Introduction

Library data is in a state of change. Google has reinvented how information is stored, searched, and displayed to users. Today, users are more likely to go to search engines to begin queries instead of using the library catalog, but library data is not present within these search results (Gonzales, 2014, p. 12). Currently, library data is held within metaphorical “silos,” only accessible through library catalogs. The breadth and substance of this information is lost to the many users only searching through online search engines. To reach these users, libraries must transform the way their data is stored in order for it to be retrieved in online searches. How to do this is a topic of great importance within the library and cultural institution community, and one in which there is much discussion.

In 2009, Tim Berners-Lee introduced the concept of Linked Data (LD) and the Semantic Web. Instead of being constructed of individual webpages like the internet we know today, the Semantic Web will be individual data points connecting to other data points. These linked data points are called Linked Data. Constructed using Resource Description Framework (RDF), LD is made up of triples; a subject, predicate and object. Each part of the triple can be represented by a Uniform Resource Identifier (URI), or more particular a Uniform Resource Location (URL) (Hallo, Luján-Mora, Maté, & Trujillo, 2016).

This essay provides an overview of the literature about the LD movement and libraries by a range of sources published between 2013 and 2016 and is organized in three sections: Moving Away from MARC and ISBD, Linked Data and Controlled Vocabularies, and Problems with Linked Data and Libraries.

Transitioning Away from MARC and ISBD

Libraries hold a vast amount of information, although much of this information is unable to be used outside of a library environment. Transitioning to LD will enable library data to be used by other institutions. Hallo, Luján-Mora, Maté, and Trujillo (2016) state, “The Semantic Web encourages institutions, including libraries, to collect, link and share their data across the Web in order to ease its processing by machines to get better queries and results” (p. 117).

Moving to LD will require libraries to change the way their data is constructed.

Ever since the invention of Machine-Readable Cataloging (MARC) in the 1960s by Henriette Avram, most, if not all, of a library’s data was placed within the strict confines of this system. Gonzales (2014) discusses the limitations of MARC’s rigid structure and it’s disuse outside of the library community. Gonzales suggests moving on from MARC, and placing all library data in RDF. “Clearly needed is a more flexible and extensible format that allows for the expression of relationships between points of data and the ability to link that data to other information outside of the presently insular library catalog” (Gonzales, 2014, p.12). A new framework, developed by the Library of Congress, called the Bibliographic Framework or BIBFRAME, is touted as MARC’s replacement. Tharani (2015) provides an overview on how BIBFRAME will enable libraries to use the LD technology.
Creating a new system to facilitate Linked Data creates problems on how to handle the legacy data already in MARC. Cole, Han, Weathers, and Joyner (2013) describe their efforts in transferring legacy data into Metadata Object Description Schema (MODS) at the University of Illinois. Needing a system that would manage the inclusion of URLs, they rejected MARC21XML, and chose Schema.org as the bibliographic record carrier.

Mapping to new schemas can lose important information contained within the MARC record. Thompson, Baxmeyer, Bell, and Green (2016) used the World Wide Web Consortium (W3C) Annotation Data model and vocabulary to transfer personal dedications found in the volumes of philosopher Jacques Derrida’s library. Originally coded in notes, Thompson et al., moved the textual information found in the MARC 500 fields into structured data points that could then create links for the people or places found in the descriptions. Weitz, Toves, Vizine-goetz, Naught, and Bremer (2016), searched for roles of the performers or contributors to musical sound recordings or scores in the 500 fields that were not included in the authority files. Adding the structured codes of these roles into the XML records helped match individuals to their correct URLs and provided more description within the record.

Along with MARC, the International Standard Bibliographic Description (ISBD) may need to be replaced. Dunsire (2014) discusses the future of ISBD in a linked data environment. While ISBD can benefit linked data by supporting hierarchical relationships between elements and attributes, “linked data is best served if semantic data is completely free of syntactic content; that is, if the data values are independent of data structure” (Dunsire, 2014, p. 866). This allows the values to be displayed more freely. The continued use of ISBD may hinge on how well the system adapts to the changing needs brought on by linked data.

Linked Data and Controlled Vocabularies

Linked Data works best when the vocabularies used are standardized. Controlled vocabularies are required so that matching resources can be linked together under one name. If multiple names are used for the same person, place, or thing, how will catalogers and users now which name to use, or which name is the most accurate? Ryan et al. (2015) stresses the importance of controlled vocabularies in the future of cataloging, “To implement linked data technologies, librarians need to begin focusing on data element rather than the record” (p. 77).

Much of the literature focused on moving to controlled vocabularies in various institutions. Bushman, Anderson, and Fu (2015) described the process of creating a controlled vocabulary for the medical subject headings found in the records of the National Library of Medicine. Ryan et al. (2015) discussed creating a program, Linked Logainm, to create authority records for names of Irish places. The initiative to create a controlled vocabulary for video games through the Game Metadata and Citation Project (GAMECIP), a collaboration between the University of California, Santa Cruz and the University Libraries at Stanford, is detailed in Kaltman et al. (2016).

When creating controlled vocabulary, it is important that the terms map to other datasets to implement interoperability. Ryan et al. (2015) found a 97% rate of precision in the Linked Logainm vocabulary to other datasets (p. 81). Most of the problems occurred between the
specificities of each dataset. Loganim used more specific terms to distinguish between towns, villages, and counties, whereas some other datasets only used one term “town” to encompass all different term options. In mapping authority file names in MARC to the Virtual International Authority File (VIAF), Cole et al. (2013) found 81% of personal names, 75% of corporate names and 59% of family names had matching VIAF URIs (p. 180). Alternatively, Radio and Hanrath (2016) found the Online Computer Library Center’s (OCLC) Faceted Application of Subject Terminology (FAST) was not a suitable controlled vocabulary for their collections at the University of Kansas, and that an increase discoverability of resources was not as noticeable as thought of by implementing LD aspects into the catalog.

Problems with Libraries and Linked Data

While linked data can improve library data, libraries need to improve their ability to use and create linked data. Hallo et al. (2016) observed problems with libraries and linked data. First, libraries do not have the technical tools available to create or use linked data. Integrated library systems (ILS) are not able to utilize URIs, and many cannot even display RDA terms yet. Moulaison and Million (2014) argue, that as of now, linked data is still in its conceptual phase, and that while libraries should persuade vendors to create ILSs that can adapt to changing technology, linked data may not work within a library framework.

Second, Linked Data is still in its infancy and libraries lack experience with Linked Data models. With this problem in mind, Southwick (2015) described the step-by-step plans and challenges faced by the University of Nevada, Las Vegas Libraries when they initiated a Linked Data project for their digital materials. Along with this, too many vocabularies and schemas were used. The use of multiple schemas and vocabularies lead to lower interoperability. Standards need to be written to prevent this problem in the future. Third, with the increase in using digital materials, libraries are owning less and less of their collection (Gonzales, 2014, p. 17). Libraries need to be mindful of the publishing and licensing rights of the materials in their collection, and how those rights might impact sharing the materials in a linked data environment. As the community continues to push towards LD and the Semantic Web these issues will need to be addressed in the future.

Conclusion

After reviewing the literature from 2013 to 2016, it is apparent that libraries are greatly invested in moving towards a LD environment. To do this, new carriers for bibliographic carriers need to be created, standards need to be written, and ILS systems need to be reimagined. Future research should continue to focus on the implementations of controlled vocabularies and how to emigrate legacy metadata into the new schemes. Libraries are a vital part of the information community, and the advancements made to increase discoverability and usability of the information libraries possess can only benefit the public.
Resources


Prepared by Nichole Kelley, MLIS student at Dominican University, River Forest, Illinois