Era, Part 1: If It's Televised, It Can't Be the Revolution," The Digital Shift, Library Journal, April 17, 2013, www.thedigitalshift.com/2013/04/roy-tennant-digital-libraries/the-post-marc-era-part-1.
3. 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.
4. Murtha Baca, ed., *Introduction to Metadata*, ver. 3.0 (Los Angeles: Getty Research Institute, 2008), www.getty.edu/research/conducting_research/standards/intrometadata.
5. Patrick Le Boeuf, "A Strange Model Named FRBROO," in "The FRBR Family of Models," special issue, *Cataloging and Classification Quarterly* 50, no. 5–7 (2012): 422–438, doi:10.1080/01639374.2012.679222.