The Online Catalog as Data Repository

Nathan Rupp

Introduction

Libraries in the early twenty-first century provide a number of tools with which their users can locate library resources. These tools include systems which provide access to print resources, digitized versions of print resources, and born-digital resources. The most common method of providing access to print library resources is, and has been for the last quarter century, the library's online catalog. The library's online catalog can be defined as a database of bibliographic records in machine readable format. In contrast, the library's Online Public Access Catalog, or OPAC, is used to provide library patrons with access to the database of bibliographic records. The first OPACs were telnet-based text interfaces; more recently, they have taken the form of graphical user interfaces available via the World Wide Web. These OPACs have enabled library users to access networked resources the library has cataloged and made available.

Despite the ubiquity of online catalogs, a number of emerging trends have caused information professionals to question the continuing usefulness of these tools. First, libraries have begun to provide access to an increasingly wider variety of resources—adding electronic texts, online journals, learning objects, geospatial data, and other digital resources to their traditional mix of books, journals, and audio-visual materials. Second, a number of library competitors such as Amazon and Google have emerged whose interfaces seem to offer some marked

improvements over the library's OPAC.[1],[2] Third, a number of metadata standards such as Dublin Core and the Metadata Object Description Schema (MODS) have been introduced in addition to the one on which most online catalogs are based, MARC21.[3],[4],[5]

It may be beneficial for libraries to re-examine their OPACs in the light of the development of other user interfaces, such as those at Amazon and Google. Indeed, some libraries are doing just that. For example, the Catalog User-Interface Platform for Iterative Development (CUIPID) developed at the River Campus Libraries at the University of Rochester uses data from a number of different sources, including the library's online catalog, to support a re-imagining of the OPAC.[6] Yet, at the same time, others are going one step further—they are suggesting that MARC, the metadata schema used in the online catalog, be discarded.[7],[8] They point out that MARC is inadequate for describing some of the complex new resources that libraries are collecting, like learning objects; that MARC-based online catalogs do not support some of the more recent features of library competitors' systems; that MARC is a proprietary format and "is at odds with open systems"[9]; and that MARC is neither as easy to use nor as flexible as

[1] Amazon.com, http://www.amazon.com/ (8 Mar. 2008).
[2] Google, http://www.google.com/ (8 Mar. 2008).
[3] DCMI, *Dublin Core Metadata Initiative*, http://www.dublincore.org/ (8 Mar. 2008).
[4] Library of Congress, *Metadata Object Description Schema (MODS)*, http://www.loc.gov/standards/mods/ (8 Mar. 2008).
[5] Library of Congress, *MARC Standards*, http://www.loc.gov/marc/ (8 Mar. 2008).
[6] David Lindahl and Jeff Suszczynski, "CUIPID Project: Catalog User-Interface Platform for Iterative Development," paper presented at forum of the Metadata Working Group, Cornell University, Ithaca, New York, February, 2005, http://docushare.lib.rochester.edu/docushare/dsweb/Get/Document-18040/CUIPID Project.ppt (8. Mar. 2008).
[7] Roy Tennant, "MARC Must Die," *Library Journal* 127, no. 17 (15 Oct. 2002): 26, 28.
[8] Dick R. Miller, "Bibliographic Access Management at Lane Medical Library: Fin de Millennium Experimentation and Bruised-Edge Innovation," *Cataloging & Classification Quarterly* 30, no. 2/3 (2000): 139-166.
[9] Ibid, 163.

some of the newer metadata standards. Considering these factors, one might conclude that the MARC-based online catalog should be discarded as well. This conclusion, however, may be premature; the MARC-based online catalog may still have a role to play in providing access to the increasing variety of resources collected by libraries.

A Brief History

MARC was originally developed at the Library of Congress in the 1960s so that LC could share catalog data with other libraries, not necessarily to enable the location of library resources, even though the library catalog cards created from MARC records served this function. It was not until the late 1970s and early 1980s that online catalogs began to be developed, allowing users to discover and locate materials; MARC was used as the bibliographic metadata standard in these systems.[10] Meanwhile, libraries, which had been providing access primarily to print materials, began providing access to other types of materials as well; these included audio-visual materials, maps, and many types of networked resources, including digital texts, electronic journals, and geospatial resources. Although MARC and AACR2 were updated and expanded to describe and provide access to networked resources, these updates and expansions often seemed inadequate; these difficulties in updating MARC and AACR2 may have stemmed from the fact that the standards were originally developed to describe print resources, not digital ones. By the turn of the twenty-first century, some library professionals

[10] Henriette D. Avram. *MARC: Its History and Implications* (Washington: Library of Congress, 1976), 2-4.

claimed that MARC had outlived its usefulness and that libraries should begin considering replacing MARC with other metadata standards like MODS or Dublin Core.

Online Catalogs as Data Repositories

As the number and types of information resources have increased, library users have begun to use other tools in addition to the OPAC to find them. These additional tools include systems external to the library like commercial search engines, as well as other systems developed within the library but not related to the OPAC, like Cornell University Library's FindArticles/Find Databases/Find e-Journals suite of services.[11],[12],[13] These other systems all use various types of metadata to provide access to information resources; these include bibliographic, administrative, and preservation metadata. Online catalogs are vast reservoirs of bibliographic metadata; for example, Cornell University Library's catalog contains over five million bibliographic records.[14] Bibliographic metadata is now being delivered by a number of other library systems, and it would be inefficient to store this bibliographic metadata separately in each one of the library systems. Libraries need to begin viewing MARC-based online catalogs as repositories of bibliographic metadata that can be combined with other types of metadata from other systems to support new types of library access tools. The encoding scheme for online catalogs need not be MARC, but the encoding scheme already *is*

[11] CUL Gateway: Find Databases, http://encompass.library.cornell.edu:20028/ (8 Mar. 2008).
[12] CUL Gateway: Find Articles, http://encompass.library.cornell.edu:20028/(8 Mar. 2008).
[13] CUL Find e-Journals, http://erms.library.cornell.edu/ (8 Mar. 2008).
[14] A query of Cornell's Voyager database on April 20, 2005 showed that there were approximately 5.1 million bibliographic records in the catalog.

MARC and it may be advantageous for libraries to work with it since it has been in development for nearly forty years. It may be more advantageous to determine ways in which an existing metadata schema—one that has proven quite extensible—can be used to meet the needs of other digital library systems designed to assist library users in locating information.

There are a number of reasons why it makes sense for libraries to continue to store bibliographic metadata in MARC-based online catalogs. First, libraries have made major financial investments in these systems; integrated library systems, of which online catalogs are a major part, have an initial cost anywhere from $72,000 to over $300,000;[15] in addition, the cost of maintaining those systems on a yearly basis is not inconsequential. Since the majority of library system vendors are still basing their products on MARC, to go with a non-MARC based system would mean that a library may have to develop a system on its own. Local development of such a system would be an additional, and potentially risky, investment of significant resources, and libraries have been moving away from the development of such systems to using what is available in the marketplace.[16] Even the implementation of an open-source library management system would involve a significant investment, and with many libraries already in the process of developing library systems to provide access to digital library resources, libraries should consider whether or not they really want to spend the additional resources

[15] Marshall Breeding, "Migration Down Innovation Up," *Library Journal* 129, no. 6 (1 Apr. 2004): 46-50+.
[16] Lib-web-cats ("A directory of libraries throughout the world"), http://www.librarytechnology.org/libwebcats/ (8 Mar. 2008). A search of lib-web-cats reveals that over the last decade, Penn State University, Stanford University, UCLA, and the University of Georgia have all switched from locally developed library systems to ones developed by library vendors.

necessary to re-develop a system for storing bibliographic data when one already exists. Second, added to the costs of purchasing and developing library systems are the costs that have been involved in creating the metadata stored in MARC online catalogs. As noted earlier, MARC was created to share bibliographic metadata; if it weren't for utilities such as OCLC and RLIN and their repositories of bibliographic metadata—stored in MARC—individual libraries would have to create this metadata themselves. With cooperative cataloging and the use of data from these utilities, libraries have greatly reduced the amount of money spent on the creation of original bibliographic metadata. Third, the MARC-based cataloging module is often just part of the larger integrated library system. If libraries were to move away from a MARC-based online catalog to one based on another metadata schema, they might “orphan” that module from the rest of the integrated library system. Ensuring the interoperability of a non-MARC based cataloging module with the rest of an integrated library system might be another large expense. As can be seen, much time and money has been invested in the creation and support of library MARC-based online catalogs; moving to non-MARC-based systems would not only be costly but would ignore a substantial investment that has already been made.

Non-traditional Library Finding Tools

Before moving away from MARC-based online catalogs, libraries should consider repurposing MARC-encoded bibliographic metadata to support other library systems. There are a number of ways in which this can be done. One example is the generation of title- or subject-sorted, web-accessible lists of

electronic journals, requested by many library patrons. These lists are automatically generated either by scripts developed in-house or by commercially available electronic resources management (ERM) systems, but in both cases, MARC-encoded data from the online catalog is used.[17]

Another method of re-using MARC-encoded data is to use it in digital library projects that provide full-text access to library resources on the Web. These projects can be quite complex, utilizing—in addition to bibliographic metadata—rights and structural metadata to describe the information objects they contain. Standards such as the Metadata Encoding Transmission Standard (METS) have been developed to tie together all the metadata associated with an information object.[18] These digital library projects pull together the metadata that describes or administers the information objects from a number of different sources. For example, rights metadata can come from a database that tracks copyright clearance, while bibliographic metadata can be extracted from the online catalog. Cornell University Library has been successful in extracting bibliographic metadata from its online catalog for use in a number of digital library collections, including the Home Economics Archive: Research, Tradition and History (HEARTH), the Core Historical Literature of Agriculture (CHLA), and the Making of America projects.[19],[20],[21]

[17] David Banush and Nathan Rupp, "Staying Afloat in the Sea of e-Journals: An Automated Process for Cataloging Electronic Serials," paper presented at the 2004 EndUser Meeting, Chicago, Ill., April 2004.
[18] Library of Congress, *Metadata Encoding and Transmission Standard (METS)*, http://www.loc.gov/standards/mets/ (8 Mar. 2008).
[19] Albert R. Mann Library, *Home Economics Archive: Research, Tradition, History (HEARTH)*, http://hearth.library.cornell.edu/ (8 Mar. 2008).
[20] Albert R. Mann Library, *Core Historical Literature of Agriculture (CHLA)*, http://chla.library.cornell.edu/ (8 Mar. 2008).
[21] Cornell University Library, *Making of America*, http://cdl.library.cornell.edu/moa/ (8 Mar. 2008).

In recent years, OPACs have been unfavorably compared to systems provided by online retailers such as Amazon. Amazon's system has a number of features that many OPACs do not, such as cover images, user recommendations, full-text searching, and suggestions for other titles about similar topics.[22] While the inclusion of user recommendations might be inconsistent with the traditional descriptive, rather than prescriptive, nature of academic OPACs, other Amazon-like features, such as suggestions for other titles about similar topics, would probably be welcome additions. However, most library system vendors have not updated the design of their OPACs to compete with the systems of online retailers. This can be partly attributed to the fact that the MARC-based catalog is not structured to provide some of the same content provided in the systems of online retailers. As Roy Tennant of the California Digital Library has observed, "Although it is possible to smash the table of contents into a MARC record . . . it's not pretty. By its very nature, MARC is flat, whereas a table of contents is hierarchical."[23] To provide access to these types of features, OPACs could be developed that pull together data from a number of sources: tables of contents from one source, topic suggestions from another, and the bibliographic metadata from still another—the online catalog. While some of these sources would have to be developed, the source for bibliographic metadata would not.

Another way to provide access to library resources is through a hierarchical display linking various iterations of a resource together, rather than

[22] Amazon.com, http://www.amazon.com/ (8 Mar. 2008). Amazon enables searchers to "write online reviews," "explore similar items," and "search inside the book."
[23] Tennant, "MARC Must Die," 26.

the more typical sequential listing of those iterations. The Functional Requirements for Bibliographic Records (FRBR) model developed by the International Federation of Library Associations and Institutions (IFLA) provides the theoretical groundwork for this.[24] The FRBR hierarchy consists of four different elements, or levels: work, expression, manifestation, and item. For example, Shakespeare's *Hamlet* is considered a "work." Various "expressions" of *Hamlet* could include the written play itself or a cinematic version of the play. "Manifestations" of the written play could include the version edited by Peter J. Smith and Nigel Wood and published by Open University Press in 1996 and the version edited by Harold Bloom and published by Chelsea House in 1990. A library could have two copies of the Bloom edition; each one of these would be an "item." Structuring these different versions of *Hamlet* in a hierarchical manner and showing their relationships to one another may enable library users to more easily identify the version they are looking for. Although some difficulties have been encountered in mapping MARC to FRBR,[25] at least one library systems vendor has introduced a FRBR interface,[26] and the staff of some libraries have begun extracting the bibliographic metadata from their traditional MARC-based catalogs and presenting it to the user in a FRBR-like fashion.[27] In both cases, the MARC metadata from the online catalog is being used.

[24] IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report* (Munich: Saur, 1998), http://www.ifla.org/VII/s13/frbr/frbr.pdf (8 Mar. 2008).
[25] Knut Hegna, Eeva Murtomaa, "Data Mining MARC to Find: FRBR?" http://folk.uio.no/knuthe/dok/frbr/datamining.pdf (8 Mar. 2008).
[26] VTLS, "VTLS Announces First Production Use of FRBR," http://www.librarytechnology.org/ltg-displaytext.pl?RC=10714 (15 Mar. 2008).
[27] Lindahl & Suszczynski. *CUIPID Project*.

In addition to e-journal lists, digital library systems, Amazon-style catalog interfaces, and FRBR organizational tools, libraries can provide access to resources through visualization tools that allow users to locate library resources via visual displays. For example, the *D-Lib Magazine Concept Space* "automatically generates the terms and their semantic relationships representing relevant topics covered in the corpus of a digital collection"—the articles in *D-Lib Magazine* itself.[28],[29] Stanford University's Highwire Press, an online tool for producing online versions of scholarly content, uses *TopicMap*, which is "a special Java applet designed to display standardized topics and subtopics in a graphical form that provides a 'sense of context' while navigating a large, tree-structured database."[30],[31] Other visualization tools include the Hierarchical Interface to LC Classification project (HILCC)[32],[33] and Virtual Book Spine Viewer.[34] Although most OPACs do not use visualization tools, libraries have begun experimenting with projects that extract bibliographic metadata from the online catalog and map it to the schema of another system which uses a visualization tool.[35]

[28] Junliag Zhang, Javed Mostafa, and Himansu Tripathy, "Information Retrieval by Semantic Analysis and Visualization of the Concept Space of *D-Lib® Magazine*," *D-Lib Magazine* 8, no. 10 (October 2002), http://www.dlib.org/dlib/october02/zhang/10zhang.html (8 Mar. 2008).
[29] Gerry McKiernan, "New Age Navigation: Innovative Information Interfaces for Electronic Journals," *Serials Librarian* 45, no. 2 (2003): 88.
[30] Highwire Press, *TopicMap*, http://highwire.stanford.edu/help/hbt/index.dtl (8 Mar. 2008).
[31] McKiernan, "New Age Navigation," 100-101.
[32] Adam Chandler and Jim LeBlanc, "Exploring the Potential of a Virtual Undergraduate Library Collection Based on the Hierarchical Interface to LC Classification (HILCC), http://ecommons.library.cornell.edu/bitstream/1813/2223/2/HILCC-LRTS-Preprint.pdf (15 Mar. 2008).
[33] Stephen Paul Davis. "HILCC: A Hierarchical Interface to Library of Congress Classification," *Journal of Internet Cataloging* 5, no. 4 (2002): 19-49.
[34] Naomi Dushay, "Visualizing Bibliographic Metadata – A Virtual (Book) Spine Viewer," *D-Lib Magazine* 10, no. 10 (Oct. 2004), http://www.dlib.org/dlib/october04/dushay/10dushay.html (8 Mar. 2008).
[35] Chandler and LeBlanc, "Exploring the Potential of a Virtual Undergraduate Library Collection Based on the Hierarchical Interface to LC Classification (HILCC)."

Implementing the Online Catalog as Data Repository

There are a number of ways to provide access to library resources beyond the library's OPAC. Yet, in each case, the bibliographic metadata that supports the OPAC can also be extracted from the online catalog to support these other systems. To realize the MARC-based online catalog as a data repository supporting numerous library systems in addition to the OPAC, libraries need to involve themselves with a number of activities, some of which they are currently doing and others that would be new enterprises. They need to explore schemes and tools for extracting bibliographic metadata from the online catalog and converting it to the forms used by other systems, as well as tools for relating bibliographic metadata from the online catalog to other types of metadata from other systems. Libraries also need to develop systems for recording these schemes and tools since they may be reused in multiple projects. Lastly, since one of the main components of all these systems and tools is the data in the online catalog, libraries will need to continue to provide and expand upon existing mechanisms for systematically maintaining the online catalog.

Before being converted into a form that can be manipulated and loaded into other library systems, bibliographic metadata must first be extracted from the online catalog. This is often easier said than done. For example, while Cornell University Library staff utilize a number of tools, including Microsoft Access, VgerSelect, and Harvest, to interact with catalog metadata, these tools merely enable them to report on or analyze the data. They do not extract entire metadata

records that can be converted to other metadata schemes for use in other systems.[36] Some tools have recently been introduced, however, that enable digital library developers to easily retrieve entire metadata records—or a set of entire records—from the online catalog. For example, the SRW/SRU (Search/Retrieve by Webservice or Search/Retrieve by URL) protocol, "designed to be a low barrier to entry solution to performing searches and other information retrieval operations across the internet," enables this.[37]

Once bibliographic metadata has been extracted from online catalogs, there are a number of tools available for converting it into forms used by other library systems. The Library of Congress has developed schemes to convert MARC metadata into an XML format, making it more interoperable with other XML-based metadata schemas like MODS and METS.[38] The FRBR Display Tool, based on analysis done by the Library of Congress' Network Development and MARC Standards Office, "transforms the bibliographic data found in MARC record retrieval files into meaningful displays by grouping the bibliographic data into the 'Work,' 'Expression' and 'Manifestation' FRBR entries."[39] Cornell University librarians have created mappings and scripts to repurpose MARC-encoded metadata for use in digital library systems; for example, the library created a local plan for mapping MARC elements to Dublin Core elements.[40]

[36] David Banush, "Raiders of the Lost MARC: Mining the Voyager Database for Fun and Profit," *Backstory* 1, no. 1 (2004), http://www.library.cornell.edu/cts/backstory/v1n1/raidersfeature.htm (8 Mar. 2008).
[37] Rob Sanderson, "A Gentle Introduction to SRW," version 1.1, 12th January 2004" http://srw.cheshire3.org/docs/introduction.html (15 Mar. 2008).
[38] Library of Congress, MARC in XML, http://www.loc.gov/marc/marcxml.html (8 Mar. 2008).
[39] Library of Congress, *Displays for Multiple Versions from MARC 21 and FRBR*, http://www.loc.gov/marc/marc-functional-analysis/multiple-versions.html (8 Mar. 2008).
[40] Dublin Core Mapping Group, Cornell University Library, *Cornell University Library MARC to Dublin Core Crosswalk*, http://metadata-wg.mannlib.cornell.edu/forum/2002-09-20/CUL_MARC_to_DC_Crosswalk.htm (8 Mar. 2008).

Extracting bibliographic metadata from the online catalog and repurposing it for use in other systems are just two steps in the process of using the metadata to describe information objects. That bibliographic metadata often needs to be tied together with other types of metadata for complete descriptions of information objects. The most familiar standard used for this purpose is METS, which provides a means to record not only an object's bibliographic metadata, but also its administrative metadata and the files that comprise it. METS also provides a mechanism for structuring the metadata and tying it together into a single metadata "package." One of the most useful features of METS is that it can either contain the metadata itself, or else point to a metadata source that is external to the METS record. METS' ability to point to external metadata sources would be of use in tying together bibliographic metadata from an online catalog with other metadata; the bibliographic metadata could still "live" in the catalog but be a part of the METS record.

As the number of digital library systems and the tools and scripts for converting MARC-based bibliographic metadata into forms that can be used by those systems proliferates, libraries will need to organize those tools and scripts so that they can be re-used by others working on similar projects. This could include the creation of a metadata repository which would store the tools used in every step of the metadata mapping, scripting, and transformation process.[41]

[41] Martin Kurth, David Ruddy, and Nathan Rupp, "Repurposing MARC Metadata: Using Digital Project Experience to Develop a Metadata Management Design," *Library Hi Tech* 22, no. 2 (2004): 153-165.

Such a repository could be open to numerous institutions for sharing tools for converting MARC-encoded data into other schemas.[42]

If the online catalog is to be the source of bibliographic metadata for a number of library projects, systems, and resource location tools, there must be a way to ensure that the catalog data is of good quality. Existing library cataloging efforts like authority control can help ensure data quality; this is one feature which has been built into online catalogs. Other metadata schemas such as Dublin Core are not accompanied by a host of supporting structures such as authority control, encoding standards, and maintenance agencies. In much the same way that bibliographic metadata should not be recreated if it already exists in the catalog, authority control mechanisms should not be duplicated in other digital library systems if they already exist as part of the MARC-based online catalog. On the other hand, while authority control is built into most online catalogs, error checking is not. Although there are efforts within library technical services departments to ensure quality control, work needs to be done in creating automatic error checking mechanisms for library cataloging clients. This would help to ensure more consistent checking of catalog data for errors and ensure that good catalog data is maintained, so that errors do not cause problems down the line when the data is repurposed.

[42] Michael Pelikan, Nathan Rupp, and Jeff Young, "Designing a Metadata Management Repository," paper presented at the Digital Library Federation Fall 2004 Forum, October 2004, http://www.diglib.org/forums/fall2004/pelikanruppyoung1004.htm (8 Mar. 2008).

Conclusion

Although a number of library professionals have suggested that the traditional online catalog is nearing extinction, it may still have some life. Rather than solely being viewed as an access tool for locating materials within the library collection, the online catalog should also be viewed as a data repository from which bibliographic metadata can be extracted for other library projects. The initial purpose for which MARC was developed—enabling libraries to share bibliographic metadata—will continue, as will additional purposes for which MARC has been used since its introduction, including providing access to library materials. Neither one of these efforts is insignificant. In addition, much of the author, title, topical, and location information that is described by MARC in online catalogs will continue to be essential for identifying library resources. The need to identify library resources in this manner is *not* going to disappear; even Amazon and many of the other newer bibliographic information systems described here provide access to their catalogs or collections using these descriptive elements. Rather than discarding a large repository of bibliographic metadata that has been created at no small investment of time, money, and effort, libraries should leverage that repository to its fullest potential. Library resources are not infinite—in fact, libraries are constantly being asked to do more with less. Rather than spending limited resources to develop a new bibliographic metadata schema or storage mechanism for their metadata, libraries may be better off spending their resources on new mechanisms and tools that can reuse the bibliographic metadata they already have.