Break On Through to the Other Side: The Library and Linked Data

by F. Tim Knight

> *Although there are a number of ontologies available to describe bibliographic data, the data contained in library systems are not generally available. The access mechanisms described in Linked Data need to be implemented for libraries to truly be "part of the semantic web".*[1]

This article will discuss the barriers that exist between our bibliographic data and other data available on the World Wide Web. The isolation of this bibliographic data is a problem that impacts on the successful integration of the library catalogue into the potential semantic Web of the future. It will look at two available data models for bibliographic data and introduces the Resource Description Framework (RDF) which has emerged as the preferred data model for enabling linked data on the Web. The article concludes with a brief look at some current activities related to linked data that are occurring as part of the development of Resource Description and Access (RDA).

The Bibliographic Record as a Data Model

The bibliographic record provides us with a view of our library data originally derived from the drawers full of cardboard catalogue cards. This data model collects all of the descriptive and analytical data elements about library resources and places them together into a single bibliographic record. Searching catalogue cards in an online catalogue environment has made it much easier to bring together and identify related resources held in the library. Presented with a standardized system of resource description, a classification scheme and some controlled vocabulary terms, a library user can fairly easily find all of the books written by a particular author or all of the resources written about a particular subject available in or through the library.

When library catalogues became accessible on the internet through the World Wide Web the ability to present access points as hyperlinks made it possible for catalogue users to click on an author's name and quickly retrieve all of the bibliographic records listing that author. In the online catalogue these links are short cuts that allow searches to be performed without returning to the search screen and manually entering an author's name. This functionality provided a useful way for catalogue users to move from one relevant record to another set of potentially related records. However, although this feature appears to free up the data and make the catalogue more web-like, the searches are still only finding what's available in the library catalogue and therefore in the library.

This is an important distinction to realize. A search for Peter Hogg in my library catalogue will retrieve only those works that our library has actually acquired or had the time to reference on the Web. It will not provide a list of all of the works that Hogg has ever written. For example, my library might own titles A and C but lack title B; title B may be held by another library, or it may be available for purchase, or as an electronic book on the Internet Archives, but the search in my library catalogue will not tell me that. This holds true for subject searches too where only books on constitutional law held in my library will be retrieved and not everything written about constitutional law.

Barriers Between Our Bibliographic Data

By its very nature the library catalogue effectively creates a barrier between the data we use to describe the resources found in our library and any other potentially useful data about these same resources that

[1] Malmsten, Martin. 2008. "Making a Library Catalogue Part of the Semantic Web", Proc. Int'l Conf. on Dublin Core and Metadata Applications, p. 149 <http://dcpapers.dublincore.org/ojs/pubs/article/view/927/923>.

exists outside of our library. In a recent webinar on the new cataloguing rules[2] Diane Hillmann spoke of the “tyranny of the records” and commented that the bibliographic record data model is “inward facing” and I think this is what she meant. The library catalogue is an inwardly focused, self contained data silo that isolates our data, and because of that it does not, and cannot, connect to data that exists outside itself.

Another problem associated with using the record as a data model is our commitment to the MARC bibliographic standards. The MARC21 standard is used to encode our data and to transfer data between library systems. This means that anyone who wants to make use of our data must know how to either handle MARC directly or be able to translate it into terms their information system can make use of. MARC has been an invaluable tool and has enabled us to share our professional cataloguing activities, but its card-based record-centric focus has also effectively locked our data in a tower and only those closely aligned with the library profession have the key.[3] Researchers and other information seekers may come across the tower when they search for information on the web, but once they have found it they will need to perform additional searches in the library database to discover what they need.

“... of the Web”

This specialized knowledge also places a wall or barrier between our data and anyone who may want to use the rich bibliographic data created and cared for by the library profession. This is a shame because the high quality of our metadata would be useful if it could interact freely with other information sources available through the Web. Ross Singer summarizes the situation well in his article Linked Library Data Now!:

> “The scope and breadth of the information universe, however, has expanded so far and so quickly that librarians can no longer claim exclusive dominion over how useful and legitimate data are built, discovered, collected, or used. The library *must* integrate itself into the weave of the Web if it is to remain relevant. ... It is neither pragmatically nor economically feasible for libraries to go it alone, so how must we adapt to this reality? The information the library contains also would be a welcome and heavily used resource if it was *of the Web* as opposed to standing apart from the rest of the information universe bridged by rickety connections into its silos, or as an island, inaccessible from the mainland.”[4] [original emphasis]

There are tools now like LibX[5] that can provide direct connections from the Web to our data but it's not the same as having our data be part *of the Web*. And, since the web has become the primary place to search for information, this is a major problem for libraries.

Our user's expectations have changed but the library, with its electronic version of those drawers of catalogue cards, really hasn't changed with them. Karen Coyle, in her recent introduction to the semantic Web and bibliographic data, holds a similar view:

> “The change that libraries will need to make in response [to today's user] must include the

[2] Hillman, Diane. 2010. “RDA Vocabularies in the Semantic Web”, ALA TechSource (the slides from this webinar The RDA Vocabularies: What They Are, How They Work are available at <http://www.slideshare.net/ALATechSource/diane-hillmann-rda-vocabularies-in-the-semantic-web>).

[3] Consider too that our current library systems still do not take full advantage of all of the capabilities of the MARC format.

[4] Singer, Ross. 2009. “Linked Library Data Now!”, Journal of Electronic Resources Librarianship, v. 21, no. 2, p. 121.

[5] LibX is a browser plugin for Firefox and Internet Explorer that provides direct access from the Web to your library's catalogue <http://libx.org/>.

> transformation of the library's public catalog from a stand-alone database of bibliographic records to a highly hyperlinked data set that can interact with information resources on the World Wide Web. The library data can then be integrated into the virtual working spaces of the users served by the library."[6]

Not until we have succeeded in making the library catalogue an actual part of the Web can we hope to regain our place as a valued information resource.

Using Search Engines to Find Information

Let's take a quick look at what *is* happening outside of the library catalogue. Well, for one thing the process of finding information on the Web is not difficult at all, or at least it is not perceived as being difficult.[7] Google provides the recipe for a very simple search process: take one search box, add a few search terms thought up in the spur of the moment and voila: thanks to Google's page rank algorithm you retrieve a few hundred thousand fairly relevant results. This is, of course, absolutely fantastic. But unlike a library catalogue attempts to find all of the resources written by a particular author can be difficult and unreliable. And because there is no classification system and no terms controlled by a standard thesaurus it is especially difficult to retrieve a comprehensive set of relevant resources on a given subject.

Keyword searching can be a very powerful tool but without predictable data fields to aid in focusing our searches retrieving the most relevant results is challenging. Without that predictability the burden of dealing with the idiosyncrasies of natural language (e.g. synonyms, homonyms, etc.) and the volume of information is shifted away from the information system and on to the researcher. And, if there is no reliable way to retrieve the most relevant resources available you can never be sure that you haven't missed a relevant information resource. In a library catalogue where, in addition to keyword searching, you can search on particular data fields, you can at least be reasonably satisfied that you have found a comprehensive and relevant set of results.

Links and Relationships

There are two fundamental problems at play here: 1) the library catalogue, with all of its wonderful metadata, stands on its own isolated from the web, and 2) the web, which although easily searched, provides unreliable search results because it lacks the characteristics of a database. But there is a third problem. Neither the library catalogue or the World Wide Web can reliably identify and reveal the numerous relationships that exist between documents and resources.

There are many, many links in web documents but no meaning associated with them. Karen Coyle makes the following observation:

> "Today's web is a web of documents that link to each other. ... The links generally go from a point in one document to another document, with some pages consisting almost entirely of links that serve as entry points to collections. The links themselves, however, are not very informative: they have no meaning beyond *link*. They do not explain why you have linked, nor what the link itself could mean. A link could be a citation that supports a quote, it could be [a] document that gives further information, or it could be critical ('Whatever you do, don't believe what this person says here!'). Note also that links go in

[6] Coyle, Karen. 2010. "Understanding the Semantic Web: Bibliographic Data and Metadata", Library Technology Reports, v. 46, no. 1, p. 5.

[7] This reminds me of a line from Thomas Mann's seminal book on library research: "... people tend to choose perceived ease of access over quality of content in selecting an information source or channel; that is, they usually follow the slope of the system **regardless of whether it is leading them to the best sources**. [original emphasis], *Library Research Models*, New York, Oxford University Press, 1993, p. 93.

> only one direction, so a document that has links to it is not aware of those links."[8]

What we are missing, and what the work leading towards the semantic Web is all about, is "meaning".

Lack of Meaning

In a law library "meaning" might look like this. When searching for a particular legislative act one might expect to find, in addition to commentary written about the legislation, the associated regulations, or each of the readings of the bill as it went through the House or Senate, or the parliamentary debates that occurred with maybe the video or audio recordings as well as the print transcripts. Maybe we'd want to focus on a particular role that an author had, for example as a primary or secondary author in a law review article, or as the editor of a collection of continuing legal education papers, or the commentator on a particular area of the law. But if these roles have not been identified by the information system we will not be able to search for them.

And then there are any number of other meaningful relationships between resources that we may not even be aware that we want. Robert Darnton offers a nice image when speaking about his archival research:

> "The manuscripts seem to stretch into infinity. You open a box, take out a folder, open the folder, take out a letter, read the letter, and wonder what connects it with all the other letters in all the other folders in all the boxes, not just in this repository but in all the archives everywhere."[9]

Our information systems should make these connections or at least allow researchers to add and share the necessary bread crumbs that make meaningful connections as they do their research.

Breaking Records

There are other ways to model data that does not use a record as the core feature. The record, for example, can be broken apart separating out the individual data elements. When a search is performed the individual data elements can be gathered up and presented to the user. Diane Hillmann delivered a great presentation last year at the American Association of Law Libraries annual meeting describing this very idea. She considered the future of our bibliographic environment and asked the following questions:

"What Would Happen If …

- We stopped thinking about our data as 'records'
- Instead, we started thinking of our data as 'statements'
- We started thinking of these statements as able to be aggregated in a variety of ways, for a variety of purposes
    - Including sharing with others, both within the library and beyond"[10]

These are great questions and hit on an important consideration necessary for the successful use and integration of our bibliographic data in the future. Instead of a bibliographic record a series of independent data statements from many sources could be brought together as part of the search process. This can allow connections between resources and across information communities to be made more

[8] Ibid., p. 18.

[9] Darnton, Robert. 2009. "Lost and Found, in Cyberspace" *in* The Case for Books, p. 61.

[10] Hillmann, Diane. 2010. "How Do We Get There From Here", Annual Meeting of AALL, slide 34. <http://www.slideshare.net/smartbroad/aall-denver-2010>.

easily bringing things together in ways that depending on our information needs at the time.[11]

RDF Data Statements on the Web

The data model that has emerged through the ongoing work on the semantic Web is the Resource Description Framework (RDF). RDF provides the means to create and use data statements like those Hillmann refers to above. In RDF these data statements are expressed as subject--predicate--object and because of the three parts have come to be known as RDF triples. In her very useful book on metadata for librarians, Priscilla Caplan talks about the importance of RDF:

> “RDF is key because it allows metadata to be represented as assertions that can be specified wholly in terms of URIs, or links to the definitions of the subject, object and predicate of the assertion. For example, RDF can make the assertion that document A (subject) is created by (predicate) John Smith (object), where the object is represented as a link to an authority record or document portion identifying Mr. Smith, the predicate is represented as a link to the creator element in the Dublin Core Metadata Specification, and the subject is represented as a link to the document.”[12]

We'll look at what Caplan calls “URIs” in a moment, but let's first examine this RDF data structure a little more closely.

If we draw on the Peter Hogg example used earlier we could write the following data statement or RDF triple:

Peter W. Hogg → isAuthorOf → Constitutional Law of Canada
subject *predicate* *object*

This subject--predicate--object expression is equivalent to the information found in our library record view:

Author:	Peter W. Hogg
Title:	Constitutional Law of Canada

These statements can be considered equivalent because the human reader can interpret the information displayed in the catalogue and based on experience conclude that Hogg is the author of the book. In fact both of these examples are data expressions that would make sense to a human reader. However, the beauty of RDF, and the goal of the semantic Web is to make statements that can also be “interpreted” and “understood” by software applications.

URIs and URLs

To accomplish this in RDF a Uniform Resource Identifier (URI) is substituted for each part of the RDF triple. A web application can use the information represented by the URI to process the data assigning a level of understanding to each part of the data statement. This process is also known as making the data “machine actionable.” For example, a machine actionable version of the above RDF triple using URIs might look like this:

http://viaf.org/processed/LAC|0103C0296 ⟶ http://dublincore.org/2010/10/11/dcelements.rdf#creator ⟶ http://www.carswell.com/description.asp?docid=2690
subject *predicate* *object*

[11] Hillmann provides a nice visual representation of how this might work in the library context in her presentation on slides 35-48 and I encourage you to take a quick look at how she represents this.

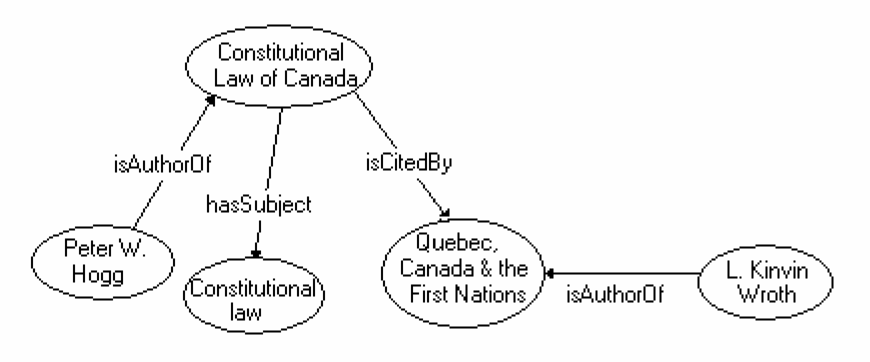
[12] Caplan, Priscilla. 2003. “Metadata and the Web” *in* Metadata Fundamentals for All Librarians, p. 52.

To the human eye this statement does not impart much meaning although we do recognize that each of the URIs here are URLs.[13] If we were to click and follow these URLs we would learn that this expression draws from the Virtual International Authority File (VIAF)[14] [i.e. the URI for Peter W. Hogg], the Creator element form the Dublin Core Element Set[15] [i.e. the URI representing the concept isAuthorOf], and the Carswell catalogue [i.e. URI for the book Constitutional Law of Canada]. And in theory this statement is also something a machine can understand and interpret without ambiguity. We'll leave the concept of URIs behind for now and continue with RDF using human readable examples.

RDF is a Graph Model

The other feature of RDF is that it is described as a "graph model" which means that the data statements can be illustrated conceptually using a visual representation. For example, some other triples that might be associated with Hogg's book could include:

> Peter W. Hogg → isAuthorOf → Constitutional Law of Canada
> Constitutional Law of Canada → hasSubject → Constitutional law
> Constitutional Law of Canada → isCitedBy → Quebec, Canada and the First Nations: The Problem of Secession
> L. Kinvin Wroth → isAuthorOf → Quebec, Canada and the First Nations: The Problem of Secession



These RDF triples could be represented in an RDF graph something like this:

Finding the trail that leads from Hogg's Constitutional Law of Canada to L. Kinvin Wroth's article is not something readily available from within a library catalogue. But if our data were described using RDF triples these kinds of relationships would begin to reveal themselves on the Web.

As more information communities become aware of the potential that this data model brings and begin to openly contribute their data on the Web in this way it won't be long before we will be browsing data as well as documents when we search for information.

> "... there is a growing community of people and organizations who have metadata available to them that they have structured using Semantic Web rules. These disparate sets of data can be combined into a base of actionable data. These sets of data are being referred to as 'linked data,' and the Linked Data Cloud is

[13] Uniform Resource Locators are a type of URI used on the Web.
[14] Virtual International Authority File <http://viaf.org/>.
[15] Dublin Core Metadata Element Set, Version 1.1 <http://www.dublincore.org/documents/dces/>.

> an open and informal representation of compatible data available over the Internet. New linked data is being added to the cloud daily. Each new resource that is added to the Web in this format increases the number of data connections possible between existing data sets."[16]

As the number of RDF statements or 'assertions' grows a cloud of links and relationships will emerge[17] and new connections will be possible between resources that would be unrealizable in any other way. This is the power of linked data and our library data can make an important contribution to this web of relationships.

There has been some progress made toward realizing the possibility of using linked data in the library environment. Leading this work are the efforts of the DCMI/RDA Task Group[18] who have begun to register the RDA data elements and vocabularies with the NSDL Metadata Registry so that URIs exist for RDA.[19] Outside of the library the W3C continue to champion the semantic Web and linked data and you will find useful information and resources on their website.[20] Raising awareness in the library community of the potential of linked data and RDF as a data model is an important next step to ensuring the successful integration of the library catalogue into the potential benefit of the semantic Web of the future.

[16] Coyle, p. 28.

[17] For an illustration of this see the Linking Open Data cloud diagram <http://richard.cyganiak.de/2007/10/lod/>.

[18] The Dublin Core Metadata Initiative/RDA Task Group grew out of work on the development of RDA, see their wiki at <http://dublincore.org/dcmirdataskgroup/>.

[19] See the RDA Vocabularies <http://metadataregistry.org/rdabrowse.htm>.

[20] The W3C is the World Wide Web Consortium an "international community that develops standards to ensure the long-term growth of the Web". For further information library standards and linked data see <http://www.w3.org/2005/Incubator/lld/wiki/Library_standards_and_linked_data> on linked data in general see <http://www.w3.org/standards/semanticweb/data> and I recommend the RDF Primer as useful document for getting better acquainted with RDF <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>.