Opening up Library Data on the Web: The “Linked Data” Bibliographic Model

Bonnie R. MacGregor

San Jose State University
School of Library and Information Science

Libr 284 Seminar in Archives and Records Management
Fall 2010
December 8, 2010
The web is a fascinating place in the 21st century. It has evolved into a global information space where we share knowledge and the barrier to access, retrieve, and examine that information has been toppled. The mere nature of the web exudes a generic, open, and extensible nature which has been seen as one of the web’s key features for its ability for unconstrained growth (Berners-Lee, Hendler, & Lassila, 2007). Even more exciting is the prospect that the web will evolve into a more highly sophisticated and dynamic web dubbed the semantic web - where data will be able to “provide a common framework that allows data to be shared and reused across application, enterprise, and community boundaries” (World Wide Web Consortium). In the semantic web’s nascent vision, data formats, ontology’s, and reasoning software would operate like one big application on the World Wide Web that would analyze all the raw data stored in online databases as well as data about the text, images, video, and communication(s) the Web contained (Feigenbaum, Herman, Hongsermeier, Neumann, & Stephens, 2007). The semantic web is viewed as a grassroots operation, placing the responsibility for its gradual advancement on working groups like the World Wide Consortium (W3C) which is an ad-hoc organization co-hosted by the Massachusetts Institute of Technology (MIT) of the U.S., the European Consortium for Informatics and Mathematics in France, and Keio University in Japan. Such organizations have already released semantic web languages and technologies needed to get the grassroots movement rolling. But while the semantic web is not different from the World Wide Web, its adoption will give the Web a far greater utility. As more and more groups develop taxonomies, link their schemes, and translate their terms, web software will begin to expand and be able to understand each other
automatically without the need for additional programming. While many things have been anticipated for the semantic web such as artificial intelligence and intelligence agent capabilities, one tool to have emerged from its semantic roots has revolutionary potential for library, museum, and archival data in terms of access and utilization in the 21st century. This new practice of publishing and connecting structured data on the Web is known as “Linked Data”.

The most radical change Linked Data can offer libraries and other bibliographic rich institutions is the ability to transform the way we model our data. With the current hypertext’ Web of unstructured documents, the links themselves are not very informative. Sure they are access points but they “have no meaning beyond the link” (Coyle, 2010, p. 18). It does not explain why you have linked or what the link means, so in that sense it is a very linear system, very one-dimensional, and limited. Supporters of the Linked Data movement are intent on changing this and transforming data into more structured, reusable, machine/human-readable, and interrelated information. This new data organization or Web content creation standard is engrossed in establishing relationships between data by mining the WWW and creating a new layer of meaningful information found within a network of related documents. By our data becoming of the Web we can make use of networked data resources and take advantage of the various data sources on the Web (Coyle, 2010). In accomplishing such goals the library catalog record must transform from its primal textual description to a set of data elements that can be machine-processed and compatible with the World Wide Web’s mainstream technologies. Today’s data design has to balance the functionality needed for machine processing while presenting the data in a user-friendly format. We need to create data that serves both parties
yet the functionality of our systems is failing us. We need to offer more in our systems and
tweak our library data to become more accessible, relevant, and useful for our communities in
the 21st century.

Karen Coyle has been a librarian with 30 years of experience with library technologies and
argues in “Changing the Nature of Library Data”, that while new bibliographic data models such
as the Functional Requirements for Bibliographic Data (FRBR) and new cataloging practices like
Resource Description and Access (RDA) have provided new conceptual foundations, how we
express our data in a 21st century data format is missing. “Moving toward a more open Web-
based model where records are addressable by programs and in formats that are easily
integrated into web applications and services” - needs to transpire in order for libraries to
remain relevant and innovative in this data age (Coyle, 2010, p. 14).

Libraries have created a great wealth of bibliographic knowledge. In fact, libraries were
some of the first communities to have developed ways to create mixed character sets with both
Latin and non-Latin data characters used to process text with computers (Coyle, 2010). In the
1960’s Machine Readable Catalog Records or MARC records were considered to be one of the
first text markup languages if not the first. Libraries often view their information as ‘records’ –
one that are mainly textual, static (self-contained) and able to exist in isolation without any
external dependencies (Singer, 2010). MARC allowed our textual documents to be coded in a
way that was readable by computers and available to be processed mechanically yet the
methods used to design such cataloging functions were designed for a particular era,
specifically the late 20th century where data was stored at fixed lengths and contents were
compact. Little has changed in our current metadata creation practices, even though such modeling techniques like MARC pose an impediment to our current data processing abilities. Some of the punctuations and other various information located within MARC fields cannot be manipulated because data is often duplicated which leads to inconsistent metadata being creating that eventually causes errors and the ability for our systems to relate records declines.

According to some, MARC is currently an arcane standard. In reading Roy Tennant’s- Library Journal’s Digital Libraries blog, in his October 2002 post Tennant states that MARC records have outlived their usefulness and are beginning to create problems for catalogers and users. Not only do MARC records operate in a flat data architecture but the granularity of MARC is also complex. Encoding often “lacks essential checks and balances to assure that the appropriate granularity (how finely the individual elements are chopped-is) achieved” (p.2). Fields too are often coded with a numbering scheme that is difficult to read by someone unfamiliar to its complicated syntax and MARC has begun to push library data toward technical marginalization because other professions are not using MARC, thus directing libraries into a niche market where vendors must design specialized systems while the majority of the information technology industry has already moved towards XML (Extensible Markup Language) to encode and transfer their information. “It is rather ironic, but Library data is so much more textually sophisticated than most business data that we simply cannot use standard business software in our system” (Coyle, 2010, p. 18), which makes library systems much more costly in terms of system development.

For Ross Singer in “Linked Library Data Now” Library data is a mess. Where duplication runs rampant and is often arcane, inefficient, and “where little control is exercised over relating
one resource in one system to another in a completely different application - much less data available outside the institution” (p.114). While libraries have moved onto the Web and made an effort to convert file formats to XML and Dublin Core, the ‘new’ data formats have essentially repackaged old data into new containers and don’t change the data model at its core. A more radical transformation is needed if we are to move from database managed search and display into an interactive use of library data on the web. “Linked data realizes the vision of evolving the Web into a global data commons, allowing applications to operate on top of an unbound set of data sources, via standardized access mechanisms” (Biszer et al., n.d., p. 24).

In order to accomplish this data rich layer of the semantic web that will be searchable by mining embedded text, can only be accomplished if we code library data differently. That coding scheme is Linked Data. By adopting the philosophy, strategies, and standards of Linked Data, library bibliographic data can become more machine use-friendly and will be detrimental in opening up our information silos in the information age. The basic tenets set forth by Tim Berners-Lee for Linked Data were published in 2006 and address the four principles Linked Data was to accomplish:

1. Use Universal Resource Identifiers (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.
4. Include links to other URIs so that they can discover more things.

“Linked data is more exposed on the Web than data available via application programming interfaces (API’s); crawlers can crawl it, app’s can link to it, and can be linked to other open data...its more discoverable and useable to applications that may not know how to access a
specific API” (Tennant, 2009, blog). Linked Data is about using the Web to publish and connect otherwise disconnected data through the use of the Resource Description Framework (RDF) and Uniform Resource Identifiers (URI’s) (Stuart, 2010). RDF is difficult to understand in its formal, mathematical definition yet it’s the most fundamental building block of the semantic web where relations can be established amongst various pieces of information regardless of format (Feigenbaum et. al, 2007). RDF can relate items and thus data in a way that computers and software can automatically understand and exchanges of that information can be made because all data will consist of things and relationships between them, i.e. “a thing = with a relationship to = another thing” (Coyle, 2010, p.20). RDF is basically a set of rules that controls how metadata can be manipulated on the web, and more specifically how it can be read by computer software and hardware. It set out to resolve the problems of encoding metadata for digital objects by using XML so software could retrieve, parse, and index information found on the web. While the rules are pretty straight-forward they can be combined to produce very complex working units. “It holds the potential to expose a growing ecology of data sources that can be linked to provide an enriched data environment that can be increasingly exploited in a variety of ways” (Tennant, 2009).

Another major key element of Linked Data requires identifiers to be assigned that will distinguish one thing from any other thing. It serves as a means to identify our things and relationships in a way that can be understood by machines. This is accomplished by identifiers being in the form of a Uniform Resource Identifier (URI). While they have a lot in common with Uniform Resource Locators (URL) the URI format is the preferred format to be used on the semantic web because in using identifiers, one can not only retrieve documents but also
retrieve descriptions of what is being identified, whereas URL’s just describe a location and
direct you to the document or to a site that is the location of where that document resides.

“An information object is “on the Web” if it has a URI. Objects which have URIs are
sometimes known as “First Class Objects” (FCOs). The Web works best when any
information object of value and identity is a first class object. If something does not
have a URI, you can’t refer to it, and the power of the Web is the less for that” (Berners-

If Library metadata is to have a better presence on the Web we must give our data an identity
in the form of a URI. This will allow our ‘things’ to be actionable on the Web. In assigning
identities to our metadata or ‘things’, identification has to take place on two levels, by its
properties and values. As each piece of data and any link connecting those two pieces of data
are identified by a URI they are suddenly connected and grouped together by what is called a
triple. The basic components of a triple include classes, properties, and values. A class is much
like a scientific taxonomy where all resources are grouped together because they belong
together and are based on some common characteristic. Applied to libraries, classes can be
created that are appropriate for a particular institutions metadata. Properties are elements that
describe your metadata. They work in the same fashion as relational database fields in that
they offer more detailed information. For example, if your metadata is describing “books in a
library”, you would have properties such as title, subject, accession number, etc. Finally, values
are the actual content of the data; so in continuing our example, if our library data property for
“title” is being explored I would also access the values of that property which may include
“Harry Potter and the Deathly Hallows” or “Harry Potter and the Chamber of Secrets” (Coyle,
2010). The advantage to using URI’s modeled in Linked Data’s fashion of triples allows the URI
to work in cohesion with URLs. So not only do URIs identify and return information about the ‘thing’, they also identify the location(s) to those ‘things’ through the Hypertext Transfer Protocol (HTTP) that needs no additional programming. The advantage to implementing such a data model for libraries would allow all of our resources to be conceptually independent from one another and yet be modeled as first class objects. By using RDF each resource gets their own identifier in the form of a URI, so in essence all Harry Potter books would use the same URI to refer to their author and all the properties and values that exist for a specific title would use the same URI to connect to their classes; etc. Furthermore other resources that have nothing to do with the catalog can use these URIs (Singer, 2009). “Although specialized programs would still be needed to handle specific workflows such as maintaining incenses or providing linking mechanisms into databases, using RDF/URIs will allow current library practices to adopt consistent and persistent URIs to identify their resources and deliver specific parts of the data they actually manage or pull from other sources” (Singer, 2009, p. 125). While this seems like what already exists, library discovery systems often lack powerful applications that allow the bibliographic resources to be mapped to an electronic inventory resource and make the connection in its interface (Singer, 2009). Linked data enhances a user’s experience because the discovery platform can easily retrieve and synthesize data from a wide variety of disparate source.

By creating identifiers for our data through the creation of Web Ontology Languages (OWL) which is just another way of creating controlled vocabularies subject headings, and/or authority naming conventions, OWL will facilitate “things” (library resources) to be distinguishable from other similar things and machines will have no problem knowing the
difference. Vocabularies and ontology creation are very important in establishing the semantic web language - operating in a linked data framework. When the values for a property come from a controlled list, such as a list of language codes like Friend of a Friend (FOAF) or Dublin Core, each value can be represented by either a string of text or a URI, whereas the URI would produce greater precision in retrieving data. For example, FOAF project is a grassroots decentralized social networking system where a semantic web vocabulary has been created for describing people’s names, ages, locations, jobs, and relationships. Users can post information and/or imagery in any format that seamlessly connects to all other types of social networking sites; which MySpace and Facebook cannot do because their fields are incompatible and not open to translation (Stuart, 2010). All in all this type extensibility allows users to grab all of the information about anything related and discover new resources by following the relationships expressive in the “Linked Data” framework.

Agencies are encouraged to create their own identifiers for their data by establishing controlled vocabularies. “It should be the maintenance agency that provides the identifier, and it should be given a name that only the agency can provide” (Coyle, 2010, p. 24). While no standard has been developed, the creation of URIs for some of the many controlled lists used in library catalog data have begun, including the Library of Congress that has created a site for the identification of the value lists under its control, and has defined over three hundred thousand entries pulled from the Library of Congress Subject Headings list (Coyle, 2010). Other factions involved with the development of RDA are pushing to align vocabularies with the Dublin Core Metadata Initiative and use RDF to better leverage the modeling of FRBR and FRAD components. They have already begun to map several existing MARC controlled vocabularies
and have made them available through the National Science Digital Library. Another project implementing RDF is Code4RDA which seeks to use the RDF model and create a code base that would support it. The overall goal for this project seeks to create RDFs within the RDA & FRBR schemes from existing MARC records. Catalogers would not be expected to remodel every record in existence but the code would allow for easy migration and markup. Public ontology’s are also encouraged that would describe the nature of the relationships made between data sets in Linked Data “because it is not about a centrally dictated universal standard” (Stuart, 2010, p. 6). The New York Times for example has published approx. 10,000 subject headings as linked open data.

While the mere mention of the semantic web within library circles sends many running for the hills, many groups, companies, and organizations have already begun to implement the data modeling technique. IBM, Hewlett-Packard, Nokia, Oracle, Chevron, Boeing, Pfizer, and Adobe are just some of the major companies and vendors that are promoting open source semantic web frameworks. Some of the most advanced progress is taking place in the life sciences and medical (healthcare) fields whom face tremendous data integration challenges. One initiative deployed in 2004 at the University of Texas Health Science Center in Houston, focused on how to better detect, analyze, and respond to emerging public health problems. Their system SAPPHIRE (for situational awareness and preparedness for public health incidences using reasoning engines) sought to pool a wide range of data from local health care providers, hospitals, and environmental protection agencies, that when integrated could better track reports on emergency room capacities, self-reported patient symptoms and updated electronic health records. “Semantic technologies integrated this information into a single view
of current health conditions across the area and provided real time report updates” (Feigenbaum et. al, 2007). One of the key features to how this system operates is based on the ontology that classifies and reports outbreaks of illnesses. Reports are thus produced and delivered faster, relieving nurses from manual paper-work and can be directed to treating patients. Another example where SAPPHIRE and its semantic web nimbleness came into play was during Hurricane Katrina. Data that had been collected on the ground via handheld computers that were surveying the spread of disease and illness and were uploaded into the system, while reports began to be output that succeeded in identifying outbreaks in survivors of the disaster much sooner than would have been possible before (Feigenbaum et. al, 2007). These examples are meant to showcase the extensibility of semantic web technologies and highlight how once the systems are configured for a general premise, they can quickly adapt to a variety of situations within that field, where a network of metadata can be added, recombined, and/or mashed up to produced new information when appropriated to a particular context.

But back to libraries. While library practices and Linked Data share a great deal of commonality in regards to naming authorities and controlled vocabularies there are serious drawbacks to consider, especially for librarians that involve issues of trust. Would libraries link to data that they had no control over and was in fact publicly editable? While the ‘crowd’ approach to content creation and folksonomies has been beneficial to sites like Wikipedia, a lack of commonality across broad communities becomes apparent and there exist limitations to the level of agreement that can be reached when creating these ontologies. Another concern is that in order for linked data to be “good” – good data must populate the system. But how can
we in fact monitor this and who determines what is good or correct? It also has been noted that in order for users of linked data to continue to implement projects related to library information, what is holding us back is the Library of Congress authority lists which are slow to be developed and released as well as standardization across institutional lines. Privacy is another concern for professionals and that is will be compromised in a number of ways yet most semantic web advocates insist that the protections are the same as those in the non-link world and if two databases joined by the semantic web have different privacy criteria’s, the software must honor both sets of rules or otherwise create a set that covers both (Feigenbaum et. al, 2007).

Libraries exist to serve the present and future informational needs of our users and the communities to which they belong. To serve them well in the 21st century we need to remain innovative in our facilitation of technology. In doing so, library bibliographic data modeling needs change. We must transform our data to enable it to operate openly on the web and not remain hidden within our information silos. How are we serving our communities well if we remain static access points? MARC at one time was radical change that improved our practices and processing. Another radical transformation is needed given the current digital environment; Linked Data is that change especially since we already have the metadata created. Libraries have such vast and wonderful holdings that it’s a shame they are not represented on the Web. While I am not suggesting that our bibliographic data on the linked cloud replace library catalogs or any other service, I’m merely suggesting that Linked Data is – simply put – another plug-in we can offer. Linked Data offers a logical direction and application for libraries to join the larger information community, manage their data, and maintain a
variety of sources from multiple creators. The possibilities offer our libraries an opportunity to exist in a dynamic and exciting virtual world where our data can foster discoverability and extensibility while remaining true to our historic role as providers of information. “We cannot go it alone, nor claim exclusive dominion over how useful and legitimate data is built, discovered, collected or used” (Singer, 2009, p. 121).
Bibliography


