Chapter 4

The Evolving Direction of LD Research and Practice

Emerging Issues in LD

In chapter 1, the author posed the question, "How should LAM institutions balance the need for generalized LD models that encourage interoperability with external community members against the need for highly granular internally focused standards?" This question is one example of a continuing discussion in the LAM community that exemplifies the current state of adoption of LD. Although the community as a whole is moving in the same direction, many paths are being taken, and clearly not all of these paths will arrive at the same place.

At the same time, while there are still many technical issues to resolve in LD adoption, the LAM community has made considerable progress in the past two years in building proof-of-concept tools, production vocabularies, and LOD-enabled services that demonstrate how data can be transformative in supporting information services rather than simply being useful. In chapters 2 and 3, this issue examined projects, tools, services, and vocabularies in more detail. The tools, vocabularies, and programs reviewed in these chapters are being informed by philosophical perspectives in the LAM community, including the value of data openness, the importance of standards and approaches for defining and maintaining standards, and approaches to system development.

Data Openness

The common practice that LAM communities forged around open-source development and licensing is now influencing how we approach making data open. In fact, while LAM institutions are choosing different open-use licenses, there is much shared practice around the creation and dissemination of open data. There are exceptions, however, as many libraries have bibliographic data from outside suppliers without having the ability to make that data available to their users under an open license. Likewise, some institutions have data policy rules that make publishing data as open data difficult. One such policy issue is often the ability to allow others to make commercial use of published data. Perhaps a much larger issue, however, is the fact that libraries are creating less metadata than they used to and are licensing much more of it from outside suppliers, meaning that the ability to drive the discussion around open metadata is being limited. This is a simple reality given the shift of information institutions to the web and the widespread licensing of electronic content. In fact, metadata generation in general is an area that requires serious consideration as information institutions and the information communities that serve them ask questions about how to afford to create metadata for the newly published information objects.

The overall lack of data openness and transparency is an influential factor in the library discovery service market. Although there is an open discovery initiative led by NISO, there is no real momentum yet behind the notion that LAM institutions should be able to make this data openly available or that data can be separated from the discovery systems that provide access to it. This creates an unfortunate circumstance in which libraries in particular are purchasing metadata multiple times and in multiple information systems. At the same time, libraries are seeking out cloud providers to make use of and manage this new metadata and must find viable commercial models to ensure that system producers are incentivized to provide the desired services. It is entirely feasible that LAM institutions should consider opting out of licensed metadata and select publishers and vendors

that produce metadata in a consistent format for open use. In fact, many publishers already build metadata for the web and are directing users to their own discovery portals, often with the purpose of selling access to licensed content that may be available through a user's institutional affiliation. This practice is having a considerable negative impact on communities of valid users whose use of the web to find resources is not supported with the systems and services required to allow them to make use of the license fees their institutions have paid.

In a legal context, the 2014 court decisions regarding uses of digitized book data by HathiTrust and Google indicated that nonconsumptive use of digitized full text falls under fair use.[1] These decisions support the efforts of LAM institutions to make new uses of copyrighted and noncopyrighted resources in new ways, with a particular emphasis on using contextualized data to support discovery and research. The related discussion about whether or not metadata is copyrightable is an important one in the LD community.[2] The DPLA took a stand in 2013 that "the vast amount of metadata is not copyrightable."[3] Such a stance is appealing in LD circles as it simplifies or removes issues associated with reusing data and making your own data available.

While many LAM institutions are turning to Creative Commons (CC) licenses that support reuse with or without attribution, reuse by commercial or noncommercial entities, and derivative or original form use only, there is no true consensus on how to ensure that data licenses are consistent and easy to apply in an automated fashion. For example, while many libraries use CC, OCLC makes use of the Open Data Commons (ODC) licenses. The ODC makes three licenses available, an "attribution" license (ODC-By), a public domain license (PDDL), and an "attribution and Share-Alike" license (ODbL). The key difference between ODC-By and ODbL is that the "Share-Alike" license allows you to adapt a dataset and rerelease it as long as you use the same license. In fact, some suggest that metadata should in fact be in the public domain and not made available via a data license, the key impact being that data licenses are in themselves restrictive and can lead to improper attribution.[4]

Open Data Commons
http://opendatacommons.org

Standards Compatibility

A similarly large issue related to LD is the issue of standards adoption and cross-community compatibility. As LAM projects are moving forward, the organizations are making highly impactful decisions about vocabularies to use, required granularity of selected approaches, and potential reuse purposes of published datasets. Without widespread agreement over how these vocabularies exchange standards should operate, LAM institutions may find themselves in a difficult-to-navigate mixed-metadata world. One such confusing area that has arisen in the past few years is the use of the BIBFRAME Lite name by Zepheria to represent an alternative to the BIBFRAME vocabulary. The reuse of the name is introducing some confusion into an already complex discussion around related standards.

Although there is yet to be a singular approach around metadata schemas, more consensus is emerging around serialization of LD. While LAM institutions are using a range of serialization standards, including RDFa, RDF/XML, Turtle, N-Triples, and JSON-LD (i.e., the predominating serialization formats for LAM LD), the stability of the RDF data model across these serialization standards as well as the growth in transformation tools, has meant that this is not as complex an issue as one might think. In fact, in the past two years, JSON-LD has grown as a standard that is more robust and appears to have a preference among the LD LAM developer community, even though it is not as granular as RDF/XML. The inclusion of JSON-LD in the RDF 1.1 specification was a signal that the issues with specificity and granularity in this serialization have largely been addressed.

Lack of Supporting Systems

It is fair to say that that LOD LAM applications are still in a "roll your own" phase of development. LAM institutions that seek to deploy LD applications are often exploring technical platforms and making localized decisions about the best systems to select. While systems do not need to be identical—in fact, it is advantageous for them to not be identical—the fact that LAM institutions are still having to select triplestores, SPARQL engines, indexing platforms, and other services means that there is still a relatively high bar for institutions to cross in taking up LD projects. A later section for this chapter explores some of the systems in use in common projects and seeks to identify some selected systems that appear to be bringing the various LD publishing tools together (e.g., triplestore, SPARQL endpoint, index, discovery interface, and creation interface).

Another area of system development that is also very much in focus is the extent of vendor support for LD applications. Library system vendors have taken different approaches over the past two years in developing the next generation of information systems. At ALA 2015, many ILS vendors expressed support

for BIBFRAME and spoke to broad roadmaps around adoption. Chapter 2 explored how some research projects focused on transforming bibliographic data are making use of existing systems, particularly open-source platforms. At the same time, there does not yet appear to be a comprehensive turnkey solution for libraries seeking to create and publish LD. On the systems front, it appears that more progress has been made in the archival and museum communities. Similar challenges still exist in these communities, although the information systems they use, such as ArchivesSpace, CollectionSpace, Fedora, DSpace, and other related tools, are already aligned around metadata standards that can be easily converted to LD for publication.

Whether or not getting to the turnkey level is necessary to see LOD adoption grow is a fair question, but it is clear that libraries are investing in LOD as a way to drive down costs as well as increase value. It is not feasible or sustainable for LOD systems to ultimately cost more than their current metadata publishing counterparts (e.g., Integrated Library Systems, Digital Asset Management systems), but it is likely that this is the reality for early adopters who need to invest in both traditional and new LD systems simultaneously.

Important Questions in the Linked Data Community

How Have Standards Evolved over the Last Two Years?

One of the key difficulties in creating LD and making it available is in defining the use cases that make sense and will have value to the community. Publishing data in some serialization of RDF is not especially useful or interesting if it does not capitalize on links to other datasets or provide new opportunities for computational analysis of data. As the LD community has grown through experimentation and project developments in the past few years, more best-case examples of how to create and publish metadata have been explored and reported. Perhaps the clearest expression of these principles is in a working group report titled "Best Practices for Publishing Linked Data." This guide surfaces ten steps for publishing Linked Data, reproduced in the list below.

1. Prepare Stakeholders
2. Select a Dataset
3. Model the Data
4. Specify an Appropriate License
5. The Role of "Good URIs" for Linked Data
6. Standard Vocabularies
7. Convert Data to Linked Data
8. Provide Machine Access to Data
9. Announce to the Public
10. Social Contract of a Linked Data Publisher.[5]

Of these ten steps, five of them focus on policy and social-good issues rather than purely technical issues or topics. This document, as well as many others, cites Hyland and Wood's work on creating Linked Data from a technical perspective,[6] and as a result, a more policy-focused document is a useful and somewhat unique contribution to the Linked Data publishing space. Although the author will not replicate the core recommendations of the document here, many key items are worth highlighting—the need for documentation using self-descriptive techniques, the importance of persistence, and the importance of supporting multiple languages.

What Are Communities of Practice Saying about the Direction of Linked Data, and How Have the Issues around LD LAM Changed in the Past Two Years?

Overall, the focus of the library community remains on a conversion to LD, and we have seen considerable development efforts to make the conversion of data as well as the creation of new data possible. In fact, new projects, such as LD4L and BIBFLOW, point to future potential production systems that may advance LD work. At the same time, libraries are challenged to demonstrate impact and prove that they have capacity. In a summer workshop held at UC Berkeley, for example, the common discussions around capacity building paralleled discussions around innovation and new projects. It is clear from the state of the projects that libraries undertaking LD efforts now must be prepared to continually convert data and to reconvert data to capitalize on new areas of development and granularity.

The state of adoption across libraries of all sizes remains limited although the tools are becoming more available and metadata standards are becoming more resolved and manageable. Whether or not simpler systems are the correct next step remains to be seen, but after several years of development it appears to be a necessary step.

With these forward steps, particularly via projects led by OCLC and the Library of Congress and grant-funded initiatives, the LAM community is pointing toward a robust future for LD. At the same time, it is also worth remembering that the community as a whole has yet to see transformative impacts from LD generation that resonate for all organization types and sizes. The goals of web visibility, research reuse, and granular preservation remain important, and it is clear that LAM institutions are driving their systems toward these purposes. Whether or not that will have a real impact in the research community remains to be seen.

What Role Do We Expect Large-Scale Projects to Play in Linked Data?

This is a difficult question to answer given the grassroots approach to LD projects in the LAM community at the moment. Traditionally, central players in the LD space, including LoC, OCLC, and NLM, are being complemented by players such as Europeana, the British Library, and multi-institutional cooperatives such as LD4L. A foundational discussion that is occurring among these groups centers on community alignment—especially how LAM institutions can make their data align with other communities of practice. OCLC, for example, has recently begun exploring the notion of a "Knowledge Vault" for libraries, a concept built on Google's work in knowledge graphs.[7] Likewise, companies such as Zepheria and its LibHub initiative continue to have a strong influence on the direction of the community, and there are a number of examples of secondary uses of metadata to create field trips in Google's mobile field trip tool to support visualization services on top of DPLA harvested data and to publish new vocabularies that aim to turn LAM data into LD.

Google: Customizing Your Knowledge Graph
https://developers.google.com/structured-data/customize/overview

Field Trip
https://www.fieldtripper.com

It should not be surprising that as organizations like DPLA and Europeana develop, that issues of sustainability and governance become important. The fact that both of these organizations included these issues in their strategic plans indicates how interesting it is timing-wise and how pressing the topic of the value of these organizations is for LAM institutions in their related countries. In fact, one of the key issues surrounding efforts of LD publishing is how to ensure that the LD that is published remains available via the published URIs over time.

The Europeana-proposed funding model is interesting in its detailed exploration of customer groups and benefit analysis.[8] The groups include end users, cultural institutions and their associated member states, project funders, and creative industries. The projected cost of Europeana during the next three years is anticipated to be €10 million annually, or approximately $10.8 million (US). While this is not an insurmountable funding challenge, gathering this level of funding for other national initiatives will likely be a focus in the coming years.

How Will LD Influence Cataloger Work and Notions of Value Moving Forward?

Seeman and Goddard explored the pressing question "what now" in relation to guiding catalogers in the creation of metadata as these LD standards are evolving.[9] Observing that much of the core work of cataloging (e.g., authority control, access point assignment, disambiguation) remains philosophically, if not functionally, the same, they suggested that this work, taken along with commonsense approaches, may make capacity for forward progress. It goes without saying that in a community driven by process and standards, the long-term discussions around a set of emerging but fluid standards without action does not serve the community well.

In fact, the LAM community as a whole has yet to tackle the true early adopter problem. Given the high level of collaboration and interoperability developed throughout the preceding century among libraries in the sharing of metadata and cooperative resource sharing, it may be that there is recognition that the stakes for early adopters are high. One such technique that is being suggested is embedding URIs in traditional MARC records. Interestingly, this notion was discussed in a 2010 LoC brief.[10]

A question related to value is whether or not LAM metadata, when transformed into LD, becomes something more than it was as unlinked metadata. Does the creation of LD, for example, make the metadata a "first class" research object? Does the publishing of LD create new streams of research or support new research methods? The fact that some institutions are publishing datasets in a more complete form points to the idea that this is possible, yet LAM metadata has typically focused on resource description and object management, areas of information that do not necessarily lend themselves to expansive research questions.

Current Education Opportunities

Challenges around bringing library staff up to speed on new approaches in metadata creation and management continue to impact the community. Some institutions have reported using the Juice Academy series, particularly the XML program. In addition, the Educational Curriculum for the Usage of Linked Data (EUCLID) project publishes a comprehensive textbook focused on Linked Data creation and use. In fact, this issue is as pressing for LIS schools as it is for practicing professionals. As a result, there is likely to be more restructuring of LIS curricula in the coming years as traditional work in resource description shifts and new concepts and skills are needed to work with LD technologies.

Library Juice Academy
http://libraryjuiceacademy.com

EUCLID
http://euclid-project.eu

Conclusion

This issue has explored current practice and emerging trends in LD LAM projects and activities and has considered some of the broad questions and topics of future exploration. In doing research for this issue, the author found that in the past two years considerable research and publication had occurred documenting specific technical projects, applications, vocabularies, and community best practices. In fact, the amount of literature and activity in this area is large enough to defy concise analysis. If anything, the exploration of trends, projects, and topics indicates that while the LAM community may be moving in a common direction, we are doing so in a number of parallel, if not identical, paths.

Notes

1. Ian Chant, "Appeals Court Upholds Wins for Fair Use in HathiTrust Case," *Library Journal*, June 12, 2014, http://lj.libraryjournal.com/2014/06/litigation/appeals-court-upholds-wins-for-fair-use-in-hathitrust-case.
2. Karen Coyle, "Metadata and Copyright: Peer to Peer Review," *Library Journal*, February 28, 2013, http://lj.libraryjournal.com/2013/02/opinion/peer-to-peer-review/metadata-and-copyright-peer-to-peer-review.
3. Ibid.
4. Timothy Vollmer, "Library Catalog Metadata: Open Licensing or Public Domain," Creative Commons, August 14, 2012, http://creativecommons.org/tag/open-data-commons-attribution-license.
5. Bernadette Hyland, Ghislain Atemezing, and Boris Villazón-Terrazas, "Best Practices for Publishing Linked Data," W3C Working Group Note, January 9, 2014, www.w3.org/TR/ld-bp.
6. Bernadette Hyland and David Wood, "The Joy of Data: A Cookbook for Publishing Linked Government Data on the Web," In *Linking Government Data*, ed. David Wood (Cham, Switzerland: Springer International Publishing, 2011), 3–26.
7. Merrilee Proffitt, Bruce Washburn, Diane Vizine-Goetz, and Roy Tennant, "OCLC Research Update" (presentation at ALA Annual Conference and Exhibition, San Francisco, CA, June 25–30, 2015), www.slideshare.net/oclcr/oclc-research-update-ala-annual-2015?from_action=save.
8. "Europeana Strategy 2020: Network & Sustainability (Draft)," Europeana, May 30, 2014, http://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana Strategy Network Sustainability.pdf.
9. Dean Seeman and Lisa Goddard, "Preparing the Way: Creating Future Compatible Cataloging Data in a Transitional Environment," *Cataloging & Classification Quarterly* 53, no. 3/4 (2015): 331–40, http://dx.doi.org/10.1080/01639374.2014.946573.
10. RDA/MARC Working Group, "Encoding URIs for Controlled Values in MARC Records," MARC Discussion Paper No. 2010-DP02, MARC Standards, Library of Congress, December 14, 2009, www.loc.gov/marc/marbi/2010/2010-dp02.html.