Chapter 1

Library Data in the Web World

Abstract

Library data has been designed to be read and interpreted by the librarians and users who are the end users of the catalog. Today's data, however, needs to be managed and interpreted by computers and integrated into myriad applications that are part of the growing web of services on the Internet. In particular, the Semantic Web technologies being developed put a new emphasis on linking data from disparate sources. To be part of the linked data network, the library world needs to transform its catalog records into true data.

> In many respects, the most important question for the library world in examining semantic web technologies is whether librarians can successfully transform their expertise in working with metadata into expertise in working with ontologies or models of knowledge. Whereas traditional library metadata has always been focused on helping humans find and make use of information, semantic web ontologies are focused on helping machines find and make use of information. Traditional library metadata is meant to be seen and acted on by humans, and as such has always been an uncomfortable match with relational database technology. Semantic web ontologies, in contrast, are meant to make metadata meaningful and actionable for machines. An ontology is thus a sort of computer program, and the effort of making an RDF schema is the first step of telling a computer how to process a type of information.
>
> –Eric Hellman[1]

As is always the case in a time of transition, it may be possible to see where we have come from, but it is very difficult, perhaps impossible, to know where we are going. This report should thus be accepted as one moment in the path of moving target. This is how it looks to me today, and tomorrow is a different story.

When I talk about library data and the semantic web, people ask me if I really think that the Semantic Web (note the case change) and RDF are "the answer." I don't. In fact, I have no more idea of what "the answer" is or could be than most people. I do think that the move toward an open declaration of vocabularies and the freeing of data from databases and even from records is key to expanding the discovery and navigation services that we can provide to information seekers. I have no reluctance in taking from the Semantic Web movement that which seems to benefit libraries without taking in the whole. Perhaps I should have written this entire report without using the "S– W–" words, but it would have been awkward to do so purely from a view of sentence construction. When I say "Semantic Web," try to understand that I mean a set of evolving techniques for presenting data in a way that could be used on networks; those networks could be the World Wide Web or a new form of library user tool.

The work to define the data elements of the new cataloging rules, Resource Description and Access (RDA), in a Semantic Web–compatible format would not have happened without the interest of the members of the Dublin Core Metadata Initiative, who have been dipping their metadata toes into the waters of Semantic Web thinking for a number of years. It also would not have happened without the interest shared by members of the Joint Steering Committee for RDA, in particular Barbara Tillett

(Library of Congress) and Tom Delsey (RDA Editor), who attended the meeting where it all happened. Diane Hillmann (Metadata Management Associates) and Gordon Dunsire (University of Strathclyde) were given the dubious honor of managing the project, and, along with Jon Phipps (Metadata Management Associates) and myself, have completed the recommended tasks that came out of the 2007 meeting referenced above. It is worth reproducing here the report from that decisive meeting, against which our work can be measured.

The picture of data, and of library data in particular, has changed considerably in the two and half years of the project. There has been a co-evolution of RDA and the Semantic Web. The only thing we can know for sure is that the evolution will continue. Please keep that in mind as you read on.

Library Data

Library data has been designed to be read and interpreted by librarians and users. Although there are some controlled data fields, most of what is in the library catalog entry is text. The emphasis is on the human user, even though the data today is stored in computer systems and displayed on a screen. The machine as user has not gotten a great deal of attention in the library cataloging environment.

Now there's yet another potential user of library data, and that user is the Web and services that function on the Web. We know that our users go to the Web to do their research, to interact with other people, and to create their works. If we are to serve our users, then we need to deliver library services to users via the Web. But delivery over the network is not enough; our services must not only be *on* the Web, but need to be *of* the Web. The services can not just pass through, but must live and interact on the Web. With Web-based data, we can use the vast information resources there to enhance our data by creating relationships between library data and information resources. This will not only increase opportunities for users to discover the library and its resources, but will also increase the value of the data by allowing its use in a wide variety of contexts. If you take the view that information has value when it is used, then greater use means greater value.

Time of Opportunity

In 1837, the British Museum found itself without a printed catalog of its books. This fact became a great opportunity that was seized upon by Sir Anthony Panizzi. It was for the creation of this catalog that he developed the "Code Panizzi" consisting of ninety-one rules for the cataloging

The following are notes from a data model meeting held at the British Library in London from April 30 to May 1, 2007[2]

A meeting was held which examined the fit between *RDA: Resource Description and Access* and models used in other metadata communities.

Participants:

- Tom Baker
- Robina Clayphan
- Tom Delsey
- Gordon Dunsire
- Diane Hillmann
- Alistair Miles
- Mikael Nilsson
- Andy Powell
- Barbara Tillett

Recommendations:

The meeting participants agreed that RDA and DCMI should work together to build on the existing work of both communities.

The participants recommend that the RDA Committee of Principals and DCMI seek funding for work to develop an RDA Application Profile -- specifically that the following activities be undertaken:

- development of an RDA Element Vocabulary
- development of an RDA DC Application Profile based on FRBR and FRAD
- disclosure of RDA Value Vocabularies using RDF/RDFS/SKOS

Outcomes:

The benefits of this activity will be that:

- the library community gets a metadata standard that is compatible with the Web Architecture and that is fully interoperable with other Semantic Web initiatives
- the DCMI community gets a libraries application profile firmly based on the DCAM and FRBR (which will be a high profile exemplar for others to follow)
- the Semantic Web community get a significant pool of well thought-out metadata terms to re-use
- there is wider uptake of RDA

Further suggestion:

The meeting further suggests that DCMI and DC Application Profile developers consider the value of using conceptual models such as FRBR as the basis for describing intellectual or artistic creations.

of books.[3] Thus, modern library cataloging was born.

We also find ourselves in a time of opportunity—not because we lack a catalog but because the cataloging community has stepped back to rethink its work. The Functional Requirements for Bibliographic Records (FRBR) and Resource Description and Access (RDA) provide new models and new rules,[4] and they come at a time when the way data is stored and managed has resulted in an entirely new technology with which we can distribute our catalog entries and make them available to users. That technology is the World Wide Web, and more specifically the burgeoning use of the linked data standard to facilitate interconnections between information resources. The Web provides a platform for linking information resources regardless of their provenance.

Neither FRBR nor RDA was developed to meet the linked data standard, but the FRBR model uses entities (things) and relationships, which is conceptually similar to the basic concepts of the Semantic Web. We are in the fortunate position of having a good model for the transformation of our data to this more modern standard.

Before setting out some steps that we can take to further this transformation, it may help us to look at our current data models and systems, with an eye to identifying those areas that are functioning today as barriers to full and open use of the great store of library metadata.

Linked Data website
http://linkeddata.org

Library Data Today

Library metadata has its purpose in the creation of the catalog. In fact, metadata creation is called "cataloging"—the development of a catalog. The catalog, which was originally physical but is now digital, uses database technology in a stand-alone system. Internet and Web access to the catalog is through a tunnel from the network to the database interface residing on the library system.

The catalog supports many library management functions: inventory control, collection development, acquisitions, new materials check-in, budget management, and many others. It also serves user functions such as circulation of materials, account management, and placement of hold requests. But the public thinks of the catalog primarily in its role in discovery, identification, and delivery of data. The discovery component, however, is used less and less as information and document seekers find that the Internet gives them a broader view of the information space and satisfies their needs more readily than the library catalog. Catalog uses by information seekers are an increasingly small percentage of discovery actions.[5]

Library catalog data could, however, be the connection between the library and the knowledge space on the Web. The library catalog data could be a source of quality bibliographic information for many user tasks like managing bibliographies, sharing with colleagues, and making connections between library and nonlibrary resources. For this to be the case, however, the library's bibliographic metadata needs to be "of the Web."

Bibliographic Control

Library cataloging has historically been all about getting control over the bibliographic universe, knowing exactly what works and editions a library holds, and making sure that all items in the library catalog are uniformly described. One of the discussion points of the Task Group on the Future of Bibliographic Control was the question the use of the term *bibliographic control*. The group defined the term as "the organization of library materials to facilitate discovery, management, identification, and access."[6] The group also said, however:

> The phrase "bibliographic control" is often interpreted to have the same meaning as the word "cataloging." The library catalog, however, is just one access route to materials that a library manages for its users. The benefits of bibliographic control can be expanded to a wide range of information resources both through cooperation and through design. The Working Group urges adoption of a broad definition of bibliographic control that embraces all library materials, a diverse community of users, and a multiplicity of venues where information is sought.[7]

In this statement, the Working Group opened up the possibility that in the future bibliographic control may be more than what we think of today as cataloging and may take place beyond the confines of the library catalog. They confirm this with another statement:

> The future of bibliographic control will be collaborative, decentralized, international in scope, and Web-based. Its realization will occur in cooperation with the private sector, and with the active collaboration of library users. Data will be gathered from multiple sources; change will happen quickly; and bibliographic control will be dynamic, not static. The underlying technology that makes this future possible and necessary—the World Wide Web—is now almost two decades old. Libraries must continue the transition to this future without delay in order to retain their significance as information providers.[8]

There is much here to frighten anyone who hopes that the library catalog will need just some minor tweaking to keep up with modern times. It's pretty clear, though, that the group was defining bibliographic control to mean

something quite different from the creation of library catalogs as we know them today. Their expanded definition and the emphasis on the World Wide Web as the appropriate platform for reaching today's users greatly broaden the role that library metadata will have to fulfill.

Bibliographic Data on the Web

There is an increasing use of bibliographic data on the Web in general, through services like Google Scholar and Google Book Search, Wikipedia and Wikimedia, LibraryThing, Open Library, and others. Some of these directly import library metadata, others create their own. The content of these sites is not limited to bibliographic data; they use bibliographic data within an information context. In some cases, the object of the bibliographic metadata is the focus of the document or page; in other cases, it serves as a pointer to other resources. In either case, though, it is clear that resources cataloged by libraries are part of the online information landscape. That landscape, however, does not make use of the MARC record, and there is no unifying standard for bibliographic data. There also is no concrete way to link data on the Web with the many instances of that data in library catalogs. Where specific item or record identifiers, such as ISBN or OCLC number, are available, it is often possible to link through WorldCat to library holdings, but that is a viable option only for OCLC member libraries and also doesn't provide links from other data elements to the bibliographic data. It is a partial implementation of integration of library bibliographic data to the Web, but only partial.

The idea that library metadata will be used widely on the open Web changes the meaning of cataloging: cataloging will no longer be limited to the creation of records for the library catalog, but will serve other functions as well, and users who may never directly make use of the library catalog. This is a true expansion of the role of library data, to the point where it can be used for any bibliographic function. However, the effort of cataloging need not increase: instead, the sharing of data can increase, and with some forethought the act of cataloging can draw on cooperative data sources. To be sure, redesign of cataloging systems will be needed.

Data in Records

Library metadata is, and has always been, a complex concept with many different points of information. Both technically and in terms of information content, the library record must be used as a whole. The record provides the context for each data element, and holds together all of the fields that describe a particular manifestation. A field taken out of this context would not be meaningful. A field like the following is not useful because it is only within the record that we know to what book it refers:

260 $a New York : $b Viking Penguin, $c 1994

The exact same information can be designed to have meaning both within a record and independently. This is done by providing explicit relationships and identifiers for the subject of the description:

New York → is place of publication of → Raintree County

Viking Penguin → is publisher of → Raintree County

1994 → is date of publication of → Raintree County

This form of data allows processing on individual data elements within their meaningful context. It creates more possibilities for machines to act on the data. For example, you may wish to know the earliest date of publication of a book, at least the earliest that the library has. With this type of data organization, it is possible to ask for all of the dates of publication for the book in question, and to receive the following as an answer:

1948

1994

Note that where the above examples have meaningful words ("Raintree County," "is publisher of"), the actual data would have identifiers for those terms and concepts. This is because the words themselves could be ambiguous ("Raintree County" is both a book and a film, and each would need its own identifier), and in any case they are not globally unique. Someone else could develop a term "is publisher of" that has a meaning slightly different from the one that I am using. The unique identification of *things* and *relationships* assures that data can mix with other data without losing its specific meaning.

All of this facilitates machine processing, of course, but it also potentially provides some new capabilities for user interfaces as well. It should be much easier to create a function that will find other editions of the same book without requiring that the user perform a search. For example, if the user has entered the catalog from a link on a professor's reading list and all copies of the book are checked out, it should be possible to expand the search to the most recent other edition of the book. In a library catalog today, the user would need to perform a search and either read through a list of retrieved items or limit by date. With the catalog record reorganized in a linked data format, this becomes an automated offering, not a user task. The main reason to organize our data as separate and unambiguous data elements, however, is to allow that data to be used outside of the context of the library catalog and to be combined with other data.

Fortunately, the data in library records is coded in some detail, so a transformation from the record-based model to a data-based model is feasible. The whole of the information about a single item may still be wrapped together as a record, but the data within that record will be usable in many different contexts.

Database as Container

The closed system model used by libraries is related to the dependence on a record-based model. This has been the dominant model for all forms of data, not just in libraries. In a database management system, data is stored in a highly controlled environment with specific functions allowed to different categories of users: those who can modify the database management structure (the system administrators), those who can modify records (the catalogers), and those who can enter only through the user interface (the users). Regardless of how rich the data, users of this system can perform only actions that are offered to them through the interface.

In addition, the data in the database cannot easily interact with data outside of the database. Libraries have added some links to Web services and are able to import book covers and reviews from other sources, but dynamic interaction is difficult to achieve. It can even be impossible to link into the library catalog from outside, thus limiting users' ability to make reference to items held in the library. This means that library data cannot participate in the highly linked and linkable information environment on the Web, and this limits the visibility of libraries to Web users.

The Search as Discovery

Users go to the library catalog to search for items in the library collection. To conduct a search, they must have something in mind: an author, a title, or a topic. Searching is a familiar first step in information seeking, but it is not the only way, and perhaps not even the predominant way, that users find the information they need. In real life—that is, "offline"—friends, colleagues, and mass media are common leads to information sources. Online, social sites have become powerful meeting grounds where users ask questions, find recommendations, and pursue a wide variety of interests. While it may be difficult to think of these actions as "information retrieval," they do provide users with a great deal of information. Just because a person stumbles upon an interesting site or reads a book that is recommended by a friend does not mean that no information has been exchanged. These informal sources of information are not at all new: many experts cite other members of their profession as their main source of information and commonly begin an investigation in a new area by contacting a colleague who is already expert in that area.

Offline, we rely on a web of human connections to help us find information. Online, that web consists of links between resources and rich social interaction that help us select and evaluate resources. The search itself is only a beginning. The library catalog, however, offers little beyond search and retrieve. Navigation is generally limited to clicking on headings, which then performs another search. The catalog, therefore, serves only limited information-seeking behavior. It is no wonder that few users report that they begin their information searches in the library catalog and that the use of library catalogs is minor compared to the use of the Internet.

It is unlikely that searching will be eliminated from our information discovery toolkit, but we can expect that navigation capabilities will become increasingly available as more and more information moves to the Web. Today we follow links found within documents, but the Semantic Web promises even richer navigation possibilities, as well as the ability to actually pose questions to the Web, treating it much like a database of information.

Moving Forward

Libraries already have the key elements for a modern metadata definition: there is a general model of the library domain provided by the FRBR entities, relationships, and attributes; there is a statement of goals in the FRBR user tasks; and a detailed set of data elements, vocabularies, and guidance rules exists in library cataloging standards, most recently in RDA. The FRBR model and the cataloging rules are coherent with each other to the extent that RDA assigns data elements to the FRBR entities.

Both FRBR and RDA are realized as documents, which means that they are presented as human-readable concepts, not as computer code. In their document forms, neither can be acted on by computers, and neither can be moved seamlessly into the Web. It may not even be possible to turn them into code without some significant changes. But the use of entities and relationships gives this whole that is FRBR + RDA some basic conceptual compatibility with the technology that is developing for the realization of the Semantic Web.

One of the first steps that needs to be taken is to tease out the many components that are encompassed by the RDA text. RDA is not a single unit but in fact a combination of

- the elements of bibliographic description
- the relationships between those elements and between areas of description
- the rules for deciding what data will be used to describe bibliographic items

All of this is wrapped up together in our catalog-

ing rules, which makes it very hard to turn them into a machine-usable set of elements. The RDA development committee did create a list of data elements and relationships, but it pulled these out of the text; it didn't build the text on them. Machines, however, will act on the data elements, not on the explanatory text, so it will be necessary to look at the data elements carefully to discover any areas where the creation of machine processing may not fit in with what is written in the text. This is because there are concepts you can create in text that you cannot automate directly. When creating text, it is hard to know when you are relying on human intelligence to make leaps that a computer cannot. The guidance rules and the data structures need to be developed together if the machine processing is going to be successful. There is going to have to be some back and forth between data structures and rules for decision making to be sure that we've covered both the human and the machine needs.

These functions need to be teased apart in order to create Semantic Web–compatible data. This is because the Semantic Web requires certain information about data elements. Some of this information may be inherent in the RDA text, such as a rule that a cataloged item will have only one preferred title, or that dates may be entered in a structured format. All of this information needs to be made explicit in the definition of the elements for use in machine processing.

Steps to Linked Data

There are four basic steps that one needs to take to enter into the world of linked data, data that can play well on the Semantic Web. The first is to design the basic data model. We have that already in the form of the definitions of functional requirements in FRBR, FRAD, and others. These models will undoubtedly undergo some evolution as library data and the data environment change. The second is to define the data elements (or, in Semantic Web parlance, *properties)* of our metadata. Part of this process is making those definitions available on the Web in a machine-actionable format. The third major step is to define all of our controlled lists in a linked data compatible format and to also make those available on the Web for anyone to use and to provide definitions and display capabilities.

1. Define the Model

We talked above about the FRBR model and the user tasks that have guided its development. A great value of the family of functional requirements is that they begin to define the entities that our metadata addresses: bibliographic resources, agents, topics. At this level, we can see some similarities already with linked data standards being developed elsewhere. For example, an early Semantic Web project is "Friend of a Friend" (FOAF), a metadata format for persons that can be used in social network situations.[9] It is not identical to the way that libraries define persons, but it points to an area where data could be exchanged among different communities.

2. Define Data Elements

This step is similar to the initial creation of a database structure, where you define all of your data elements. Each data element will need to be defined according to certain requirements posed by the Semantic Web concepts. The Semantic Web view of data differs from that of a database, so developers will need some level of learning and adjustment.

3. Define Vocabularies

One of the great advantages that we have in transforming our catalog data to the Semantic Web is that we have already made much use of controlled vocabularies. These help greatly in communicating with other communities because we can clearly delineate the possible meanings of certain elements through the finite vocabulary list that the element can carry.

Vocabularies can be simple lists of terms, but it is also possible to define each term in a vocabulary with a unique identifier. Identifiers are less ambiguous than language and also often create links back to the identifying agency and documentation about the term. For example, the term *green* can mean different things: in the context of politics, it may indicate an approach toward environmental issues or even the name of a political party. In a shoe catalog, it may be the color of the product. The use of an identifier provided by the entity that has developed the vocabulary means that each of these meanings will have a different identifier.

4. Develop Application Rules

There is another level of definition that will usually be undertaken, although it is not strictly required by the Semantic Web, and that is the creation of application rules or an application profile. Application rules generally add constraints to your metadata, such as whether your element will be mandatory or optional, and rules for repeatability. For example, although you may define the element for title in your list of elements, yet in your actual application you may wish to limit the use of the title to one per description. Or you may wish to say that in your application the title is mandatory. The definition of these rules in a machine-readable form will allow others to understand the output of your applications.

The remainder of this report will illustrate these steps in greater detail.

Notes

1. Eric Hellman. "Can Librarians Be Put Directly onto the Semantic Web?" Go to Hellman blog, Aug. 4, 2009, http://go-to-hellman.blogspot.com/2009/08/can-librarians-be-put-directly-onto.html (accessed Dec. 14, 2009).
2. British Library, "Data Model Meeting," www.bl.uk/bibliographic/meeting.html (accessed December 14th, 2009.
3. Anthony Panizzi, "Rules for the Compilation of the Catalogue," *Catalogue of Printed Books in the British Museum,* vol. 1 (London: 1841), v–ix.
4. IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report,* Sept. 1997, as amended and corrected through Feb. 2009, http://archive.ifla.org/VII/s13/frbr/frbr_2008.pdf (accessed Dec. 14, 2009); Joint Steering Committee for Development of RDA, "RDA: Resource Description and Access," www.rda-jsc.org/rda.html (accessed Dec. 14, 2009).
5. Cathy De Rosa, *Perceptions of Libraries and Information Resources: A Report to the OCLC Membership* (Dubin, OH: OCLC Online Computer Library Center, 2005), 1–17.
6. *On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control* (Washington, DC: Library of Congress, 2008), 6; available online at www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf (accessed Nov. 6, 2009).
7. *On the Record*, 10.
8. *On the Record*, 4.
9. Dan Brickley and Libby Miller, "FOAF Vocabulary Specification 0.97," The FOAF Project website, http://xmlns.com/foaf/spec (accessed Jan. 5, 2010).