Chapter 3

Applied Systems, Vocabularies, and Standards

In the previous *LTR* issue on LD (July 2013), one of the compelling comments in the NISO community forum indicated that the important work in metadata and LD should focus on "mapping not migration."[1] The notion that the future of bibliographic or other types of metadata would involve the ability to round-trip metadata rather than a wholesale adoption of Linked Data models and vocabularies is not entirely in sync with some of the efforts we have seen in the review of projects that have taken shape since 2013. The research in the field around LD for the past two years has focused largely on surveys of adoption and specific technical works focused on defining best practices and proof-of-concept services. As the exploration of example projects and research initiatives in chapter 2 indicates, the LD LAM community is reaching a level of maturity that may be shaping next steps in LD adoption toward production systems and permanent migration.

Chapter 3 explores trends around specific tools, vocabularies, systems, and approaches employed by the projects mentioned in chapter 2. While the space allotted limits this section to providing pointer and brief descriptive information, the chapter seeks to provide references to literature and project approaches that may provide sufficient detail for organizations seeking to get started in their own LD projects. Readers seeking a more in-depth understanding of how to approach Linked Data projects would be well served by spending time with one of the growing sets of implementation guides. These include *Linked Data* by Wood, Zaidman, and Ruth and "The Joy of Data" by Hyland and Wood, in addition to a range of other resources.[2]

As explored in chapter 1, the growing list of LD adopters is laying important groundwork for those taking on LD creation next by developing tools and approaches as well as establishing more robust vocabularies to draw on. In chapter 3, we explore a few representative vocabularies and some tools that are increasingly used in LD projects.

Vocabularies and Schemas

The LAM community has largely centered on RDF and RDFS as a main representation data model for LD but varies in its choice of serializations (e.g., RDF/XML, RDFa, JSON-LD, Turtle). RDF/XML remains popular, but N-Triples, Turtle, RDFa, and especially JSON-LD are growing in popularity. New serialization standards, such as Versa, continue to emerge but do not appear to have widespread adoption. JSON-LD's increasing use in the LD community is notable in part because of its lightweight syntax but also because of its ease of use in programming languages. In fact, over the past two years, more programming languages have built libraries to make use of JSON-LD, and a more robust vocabulary has been developed within the standard to support lossless encoding of RDF. More information on JSON-LD is available at the JSON for Linking Data website, including a demonstration site. One common application of JSON-LD is to use the data in a framework such as AngularJS, a JavaScript-based development framework primarily oriented at using HTML to express web applications. AngularJS has been used by the British Museum, for example, to deploy a SPARQL search demonstration.

JSON for Linking Data
http://json-ld.org

British Museum AngularJS SPARQL Demo
http://collection.britishmuseum.org/angularsparqldemo/#

As more projects advance around LD standards, there are a growing number of vocabulary-aware tools built into common scripting languages that are lowering barriers to adoption. Python includes libraries like RDFLib, a library for working with RDF, and Django-RDF, a Django-based RDF engine. Other tools include html5lib, an HTML library for publishing data; Apache Jena and Fuseki, an in-memory database for processing RDF; and Callimachus, a Linked Data management system or an application server for Linked Data.

Django-RDF
https://code.google.com/p/django-rdf

Html5lib
https://github.com/html5lib

Apache Jena and Fuseki
http://jena.apache.org/index.html

Callimachus
http://callimachusproject.org

In addition to RDF, common organizing vocabularies include RDFS, OWL, and SKOS, within which FOAF, GeoNames, Dublin Core, and MODS are vocabularies commonly implemented. In several cases, these vocabularies are implemented in more comprehensive Semantic Web services such as sameAs.org, a service to support disambiguation and URI identification of data; DataHub, a site for publishing datasets; and DBpedia, a Linked Data platform for *Wikipedia* data. Another popular source for discovering datasets is Wikidata, an LD platform for collecting structured data that is also used in other Wikimedia projects.

sameAs.org
http://sameas.org

DataHub
http://thedatahub.org

DBpedia
http://dbpedia.org

DataHub: Datasets
http://datahub.io/dataset

Wikidata
https://www.wikidata.org

Of all of the vocabularies that are of interest to the LAM community, BIBFRAME and BIBFRAME Lite are certainly among the most discussed. The BIBFRAME Lite vocabulary is available online and includes four base terms: Work, Instance, Authority, and Event. These terms mirror those in BIBFRAME but do not entirely overlap with BIBFRAME vocabulary meanings. The BIBFRAME Lite site includes interoperability maps showing the overlap and interoperability with other LD schemas, including Schema.org and BIBFRAME. The author found, in his research about the status of LD adoption and services, that there is a wealth of resources that document the structure and application of these vocabularies. As a result, this issue of *LTR* does not attempt to replicate this information.

BIBFRAME Lite vocabulary
http://bibfra.me/view/lite

A vocabulary that is becoming more common in the LAM community is BiblioGraph.net, an extension to Schema.org designed to add bibliographic-specific content to Schema.org. As the Schema.org vocabulary matures, it is developing methods for representing videos and music in ways that allow computers to embed the media in web pages as well as capturing and promoting events. Such new structured data elements in the Schema.org vocabulary pose opportunities for LAM institutions to embed not only descriptive metadata centered on resources but also actual media and activity information in their sites. Another vocabulary related to Schema.org practices is called GoodRelations. GoodRelations provides a semantic structure for dealing with product data, sales locations, and other commercially focused concepts.

BiblioGraph.net
http://bibliograph.net/docs/bgn_releases.html

Schema.org: TV and Movie Watch Actions
https://developers.google.com/structured-data/actions/watch-movies

Schema.org: Event Markup
https://developers.google.com/structured-data/events/venues

GoodRelations wiki
http://wiki.goodrelations-vocabulary.org

In the cultural heritage community, a more established cultural heritage vocabulary, Lightweight Information Describing Objects (LIDO), has seen many adopters. Tsalapati as well as Van Keer, for example, studied the migration of LIDO using the CIDOC CRM model.[3] The CRM model is a conceptual model that defines semantic relationships for cultural heritage resources. CIDOC continues to enjoy adoption across a range of communities. The FRBRoo model represents FRBR relationships using the CRM model. Likewise, PRESSoo extends FRBRoo for serials and other continuations.

LIDO: XML Schema for Contributing Content to Cultural Heritage Repositories
www.lido-schema.org/schema/v1.0/lido-v1.0-schema-listing.html

CIDOC: FRBRoo Introduction
www.cidoc-crm.org/frbr_inro.html

PRESSoo
www.issn.org/the-centre-and-the-network/our-partners-and-projects/pressoo

Portland Common Data Model

A commonly mentioned schema around LAM applications of LD is the emerging Portland Common Data Model (PCDM). The PCDM is growing out of the digital asset management system (DAMS) community in particular to serve Hydra-based systems but with a focus on supporting other RDF and Fedora-based services as well. PCDM is primarily focused on structural and administrative metadata and includes provisions for access control. As with many current data models, PCDM draws heavily on Dublin Core, RDF, FOAF, Internet Assigned Numbers Authority (IANA), and other related vocabularies. At its core, PCDM implements collections and objects that are subclasses of Object Reuse and Exchange (ORE) vocabularies. The PCDM also includes an access control notion that provides a granular rights-granting platform that includes read, write, append, and control methods. The PCDM is under development and is envisioned as an important part of the Fedora 4 deployment in the LAM community. More developments are expected in this area.

Portland Common Data Model
https://github.com/duraspace/pcdm/wiki

Linked Data Platform 1.0 Specification

The Linked Data Platform (LDP) 1.0 specification, released in December of 2014, defines a standardized method of interaction for LD applications. The LDP refers to resources that have relationships via containers and that can be manipulated through web standard behaviors (e.g., get, post, put, patch, delete, options head) and returns data in a prescribed way using Turtle and JSON-LD.[4] The LDP specification is published as a working group recommendation at this point, meaning that it is not yet endorsed as a specification by the W3C. The goal of LDP is to define a standard set of application behaviors and response formats. This would be a useful next step in standardizing LD applications. In addition, the fact that the LDP standard focuses on tracking direct and indirect relationships between resources and containers of resources means that the data model that it employs may be a good fit for LAM institutions seeking to create LD applications. Fedora 4 has adopted the LDP model with these goals in mind and uses the LDP specification to inform its implementation of create, read, update, and delete (CRUD) functions.[5]

FRBR Library Reference Model

The Functional Requirements for Bibliographic Records (FRBR) model has been in development and discussion since the 1990s, with Functional Requirements for Authority Data (FRAD) and Functional Requirements for Subject Authority Data (FRSAD) having been defined more recently. The IFLA FRBR working group has recently undertaken the consolidation of these three models to create the FRBR Library Reference Model (FRBR-LRM). This model incorporates authority and subject authority relationships without modifying the core works, expressions, manifestations, and items (WEMI) model that has guided FRBR. In combining the models, the user task Explore is drawn in from FRSAD but is also expanded to include the FRAD task Conceptualize.[6] Although this model is in early draft form and slated to be reviewed in 2016, it is worth noting that IFLA as well as other organizations are exploring how to manage the WEMI and other FRBResque relationships that are at the core of many of the LD-focused user tasks that the LAM community imagines will be impactful.

Linked Data Services

The building blocks of Linked Data platforms commonly employ an ingest and reconciliation service, a data storage platform, a SPARQL endpoint, and, in many cases, some sort of more user-focused discovery platform. The Yale Center for British Art, for example, harvests data using OAI-PMH using LIDO, indexes data using Apache Solr, provides data via an API service, and supports discovery and interaction through VuFind, websites, and other application plugins.[7] In contrast, the British Museum collection relies on a unified platform called OntoText to provide indexing and SPARQL services. OntoText provides a service called Self-Service Semantic Suite (S4), which provides a set of semantic and text analysis tools that stores output in an RDF graph database running as a database-as-a-service. S4 integrates with other knowledge graph platforms such as GeoNames, DBpedia, and Freebase.[8]

The survey of LD vocabularies in use from the systems and projects reviewed surfaced a wide range of vocabularies for LAM and other applications. As with the survey of projects and systems, the vocabularies and tools in use are too numerous to catalog comprehensively. Many of the sources used for this issue, including the OCLC survey results; websites including Linked Data and Schema.org; the BIBFRAME implementation register; the Linked Data incubator group; and research articles cited in this issue are good sources for exploring the vocabularies in use in the LAM LD community.

OCLC survey results
www.oclc.org/content/dam/research/activities/linkeddata/oclc-research-linked-data-implementers-survey-2014.xlsx

Linked Data
http://linkeddata.org

Schema.org
http://schema.org

BIBFRAME implementation register
www.loc.gov/bibframe/implementation/register.html

Linked Data incubator group
www.w3.org/2005/Incubator/lld

Tools and Systems

There has been considerable growth in available tools to convert metadata to LD, in systems to serve LD, and in applications to query LD over the last few years. Tools already well known in the LAM community, including MarcEdit, OpenRefine, and RIMMF3, all provide LD-related editing functions. SPARQL command-line tools such as ARQ are increasingly common in the literature, and there is a wide range of triplestores available to store RDF data. For interested readers, two good sources of LD-related tools include the series of OCLC surveys (see chapter 1) and survey articles on *Wikipedia* and the W3C. For the reader looking for quick suggestions, a survey of the OCLC results indicates that Dydra, OpenLink Virtuoso, Jena, SESAME, and AllegroGraph are all common tools. Increasingly, there are cloud-based services available to support RDF triplestores, including Dydra. There is another set of tools focused on providing support for viewing LD data. These viewers include rdf:SynopsViz, Tabulator, OpenLink Data Explorer, and a range of other viewers. The W3C site on Semantic Web tools remains an up-to-date catalog of tools as well as standards and best practices.

Wikipedia: Triplestore
https://en.wikipedia.org/wiki/Triplestore

W3C: Large TripleStores
www.w3.org/wiki/LargeTripleStores

rdf:SynopsViz
http://synopsviz.imis.athena-innovation.gr

W3C Semantic Web wiki, Category:Tool
https://www.w3.org/2001/sw/wiki/Category:Tool

Tabulator
https://github.com/linkeddata/tabulator

OpenLink Data Explorer
http://ode.openlinksw.com

LAM-specific tools in the LD community tend to center on a specific vocabulary or use. The BIBFRAME editor and other tools made available by LoC and Zepheria, for example, provide support for working with BIBFRAME and related metadata but are not appropriate for more generalized work. Other tools common in the LAM community, such as ArchivesSpace, do not include built-in editor support that is LD-focused but are designed around principles of linking and can make use of APIs and data integration and export tools that are useful in the LD community. Just as there was value in tools that sought to automatically catalog web pages or extract metadata from structured HTML, there is an emerging set of tools dedicated to harvesting and transformation of LD in web pages. One such tool is the RDF Translator developed by Alex Stolz. This tool supports input via RDFa, Microdata, XML, N3, NT, and JSON-LD and translates that output to RDFa, microdata, pretty-xml, XML, N3, NT, and JSON-LD formats. The service is built on a Python library (RDFLib) and also uses pyRdfa, pyMicrodata, and rdflib-jsonld libraries. As this issue finds in many cases, Python and Python-related libraries are becoming a common platform for LD work across LAM and other institutions.

RDF Translator
http://rdf-translator.appspot.com

A similar tool that facilitates working with JSON-LD data is the JSON-LD Playground. Similar to the RDF Translator, the JSON-LD Playground tool provides different serializations of JSON-LD data, including translation into N-Quads and multiple forms of JSON data. While the focus of this issue is on LD metadata, another area of interest is RDF and LD visualization tools. Tools commonly used in the community include Gephi and Tableau. Ontology-specific visualization tools, such as the WebVOWL platform, provide the ability to visualize FOAF and other ontologies (http://vowl.visualdataweb.org/webvowl.html). In addition to these client-based tools, web-based tools such as Node.js, D3.js, and MongoDB are increasingly common in helping to display LD relationships.

JSON-LD Playground
http://json-ld.org/playground/index.html

WebVOWL
http://vowl.visualdataweb.org/webvowl.html

As LD platforms mature, more "comprehensive" or end-to-end tools are becoming available. One system that is featured in Wood et al.'s *Linked Data* is the Callimachus project, an LD ingest, hosting, and publishing platform.[9] This platform includes template systems for web publication, allowing authors to create Semantic Web applications. The platform adheres to each of the five building principles of LD (i.e., open on the web, machine-readable, non-proprietary, RDF-based, linked). The publishers of Callimachus compare it to content management systems (CMSs), differentiating it from these platforms in that Callimachus primarily manages structured data. Another similar tool, Graphity, provides a unified data publishing platform that includes an LD client, publishing platform, and processing engine. Like other tools, Graphity is available under an open-source license, although a commercial provider (GraphityHQ) provides commercial services. Another such tool, Arches, is a cultural heritage inventory and management system. Although this platform was not necessarily designed around LD principles, there are an increasing number of use cases related to how this platform is making use of LD, including one connected with the city of Los Angeles, California.[10]

Graphity
https://github.com/Graphity

Issues in LD Translation

Enhancing Data via LD

A common use case for LD is the use of vocabularies and authorities to create metadata with more obvious community value. While the LAM community as a whole appears to agree on this goal and the value of the work, there is still much work to do in creating the tools that enable widespread normalization. Johnson and Estlund suggested a number of potential outcomes from LD processes, including removal of "noise," normalized presentation, assignment of URIs for curated objects, and migration from legacy metadata to new LD vocabularies.[11] By removing "noise," Johnson and Estlund mean "eliminating valueless metadata entries" such as elements without content or values that essentially say "unknown." One application of this idea of URI resolution has been documented by Klein and Kyrios.[12] The project matched VIAF records against *Wikipedia* entries using the Pywikipediabot framework, a Python-based *Wikipedia* framework. Starting with VIAF clusters with a *Wikipedia* link, associated *Wikipedia* pages were scanned for content. One of the primary outcomes of this work is the notion that the VIAF bot may be a model for application with other types of data. It successfully connected VIAF data and *Wikipedia* pages at the "hundreds of thousands" of pages level.

The generation of LD through automated text and metadata analysis is an area where research is advancing the integration of tools, including text analysis, natural language processing (NLP), and connection with existing authority vocabularies. Pattuelli et al., for example, developed a Python-based platform to match DBpedia URIs and LoC Name Authority File (NAF) records as well as applying named-entity recognition using the Natural Language Toolkit (NLTK) platform.[13] Similarly, some libraries are using programs to bring LD into discovery platforms. For example, Hatop designed a platform to create a Solr index using LD sources.[14]

Natural Language Toolkit (NLTK 3.0) Documentation
http://nltk.org

Conversion Strategies

The conversion of metadata to LD is one of the more complex topics in the LD community, often complicated by issues of scale and diversity of metadata as well as the fact that LAM institutions have not yet settled on new systems, meaning that LD systems often contain secondary or derivative instances of metadata. Two strategies in particular around conversion, iterative (i.e., retransforming metadata as new features and requirements are integrated) and cumulative (i.e., building on previously transformed metadata) are commonly used. OCLC, for example, combines data from production and experimental processes to enhance MARC records and publish new data as Linked Data using a cumulative process. OCLC's new model for representing Works is motivated by FRBR concepts and algorithms but follows its own set of relationships to express the creative work.[15] This identifier is represented via RDFa as well as via the OCLC xID service.[16] In contrast, the LoC BIBFRAME tools encourage iterative transformation through the regular incorporation of enhancements that require the complete retransformation of all data.

Although the next clear step, particularly in the bibliographic arena, is to get to a level of system and schema maturity to move away from older systems and standards, it appears that this is still an aspiration rather than a realized goal for most projects. The Oslo Public Library's transformation to LD is an example of one project that has reached that goal, moving away from its old ILS to LD metadata using the Koha ILS in early 2015.[17] The Oslo Public Library was an early innovator in RDF and LD research, having developed MARC2RDF in 2011 as well as experimenting with LD-based services.

marc2rdf
https://github.com/digibib/marc2rdf

Conclusion

Chapter 3 has explored the systems, vocabularies, and standards in use in the LAM community to generate or make use of LD and has explored key issues in LD generation—options for enhancing LD as well as approaches to conversion of existing metadata to LD. Given the number of state of adoption reports that have been completed in recent years as well as the upcoming release of new survey results on adoption, this report did not seek to provide a comprehensive listing of tools, standards, and services. Rather, this chapter focused on example tools and standards and identified themes and trends in more depth. In chapter 4, we consider several of these themes in more detail and consider what the coming year might hold in LD exploration and adoption.

Notes

1. Erik T. Mitchell, "Library Linked Data: Research and Adoption," *Library Technology Reports* 49, no. 5 (July 2013), 46.
2. David Wood, Marsha Zaidman, and Luke Ruth, with Michael Hausenblas, *Linked Data: Structured Data on the Web* (Shelter Island, NY: Manning Publications, 2014); Bernadette Hyland and David Wood, "The Joy of Data: A Cookbook for Publishing Linked Government Data on the Web," in *Linking Government Data*, edited by David Wood (Cham, Switzerland: Springer International Publishing, 2011), 3–26. See also Eric M. Hanson, "A Beginner's Guide to Creating Library Linked Data: Lessons from NCSU's Organization Name Linked Data Project," *Serials Review* 40, no. 4 (2014): 251–58, http://dx.doi.org/10.1080/00987913.2014.975887; 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; Martin Malmsten, "Exposing Library Data as Linked Data," Web Technology Laboratory, Ferdowsi University of Mashhad, 2008, http://wtlab.um.ac.ir/images/e-library/linked_data/other/Exposing%20Library%20Data%20as%20Linked%20Data.pdf; Silvia. B. Southwick, "A Guide for Transforming Digital Collections Metadata into Linked Data Using Open Source Technologies," *Journal of Library Metadata* 15, no. 1 (2015): 1–35, http://dx.doi.org/10.1080/19386389.2015.1007009.
3. Eleni Tsalapati, Nikolaos Simou, Nasos Drosopoulos, and Regine Stein, "Evolving LIDO Based Aggregations into Linked Data" (paper presented at CIDOC 2012, Helsinki, Finland, June 10–14, 2012), http://network.icom.museum/fileadmin/user_upload/minisites/cidoc/ConferencePapers/2012/simou.pdf; Ellen Van Keer, "Moving from Cross-Collection Integration to Explorations of Linked Data Practices in the Library of Antiquity at the Royal Museums of Art and History, Brussels," in *Current Practice in Linked Open Data for the Ancient World*, ISAW Papers 7, edited by Thomas Elliott, Sebastian Heath, and John Muccigrosso (New York: Institute for the Study of the Ancient World, 2014), 5–8, http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/vankeer.
4. Nandana Mihindukulasooriya and Roger Menday, eds., "Linked Data Platform 1.0 Primer," W3C Editor's Draft, September 11, 2015, https://dvcs.w3.org/hg/ldpwg/raw-file/default/ldp-primer/ldp-primer.html.
5. Greg Jansen, "Linked Data Platform (Slides)," last updated March 23, 2015, https://wiki.duraspace.org/pages/viewpage.action?pageId=68065495.
6. Pat Riva and Maja Žumer, "Introducing the FRBR Library Reference Model" (paper presented at IFLA World Library and Information Congress, Cape Town, South Africa, August 16–20, 2015), http://library.ifla.org/1084.
7. "In Depth," Yale Center for British Art, accessed July 28, 2015, http://britishart.yale.edu/collections/using-collections/technology/in-depth.
8. "British Museum Semantic Web Collection Online," British Museum website, accessed July 28, 2015, http://collection.britishmuseum.org.
9. Wood et al., *Linked Data*.
10. Annabel Lee Enriquez, "HistoricPlacesLA.org, Powered by Arches v3.0, Launches in Los Angeles," Arches, March 4, 2015, http://archesproject.org/historicplacesla-org-powered-by-arches-v3-0-launches-in-los-angeles.

11. Thomas Johnson and Karen Estlund, “Recipes for Enhancing Digital Collections with Linked Data,” *Code4Lib Journal*, no. 23 (January 17, 2014), http://journal.code4lib.org/articles/9214.
12. Maximilian Klein and Alex Kyrios, “VIAFbot and the Integration of Library Data on Wikipedia,” *Code4Lib Journal*, no. 22 (October 14, 2013), http://journal.code4lib.org/articles/8964.
13. M. Cristina Pattuelli, Matt Miller, Leanora Lange, Sean Fitzell, and Carolyn Li-Madeo, “Crafting Linked Open Data for Cultural Heritage: Mapping and Curation Tools for the Linked Jazz Project,” *Code4Lib Journal*, no. 21 (July 15, 2013), http://journal.code4lib.org/articles/8670.
14. Götz Hatop, “Integrating Linked Data into Discovery,” *Code4Lib Journal*, no. 21 (July 15, 2013), http://journal.code4lib.org/articles/8526.
15. Carol Jean Godby, Shenghui Wang, and Jeffrey K. Mixter, *Library Linked Data in the Cloud: OCLC’s Experiments with New Models of Resource Description*, Synthesis Lectures on the Semantic Web: Theory and Technology (Morgan & Claypool, April 2015), 64, http://dx.doi.org/10.2200/S00620ED1V01Y201412WBE012.
16. “WorldCat Work Descriptions,” OCLC Developer Network, accessed September 14, 2015, https://www.oclc.org/developer/develop/linked-data/worldcat-entities/worldcat-work-entity.en.html.
17. Asgeir Rekkavik, “RDF Linked Data as New Cataloguing Format at Oslo Public Library” *SCATNews*, Newsletter of the Standing Committee of the IFLA Cataloguing Section, no. 41 (June 2014): 13–16, www.ifla.org/files/assets/cataloguing/scatn/scat-news-41.pdf.