Chapter 1

Introduction

Abstract

Chapter 1 is the Introduction to this issue of Library Technology Reports. *Library activities in recent years, particularly those that have contemplated the future of bibliographic control, have given libraries a theoretical basis for the move from traditional catalogs to the Web as a data platform. Library catalogs have already evolved to federate resources from external locations and to bring in data from remote sources. FRBR, RDA, and the commitment of the Library of Congress to a new bibliographic framework all point libraries in the direction of shared, linked data.*

The world today is clearly not that of our library predecessors, of Dewey and Cutter, not even of Lubetzky or even Gorman. The changes that have taken place since the introduction of the personal computer and the globalization of communication over the World Wide Web are huge, and they affect in particular anyone involved in knowledge research and creation. Let's do a brief environment scan.

Today's world is highly computerized; we have computers on our desks, in our pockets, and even in our pens. It isn't, however, the cold, oppressive, machine-like computerization of 1950s science fiction. The amount of user interaction is astonishing, even ignoring the edge cases of the teenagers who send hundreds of text messages a day or the constant Tweeters. Users expect to have an effect on their virtual world, even if that only means creating their own Facebook page. Facebook is in fact a prime example of user-created content, along with Flickr, YouTube videos, blogs, and Tweets.

The spread of global computing has overthrown some of our previous assumptions about institutions and power. The do-it-yourself information system can be personal and political, as we have seen with efforts like WikiLeaks and with the fact that some bloggers now have the public's ear in a way that was once reserved for network news celebrities.

Today's resources are either becoming digital or already born digital. This is not the end of print, but print is definitely becoming a format of the past. Few college students today would consider print to be a modern technology, although many undoubtedly still find it useful at times. Digital resources are relatively easy to find (through keyword searching) and to obtain (because they are online), but they can be hard to use because they require specific technical skills. They also require devices that add a new cost to the use of information. Users are dependent on software tools that are not controlled by the library and may not work as well as is desired. Access often means obtaining a copy that the user then needs to manage.

Yet today's users expect to work independently without prior instruction; they have come to expect the single search box as their interface.

The entire concept of communication is changing. Communication is increasingly remote and asynchronous, not face-to-face and in real time. There is a rise of new media (especially video) over old (text): increasingly, instruction is provided in video format rather than textual documentation.

The conversation has become faster and shorter. Academics refer to books as a "slow conversation" that takes place over time and space, but a conversation today is more likely to be fast and short. Even a blog post is considered lengthy when Twitter becomes your norm.

We are capturing and storing many information events that were once transitory, from conversations

(podcasts) to conference presentations (streamed and restreamed). Communications that were once considered too informal for fixation are now part of the permanent or semipermanent record of our civilization. E-mail, which was once thought of as being no more "official" than an office-cooler conversation, is now frequently considered evidence in courts of law, as are images captured by bystanders on phones at the scene of a crime. We have almost lost our ability to be off the record, and the great increase in that record haunts the information professionals who dare contemplate the need to treat it as we have treated our more formal information products in the past.

In the face of these changes, it is obvious that libraries have to modify many of the processes that were put in place before the computer became a ubiquitous tool.

Some Library History

It may seem that the Semantic Web and its linked data technology have come blasting out of nowhere to disrupt metadata activities everywhere. This is especially true in the library world with its well-established and relatively stable metadata practices. But change rarely happens suddenly, and this change is no exception. Libraries have been struggling for years with many of the same issues that have prompted the development of linked data, although different terms have been used in the library environment.

The move from analog to digital materials was not the beginning of the change. Previous technologies, like film and disk recording, already presented challenges for libraries that wished to maintain strong bibliographic control over those materials. These technologies allowed replication of content in different formats, which led to a discussion that many of us remember as the "multiple versions" problem, or "mulver." The rapid increase in new formats and duplication of already cataloged materials proved very difficult to fit into the standard library data record known as the Machine Readable Cataloging (MARC) record. The problem, though, was only partially with the record format: the basic model of a catalogable unit had been formed in a more stable time when the book was a typical information product. The tension brought on by the increase in format change became even more intense as materials moved from analog to digital and the physical stability that had been the foundation of library cataloging began to melt away.

Economics played and continues to play an important role in the need for change. In the very early days of the World Wide Web, libraries actually contemplated creating catalog records for selected Web resources. Even though these records were simplified versions of the ones created for traditional resources, it soon became clear that the Web would grow too fast for the creation of human-crafted metadata. There is now a visible speedup of all forms of information resources, even those that are ostensibly in traditional off-line formats, and doubts are growing about the ability of libraries to afford the costs of hand-hewn bibliographic control today and in the future.

Linking and Federating

The library catalog has been losing its walls for a while. The ability to federate searches across disparate data stores, some owned by the library and some licensed for use, has allowed the library to provide results that include a mix of library-cataloged materials and materials from outside the catalog database.

Competition from websites, and in particular from sites like Amazon, whose content overlaps with library catalogs, has led libraries to move away from plain text displays, mainly by linking. Catalogs began bringing in cover images from external sources to add interesting visuals to their displays. While not technically difficult, this was a significant change, as libraries for the first time began pulling outside content into their catalog. This was a noteworthy departure from the previous concept of the separate and highly controlled library catalog. Catalogs now link to tables of contents from a variety of sources, to biographical information about authors, and even to reviews.

The OpenURL is an ingenious way to link from external resources back to library catalogs and licensed resources. From a data source that is clearly not the library catalog, the user is offered a link directly to local library resources.

What these technologies have proven possible in the context of library systems can be extrapolated to the idea of libraries on the Web. First, federated searching could combine library resources with Web resources in search and display. If nothing else, we have learned how to create displays that combine different types of data, and our users seem to navigate them without great difficulty. Next, we know that we can enhance the user experience by linking out to select Web-based resources. These resources may not be one hundred percent reliable, but the risk is not unmanageable. Libraries have formed trust relationships with information providers, a proof that linking does not have to be entirely uncontrolled or open. And finally, we are already seeing the advantages of moving discovery

of library materials out beyond the library catalog to other environments where the user is searching and interacting.

This evolution of library catalogs is like a dress rehearsal for moving library data from its storage silo in library systems and databases to the Web of linked data. Library data will link to select other data sources in order to provide more value and services for users. Other users and resources will be able to link to library data, thus making library data discoverable from a variety of points in Web space. The user view will not be a single view controlled by the library but will blend into the user's current environment, with links to the library from anywhere the user is searching or working. As information creation moves to the "cloud," so will library services, not because libraries create their own cloud but because there will be no separation between libraries and the Web.

The Future of Bibliographic Control

While library catalogs have been experimenting and evolving, sometimes in informal environments, the library world has been making official investigations into the future of bibliographic control. In 1998, the International Federation of Library Associations issued the Functional Requirements for Bibliographic Records (FRBR), a rethinking of the nature of bibliographic description.[1] In 2000, the year of the Library of Congress Bicentennial, the library held the Bicentennial Conference on Bibliographic Control for the New Millennium to address major challenges facing the cataloging community.[2] This was followed in 2007 by the report of the Working Group on the Future of Bibliographic Control.[3] Then, in 2008, Resource Description and Access (RDA), the new cataloging rules aligned with FRBR, was issued.[4]

There are some common threads running through these efforts. One was an awareness that library information needs to be more data-friendly. FRBR produces a three-step plan to move from the flat catalog record of MARC to an entity-relationship model that is more conducive to machine processing. The Library of Congress efforts had an underlying theme of finding ways to make cataloging more efficient so as to address the increase in the rate of knowledge production. The 2007 Report on the Future of Bibliographic Control urged the Library of Congress to leverage the existing Web environment to give libraries greater visibility and to take advantage of the robust technology that the Web has developed.[5]

None of these efforts directly mentioned the emerging technology of the Semantic Web, linked data. The connection between the Semantic Web and library data was made in May 2007 at a meeting at the British Library. That meeting was the result of discussions between persons involved in the Semantic Web, the Dublin Core Metadata Initiative (DCMI), and the Joint Steering Committee for Development of RDA (JSC). At the meeting, an agreement was reached between the participants that working together, representatives of the JSC and DCMI would publish the RDA element set and terms lists in Semantic Web formats.[6] There were a number of reasons why this step came about at that moment. To begin with, RDA's basis in FRBR brings it closer to the Semantic Web data format than previous domain models, such as ISBD, have been. Secondly, RDA did not yet have a record format, so there was no need to consider existing data. This made experimentation, including the inevitable trial and error, feasible. Thirdly, with the possibility that RDA would be the future of library cataloging, it made sense to begin a new era in library data with a radically new data format.

Over the next two years (and continuing to this day since metadata is a constantly evolving organism), all of the RDA elements and term lists were coded as RDA properties and value vocabularies, respectively, in the Open Metadata Registry, using the domain name selected by the JSC: rdvocab.info. These elements are now legitimate entities in the Web of linked data and are being used in a variety of bibliographic datasets on the Web.

Around this same time, the Library of Congress began to publish some of its key authority files in Semantic Web format, beginning with the Library of Congress Subject Headings. While use of these appears to still be limited to the library community, as we will see in chapter 4, a web of library authorities, both subjects and names, is beginning to grow.

Perhaps the most significant step in this process was that of testing the RDA cataloging rules, an activity that took place in 2010 in libraries in the United States and elsewhere. Many useful insights were gained from this test, but the key one for the purposes of this report was the confirmation that the flat MARC 21 record format would not accommodate the use of relationships between bibliographic entities that both FRBR and RDA internalize. Based on this conclusion, the Library of Congress announced the Bibliographic Framework Transition Initiative in May 2011.[7] In that announcement, LC stated that careful attention would be paid to Semantic Web technologies, and in the plan put forth in October 2011 LC stated:

> The new bibliographic framework project will be focused on the Web environment, Linked Data principles and mechanisms, and the Resource Description Framework (RDF) as a basic data model.[8]

This statement is a culmination of the prior work and thinking about the direction of library catalogs and cataloging. It could not have come about without the interest in next-generation catalogs, without the entity-relationship model of FRBR, and without the need for new cataloging rules as expressed by RDA. It is a statement of evolution, not revolution.

The remainder of this technical report will, I hope, give some idea of the linked data context in which library data can find its place. It begins with a brief description of some Semantic Web terms and concepts that will be useful through the following chapters. I hope that those chapters show convincingly that linkable data is available that will facilitate the creation of a library data format that is truly "of the Web."

Notes

1. IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report* (Munich: K. G. Saur, 1998).
2. Ann M. Sandberg-Fox, ed., *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web* (Washington, DC: Library of Congress), 2001.
3. *On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control* (Washington, DC: Library of Congress, 2008).
4. Joint Steering Committee for Development of RDA, "RDA: Resource Description and Access," last modified December 9, 2010, www.rda-jsc.org/rda.html.
5. *On the Record*.
6. British Library, "Data Model Meeting," accessed March 8, 2012, www.bl.uk/bibliographic/meeting.html.
7. Deanna Marcum, "Transforming Our Bibliographic Framework: A Statement from the Library of Congress," May 13, 2011, last modified June 16, 2011, www.loc.gov/marc/transition/news/framework-051311.html.
8. Deanna Marcum, *A Bibliographic Framework for the Digital Age* (Washington, DC: Library of Congress, 2011), 4, www.loc.gov/marc/transition/pdf/bibframework-10312011.pdf.