Linked Data in the Library Cloud: Technology of the Future?

Keynote presentation: 2nd International Conference on Academic Libraries (ICAL-2013)

Heather Lea Moulaison, PhD

Introduction

Technology means different things to people at different times, and it can be difficult to imagine what the future has in store for us and the technologies that we are currently using in libraries. Cloud computing in particular has been identified as a method for libraries to embrace new advances in technology (c.f. Corrado & Moulaison, 2011) both now and into the future. This paper takes that idea one step further by suggesting that linked data stored in the cloud might be a specific way for libraries to take advantage of emerging standards and web technologies. One simple way of defining the cloud is to equate it to the Internet (Anderson & Rainie, 2010); information stored in the cloud is then accessible over the Internet. What is linked data, and how does it fit into the current technology landscape of cloud computing? In exploring the question of linked data in academic libraries, this paper will examine the notion of linked data as it relates to the world of library data and metadata, situating it as a series of standards that demonstrate and represent a host of new possibilities for the future of academic libraries.

Academic libraries do not exist in isolation; they rely heavily on technologies and standards that are already being explored and pioneered in other sectors. The semantic web is one concept that is increasingly of interest. Tim Berners-Lee, the father of the World Wide Web, had a vision at the turn of the millennium. On May 17, 2001, he and two colleagues published a paper in *Scientific American* describing what it would be like if computers could actually understand each other (Berners-Lee, Hendler, & Lassila, 2001). The idea was that users would not just give and receive direct instructions through a web-enabled device, but that the computers would actually understand what data meant and make sense of it after minimal direction, not unlike humans. Berners-Lee called this idea the semantic web, tying the proposed *web* to the notion of *semantics* being *meaning*. It has also been called the *intelligent web*, *web 3.0*, and the *web of data*, and it is the basis for the linked data concepts that academic libraries and others have been exploring.

> "The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data" (Berners-Lee, 2006)

For this vision of computers understanding each other to happen, a lot of standards and protocols need to be in place. As of right now, the semantic web as such is not functioning, but things are improving all the time. Both in libraries and beyond, new advances are paving the way for something akin to the semantic web to one day work. And when something akin to the semantic web is in place, libraries will be able to integrate their resources into that network and harness its power for patrons in ways that have up to this time been unimaginable.

Citation: Moulaison, H. L. (2013). *Linked data in the library cloud: Technology of the future?* Paper presented at the 2nd International Conference on Academic Libraries (ICAL-2013), New Delhi, India, February 12-15, 2013. Retrieved from http://codabox.org/

Academic Libraries and the Emerging Metadata Ecosystem

Academic libraries traditionally provide access to scholarly knowledge for researchers and students. As early as the 1990s, libraries took the option to think about modeling their work in their field of activity, the bibliographic universe, in a document called *Functional Requirements for Bibliographic Standards* (FRBR) (IFLA Study Group, 1997). FRBR puts forth an entity-relationship model of the bibliographic universe, enumerating the various entities, grouping them, and making explicit the relationships between them. This kind of modeling exercise is just the sort of exercise necessary to understanding, among other things, library data and how each discrete element of the bibliographic universe and its attributes might relate to other elements. These connections or links that FRBR makes explicit are not unlike the links that Berners-Lee envisions as being machine understandable on the semantic web. In fact, many of the changes the international library community has been exploring in the past fifteen years or so actually dovetail with and support efforts to make library data and content accessible on the semantic web, with FRBR as a foundation. Two prime examples being explored in the United States and around the world are the creation of descriptive metadata through the use of a new FRBR-friendly cataloging code called Resource Description and Access (RDA) and the creation of a new linked data-friendly encoding scheme to replace MARC.

RDA: New Cataloging Code

Resource Description and Access (RDA), the new cataloging code to be adopted in the United States formally beginning at the end of March 2013, is based on FRBR. In considering this model of the bibliographic universe, it is possible to imagine new and different ways of representing that universe in the context of the web. FRBR models the bibliographic universe specifically with the goal of satisfying four user tasks: 1) find, 2) identify, 3) select 4) obtain. Elaine Svenonius (2000) has suggested a 5th user task to consider, 5) navigate (implying discovery). This fifth task is incredibly important, and the more recent *Statement of International Cataloging Principles* (2009) incorporates all five of these user tasks in the establishment of its "Objectives and Functions of the Catalogue". RDA was created with these user tasks and principles in mind, moving to a FRBRized version of the catalog away from an ISBD-based record system.

FRBR is a conceptual model and not cataloging code. More precisely, it is an entity-relationship model. FRBR identifies every *entity* that comes into play in the bibliographic universe. There are 10 of them including person, corporate body, place, concept, and work. FRBR defines the relationships between and among these different entities/groups of entities. Persons can create works, corporate bodies can produce manifestations, works can have as subjects other works, or persons, or places, etc.

Bibliographic records are not the only kinds of information needed for catalogs to function, and accordingly, there are other aspects to keep in mind when considering the bibliographic universe. *Functional Requirements for Authority Data* (FRAD) (Patton, 2009) defines user tasks and identifies attributes for people, corporate bodies, etc. *Functional Requirements for Subject Authority Data* (FRSAD) (Zeng, Zumer, & Salaba, 2010) identifies subject thema and nomen. Together, FRBR, FRAD and FRSAD

Citation: Moulaison, H. L. (2013). *Linked data in the library cloud: Technology of the future?* Paper presented at the 2nd International Conference on Academic Libraries (ICAL-2013), New Delhi, India, February 12-15, 2013. Retrieved from http://codabox.org/

are the three members of the FR-family and they cover in depth the entirety of the bibliographic universe's elements and their attributes in a way that demonstrates the relationships between them.

This gets to be really exciting when we start to consider all of the attributes an entity might have. In our current approach to organizing library information, Steven King is the author of several books. In the FRBR model and in the resulting RDA-based authorities, Steven King is a person with a gender, a nationality, writing in a certain genre at a certain point in time, who writes books. The attributes of his *personness* allow us potentially to create additional links to other authors with similar or different attributes, too. FRAD lists possible attributes for a person that can be included in the authority record for the person. Current authority records for people can be quite sparse in terms of attributes recorded, and even if some attributes are identified, they are not machine understandable, but are only human-readable.

The FRAD list of attributes does not look so different from the information in Wikipedia for persons that works might be by or about. See Figure 1 for a screen shot of organized data from Wikipedia. This organized data can easily be encoded so that machines understand it, bringing the focus back to Berners-Lee's notion of machines understanding data and being able to talk to each other on the semantic web.

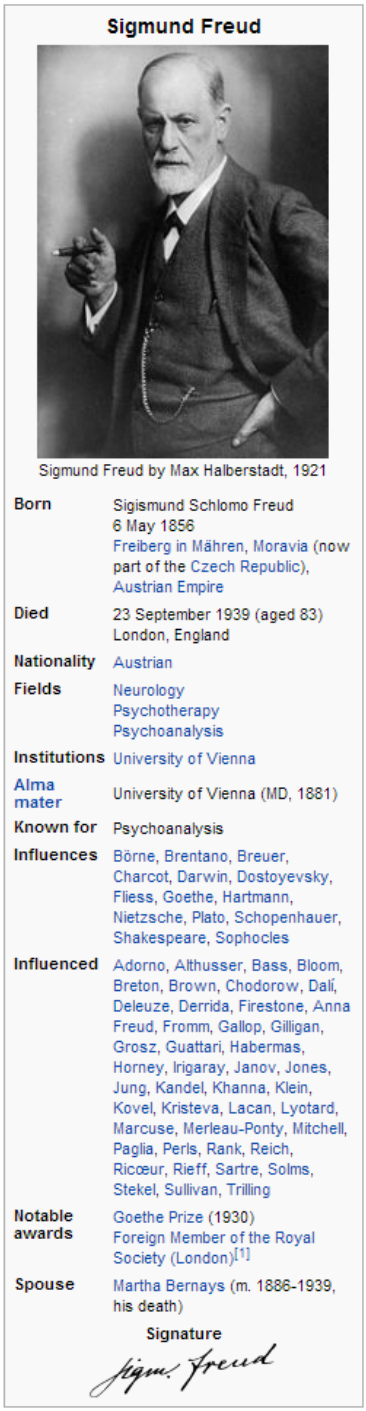

Figure 1: Encoded data about Sigmund Freud from Wikipedia (English) (http://en.wikipedia.org/wiki/Sigmund_Freud)

The Move away from MARC

As librarians move to encode information about attributes that will permit relationships and links to be made in library catalogs, the degree to which the encoding scheme, MARC, is not sufficient becomes increasingly apparent. MARC was designed to transmit cataloging record data as whole records (Ford, 2012) and does not do well at making explicit the relationships between different entities or attributes of these entities in the bibliographic universe. Currently, library systems cannot pull a set of books written by 18th century women poets. The data does not currently exist (c.f. Coyle, 2010), and even if it did, our systems could not produce a list of results.

The fact is that there are a lot of other things MARC does not do, as well. The MARC standard is ISO 2709; it has been around since before the web and consequently is not web-ready. As an encoding language, MARC is also not terribly granular (Tennant, 2002), and it does not provide for the inclusion of encoded machine understandable data other than in a few select fields. Even worse, it is difficult to integrate MARC into other systems,

Citation: Moulaison, H. L. (2013). *Linked data in the library cloud: Technology of the future?* Paper presented at the 2nd International Conference on Academic Libraries (ICAL-2013), New Delhi, India, February 12-15, 2013. Retrieved from http://codabox.org/

especially those outside of the library and publishing world. In light of these limitations, it becomes necessary to re-examine not only library data, but also library encoding.

MARC does not do what we want our systems to be able to do into the future. The Library of Congress has explained its intention to move away from MARC in May 2011 (Miller, Ogbuji, Mueller, & MacDougall, 2012). They are calling this move the Bibliographic Framework Transition Initiative (BIBFRAME) (http://www.loc.gov/marc/transition/). "The Initiative aims to re-envision and, in the long run, implement a new bibliographic environment for libraries that makes 'the network' central and makes interconnectedness commonplace" (Miller et al., 2012, p. 1). Indeed, linked data will be used in place of MARC as the way this interconnectedness will be achieved.

Bibliographic Data as Linked Data

A better way to show relationships between entities and between their attributes is through a kind of very simple machine-understandable *sentence* called an RDF triple. That triple (or sentence) will have a subject, a predicate, and an object, all of which, ideally, will be machine-understandable also.

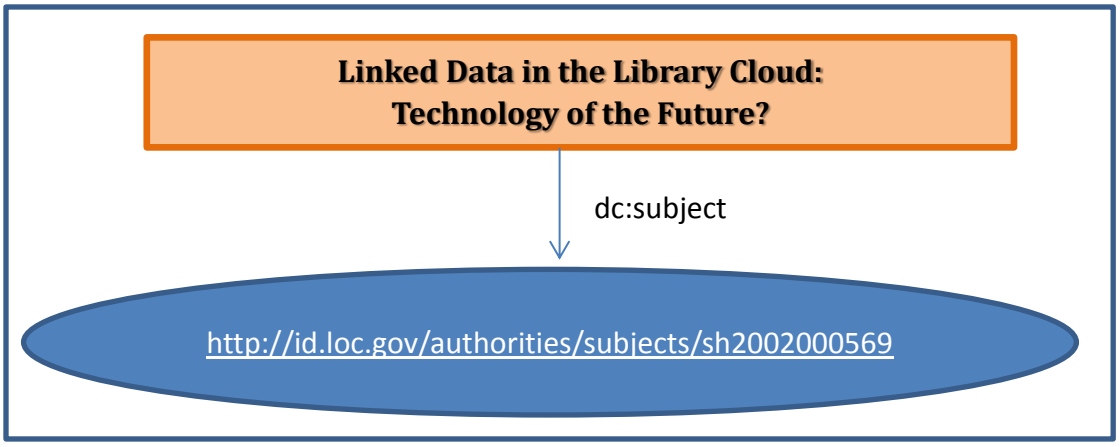

Figure 2. Graph of an RDF triple for the subject of this paper

A graph like the one depicted in Figure 2 is one way of visualizing the three-part RDF triple, with the subject as the title of the paper, the predicate as the Dublin Core element indicating subject, and the object as a URI for the linked data-readable version of the subject heading for the semantic web. Depending on the context in which this triple is stored, it might be possible to pull a set of results for other works whose topics are also the semantic web; to indicate additional topics of this presentation, to indicate information about the author, the venue, the date of publication, and other information and to discover relationships between those objects and other subjects.

When these relationships are built out visually, a quite complex graph structure can result. Even more powerful is the possibility of tapping into other related graphs using the same or similar structure. The potential for demonstrating relationships is one of the main concepts behind linked data, and this is not possible at this time with the MARC records libraries maintain. Currently, MARC records where human-

Citation: Moulaison, H. L. (2013). *Linked data in the library cloud: Technology of the future?* Paper presented at the 2nd International Conference on Academic Libraries (ICAL-2013), New Delhi, India, February 12-15, 2013. Retrieved from http://codabox.org/

readable text strings of what a record is about, who wrote it, etc. appear somewhere in the record. With linked data, libraries have the possibility of moving from a record-by-record model to a new way of thinking represented by RDF triples.

Libraries, therefore, have been moving to re-think their data in a way that is linked data-friendly since the 1990s. Because library data is increasingly hosted in the cloud, because other aspects of linked data (e.g. predicates and objects of the RDF triples) are also hosted in the cloud, linked data is a technology that is fundamentally cloud-based; one that is moving libraries into the future. Below, this paper explores some of the technical requirements for linked data that librarians should understand in a basic way.

Technical Requirements for Linked Data

Tim-Berners Lee's solution to the problem of making data machine-understandable has been to encode it as linked data in order for it to become part of the semantic web. There are four criteria for linked data according to Tim Berners-Lee (2006):

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.

Each of these four criteria is examined in further detail below.

Use URIs as names for things

Identifiers, the *nouns* pictured in the squares and ovals on the graph in Figure 2 should ideally be uniform resource identifiers (URIs), and not just strings of text. The W3C (World Wide Web Consortium) (2006) describes URIs as:

> short strings that identify resources in the web: documents, images, downloadable files, services, electronic mailboxes, and other resources. They make resources available under a variety of naming schemes and access methods such as HTTP, FTP, and Internet mail addressable in the same simple way (Learning About URIs, para. 3).

There are two kinds of URIs -- URLS (uniform resource locators) and URNs (uniform resource names). For the purpose of thinking about linked data, a URI is a URL that represents a noun. A URI for a person, for example, could be that person's homepage. A URI for a city might be the city's main .gov page.

Use HTTP URIs so that people can look up those names

Hypertext transfer protocol (HTTP) is one of the web protocols that enables machines to share content on the web. "HTTP defines how messages are formatted and transmitted, and what actions web servers

Citation: Moulaison, H. L. (2013). *Linked data in the library cloud: Technology of the future?* Paper presented at the 2nd International Conference on Academic Libraries (ICAL-2013), New Delhi, India, February 12-15, 2013. Retrieved from http://codabox.org/

and browsers should take in response to various commands" ("HTTP", 2012). HTTP was distinguished above from FTP (File Transfer Protocol) and email addresses, and here is identified as the preferable URI from among the others.

When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)*

The two main standards mentioned are RDF and SPARQL. The topic of RDF was broached earlier when investigating the triples in a preliminary way. RDF stands for *Resource Description Framework.* According to the semantic web section of the W3C: "RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed." (RDF Working Group, 2004) More specifically, RDF is written for computers in XML; it is not meant to be human-readable (hence the tendency to depict relationships using graphs) ("Introduction to RDF", 2013).

SPARQL is the query language for the semantic web, and as such, it "offer developers a way to write queries across the wide range of RDF information on the internet" ("W3C RDF and OWL Activities," 2013). It is somewhat difficult to build queries with SPARQL. This issue will be explored further below.

Include links to other URIs. so that they can discover more things.

Given the broad range of relationships that can be encoded, links have the potential to permit guided discovery like never before. The potential, given the mission and goals of libraries, is great, and libraries are working to make sure their data and contents will able to be a part of this new cloud-based library environment.

Libraries and Linked Data

The Library Linked Data Incubator Group was brought together in 2011 to define how the linked data model might affect libraries and their report sheds a great deal of light on the question of linked data and libraries (Library Linked Data Incubator Group, 2011a). Linked data "refers to data published in accordance with principles designed to facilitate linkages among datasets, element sets, and value vocabularies." In a related document, the group outlines examples of datasets, element sets, and value vocabularies (Library Linked Data Incubator Group, 2011b) that will serve in part as the basis for the discussion below.

Element sets

Element sets can be understood to be descriptive metadata used to describe a resource. Elements from element sets form the predicates in the sentences (the RDF triples) described earlier, so they are an essential part of the linked data environment.

Citation: Moulaison, H. L. (2013). *Linked data in the library cloud: Technology of the future?* Paper presented at the 2nd International Conference on Academic Libraries (ICAL-2013), New Delhi, India, February 12-15, 2013. Retrieved from http://codabox.org/

The Dublin Core Metadata Element Set (DCMES) (http://dublincore.org/documents/dces/) might seem like a quintessential element set, familiar to many librarians. Other element sets have been created by other communities to meet their particular description needs. If describing blood samples, a library will need different elements from when it is describing sculptures, or root structures of plants, or bindings on illuminated manuscripts from the middle ages.

All of the Dublin Core (DC) elements reside together in the DC namespace which is designated by a URL. That namespace is located in the cloud, and is in some ways like the house where the members of the DC element set lives. Computers can go and visit the namespace when reading an RDF statement in order to make sense of the relationship being presented.

There are many namespaces on the web -- as many as there are registered element sets. Each element in an element set has a different function. One might cook (or designate the title of the work). Another might take out the trash (or designate the creator of the work). There may be redundancies between elements in various different namespaces (multiple namespaces might have an element for creator/author). There should also be elements that are unique to each element set and its namespace or that are directed to be completed in a way that is unique. To continue with the analogy, some elements living in the namespace house might have the ability to fix a car (or allow for encoding of detailed information about the style and time period in which a painting was created); some might know how to paint a roof.

Value vocabularies

Value vocabularies are the lists of terms that are used to fill in the elements in a controlled vocabulary environment. Controlled vocabularies have been around a long time in libraries and now many of them have been published as linked data as value vocabularies.

The Library of Congress Subject Headings (LCSH), for example, have been published by the Library of Congress are now a value vocabulary that can be used on the semantic web. Not only is LCSH published in print in the big red books, it also is published in a way that is machine readable, as MARC for authority as subject headings, and a way that is machine understandable, as linked data. Data from the Name Authority File has also been published as linked data using linked data standards. And, classification schema like Dewey have been published as linked data as well.

Datasets

Datasets are the equivalent of databases containing structured linked data that are searchable using linked data standards like SPARQL. Many exist, and many of them are open, meaning the contents are free to be reused.

DBpedia. The most important dataset by far in the current linked data environment is the DBpedia (http://dbpedia.org) dataset. "The DBpedia project extracts various kinds of structured information from Wikipedia editions in 111 languages and combines this information into a huge, cross-domain

Citation: Moulaison, H. L. (2013). *Linked data in the library cloud: Technology of the future?* Paper presented at the 2nd International Conference on Academic Libraries (ICAL-2013), New Delhi, India, February 12-15, 2013. Retrieved from http://codabox.org/

knowledge base" ("DBpedia Data Set", 2012) as linked data. The highly encoded data in the sidebars for Sigmund Freud (Figure 1) exist for other kinds of Wikipedia entities as well, not just people. Geographic locations have highly encoded sidebars with information about latitude and longitude, countries have highly encoded information about the people, the government, the currency, etc. All of this highly encoded data of attributes for these entities has been harvested and is now part of the DBpedia database. Because of the encoding structure, it is now theoretically possible to run a SPARQL query and pull a list of locations based on shared attributes such as currency, religion of the people, or even population density or elevation. With such an initiative, however, the caveat is that the data needs to be of high quality: data needs to be recorded accurately, and it needs to be encoded correctly.

OpenLibrary. OpenLibrary (http://openlibrary.org/) is another dataset. For librarians, it is particularly interesting because of its focus on books and authors, but also for the hyperlinks it provides directly into online library catalogs. OpenLibrary is community-based but contains information from linked data sources.

Drawbacks to the Linked Data Model in Libraries

Although there is much potential for linked data in libraries, there are a certain number of drawbacks that should be acknowledged. Three primary ones are the need for human oversight when encoding, problems with SPARQL as a query language, and issues knowing how and where data is encoded and stored. Human intervention is necessary for de-duplicating names of entries for people; at this point, there is no way to do an automated match of personal names based solely on character strings, for example, and have the matching be accurate. Bob Jones in the library catalog may or may not be the Bob Jones in Wikipedia. Second, issues arrive with the SPARQL protocol. SPARQL is not intuitive to use at all, and is meant to be used by developers. Any kind of sophisticated query needs to be laboriously generated by hand. Web-savvy users of technology used to Google, an information retrieval system that functions as a "bag of words", will find SPARQL difficult to use (Freitas, Curry, Gabriel Oliveira, & O'Riain, 2012). As well, it is also not easy to know how individual datasets are encoded, what information they contain, and how that information is recorded. What are the different sets of standards in use in a given repository? Any experienced searcher knows that when carrying out federated searches, the results are inconsistent at best, and that's essentially what using a variety of linked data datasets proposes. With SPARQL, we're back to the days of needing to know which database to query and how, in order to get what information. SPARQL in a way seems similar to the old-fashioned DIALOG searches librarians and information professionals would carry out, except that there's no blue sheet repository to explain how to search all of these datasets, even if they are all encoded at a basic level in RDF-compliant triples.

Conclusion

Citation: Moulaison, H. L. (2013). *Linked data in the library cloud: Technology of the future?* Paper presented at the 2nd International Conference on Academic Libraries (ICAL-2013), New Delhi, India, February 12-15, 2013. Retrieved from http://codabox.org/

This is an exciting time in librarianship, as academic libraries move to the cloud and re-envision how their data should be stored there and made accessible. Allowing other institutions access to library data and datasets has the potential of placing academic libraries squarely on the scholar's radar, at the nexus of possibilities for research and discovery. As libraries prepare to move into this bright future, caution should be heeded, as issues will present themselves. Planning, oversight, training, and creative envisioning of the future are all necessary as libraries move from stand-alone information silos to cloud-based information centers. Indeed, libraries are on the cutting edge of the changes that will be made possible by the semantic web, putting libraries in the challenging position of being pioneers. The rewards of placing library data on the web and of allowing patrons linked access to the whole of the linked data web that is being built, however, is worth the preliminary issues that libraries will potentially face in adjusting to the changes that the future surely has in store.

References

Anderson, J.Q., & Rainie, L. (2010, June 11). *The future of cloud computing*. PEW Internet &American Life. Retrieved from http://pewinternet.org/~/media//Files/Reports/2010/PIP_Future_of_the_Internet_cloud_computing.pdf

Berners-Lee, T. (2006, rev. 2009). *Linked data*. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J.& Lassil O. (2001, May 17). The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American.* Retrieved from http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html

Corrado, E.M., & Moulaison, H.L. (Eds.). (2011). *Getting started with cloud computing: A LITA guide.* (Guide #16). New York: Neal-Schuman. ISBN: 978-1-55570-749-1

Coyle, K. (2010). Library data in a modern context. *Library Technology Reports*, *1*, 5-13.

DBpedia data set. (last update: 2012). DBpedia. Retrieved from http://wiki.dbpedia.org/Datasets

Ford, K. (2012). LC's bibliographic Framework Initiative and the attractiveness of linked data. *Information Standards Quarterly*, *24*(2/3): 46-50.

Freitas, A. Curry, E., Gabriel Oliveira, J., & O'Riain, S. (2012). Querying heterogeneous datasets on the liked data web: Challenges, approaches, and trends. *IEE Internet Computing*, *16*, 24-33.

HTTP. (2012). In *Webopedia*. Retrieved from http://www.webopedia.com/TERM/H/HTTP.html

Citation: Moulaison, H. L. (2013). *Linked data in the library cloud: Technology of the future?* Paper presented at the 2nd International Conference on Academic Libraries (ICAL-2013), New Delhi, India, February 12-15, 2013. Retrieved from http://codabox.org/

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1997, 2009 update). *Functional requirements for bibliographic records.* IFLA. Munich: K.G. Saur Verlag. Retrieved from http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

Introduction to RDF. (2013). W3C Schools.com. Retrieved from http://www.w3schools.com/rdf/rdf_intro.asp

Miller, E., Ogbuji, U., Mueller, V., & MacDougall, K. (2012, November 21). *Bibliographic framework as a web of data: Linked data model and supporting services.* Library of Congress. Retrieved from http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf

Library Linked Data Incubator Group. (2011a, October 25). *Library Linked Data Incubator Group final report*. W3C Incubator Group. Retrieved from http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

Library Linked Data Incubator Group. (2011b, October 25). *Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets*. W3C Incubator Group. Retrieved from http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

Patton, G.E. (Ed.). (2009). *Functional requirements for authority data: A conceptual model*. IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). München: K.G. Saur.

RDF Working Group. (2004). *Resource description framework (RDF).* W3C Semantic Web. Retrieved from www.w3.org/RDF/

Statement of international cataloging principles. (2009). IFLA. Retrieved from http://www.ifla.org/files/assets/cataloguing/icp/icp_2009-en.pdf

Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, MA: MIT Press.

Tennant, R. (2002, October 15). MARC must die. *Library Journal*, Retrieved from http://www.libraryjournal.com/article/CA250046.html

W3C RDF and OWL Activities. (2013). W3C Schools.com. Retrieved from http://www.w3schools.com/w3c/w3c_rdf.asp

W3C. (1993, rev. 2006). *Naming and addressing: URIs, URLs, ...* Retrieved from http://www.w3.org/Addressing/

Zeng, M.L., Zumer, M., & Salaba, A. (Eds.). (2010, June). *Functional requirements for subject authority data: A conceptual model (FRSAD)*. IFLA. Retrieved from http://www.ifla.org/files/assets/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf

Citation: Moulaison, H. L. (2013). *Linked data in the library cloud: Technology of the future?* Paper presented at the 2nd International Conference on Academic Libraries (ICAL-2013), New Delhi, India, February 12-15, 2013. Retrieved from http://codabox.org/