A Bibliographic Framework for the Digital Age

The Working Group of the Future of Bibliographic Control, as it examined technology for the future, wrote that the Library community's data carrier, MARC, is "based on forty-year-old techniques for data management and is out of step with programming styles of today."[1] The Working Group called for a format that will "accommodate and distinguish expert-, automated-, and self-generated metadata, including annotations (reviews, comments, and usage data."[2] The Working Group agreed that MARC has served the library community well in the pre-Web environment, but something new is now needed to implement the recommendations made in the Working Group's seminal report. In its recommendations, the Working Group called upon the Library of Congress to take action. In recommendation 3.1.1, the members wrote:

> "Recognizing that Z39.2/MARC are no longer fit for the purpose, work with the library and other interested communities to specify and implement a carrier for bibliographic information that is capable of representing the full range of data of interest to libraries, and of facilitating the exchange of such data both within the library community and with related communities."[3]

This same theme emerged from the recent test of the Resource Description and Access (RDA) conducted by the National Agricultural Library, the National Library of Medicine, and the Library of Congress. Our 26 test partners also noted that, were the limitations of the MARC standard lifted, the full capabilities of RDA would be more useful to the library community. Many of the libraries taking part in the test indicated that they had little confidence RDA changes would yield significant benefits without a change to the underlying MARC carrier. Several of the test organizations were especially concerned that the MARC structure would hinder the separation of elements and ability to use URLs in a linked data environment.

With these strong statements from two expert groups, the Library of Congress is committed to developing, in collaboration with librarians, standards experts, and technologists a new bibliographic framework that will serve the associated communities well into the future. Within the Library, staff from the Network Development and Standards Office (within the Technology Policy directorate) and the Policy and Standards Division (within the Acquisitions and Bibliographic Access directorate) have been meeting with Beacher Wiggins (Director, ABA), Ruth Scovill (Director, Technology Policy), and me to craft a plan for proceeding with the development of a bibliographic framework for the future.

Below this cover note, you will find our thoughts about the way ahead. We have identified the requirements for the new bibliographic framework, based on the recommendations made by both the Working Group on the Future of Bibliographic Control and the final report on the RDA Test.

We at the Library are committed to finding the necessary funding for supporting this initiative, and we expect to work with diverse and wide-ranging partners in completing the task. Even at

[1] *On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control,* January 9, 2008, p.24, http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf
[2] *On the Record,* p. 24.
[3] *On the Record,* p. 25.

the earliest stages of the project, we believe two types of groups are needed: an advisory committee that will articulate and frame the principles and ideals of the bibliographic framework and a technical committee that has the in-depth knowledge to establish the framework, itself.

When MARC was created in the late 1960s, early 1970s, the Library community, along with computer scientists, took a bold step that led to libraries being able to share bibliographic data. This was an extraordinary achievement in that individual libraries became nodes in a much larger network of library resources. A side benefit is that the cost of cataloging was significantly reduced. The new bibliographic framework we are aiming for will broaden participation in the network of resources, librarians will be able to do a much better job of linking their patrons to resources of all kinds (from the library and from many other sources), and costs can be better contained.

The MARC standard is responsible for the creation of millions of bibliographic records from all parts of the globe. We recognize the need to continue supporting MARC during the transition, and, most likely, for years to come as libraries determine their timetable for making a change. The amount of legacy data, though, does not deter us from taking responsible actions for the next generation of libraries and librarians. The problem has been well defined by our partners. We now turn to partners of many types to help us find a durable solution.

We are posting this general plan for your comments. Please let us know what you think. We are grateful for your interest, and we appreciate suggestions for improvement. We encourage you to post your thoughts to the Bibliographic Transition listserv:

http://listserv.loc.gov/listarch/bibframe.html

Your and others' comments will be publicly available for all to read. It is in this spirit of openness and transparency that we will proceed with the development of a bibliographic framework for the 21st century.

We would also like to solicit your recommendations for members of either the advisory or technical committees. However, in the interest of privacy, we ask that you submit these (names and contact details) by email to ndmso@loc.gov.

Links to the listserv, contact details, and all other official information, announcements, and resources related to the Bibliographic Framework Initiative are available at:

http://www.loc.gov/marc/transition/

We're excited about this transition, and we hope you are too.

Deanna Marcum
Associate Librarian for Library Services
Library of Congress
October 31, 2011

Library of Congress
Bibliographic Framework Initiative General Plan

A central activity to the Bibliographic Framework Initiative is the development of a new means for capturing and sharing bibliographic data. Included in this activity is pursuing a replacement of the MARC format as the common exchange currency for bibliographic data. This was one recommendation of the 2008 report from the Library of Congress' Working Group on the Future of Bibliographic Control, On the Record, and has been discussed in the community for a number of years. Although the format is deeply embedded in the infrastructure, changing technologies and changing resource description practices mandate a transition to a more current and forward looking data creation and interchange environment. The semantic web and related linked data model hold interesting possibilities for libraries and cultural heritage institutions. (Please see the Appendix for a brief history MARC, the issues arising from its incredible success, and LC experimentation with alternate record formats, all of which inform the following Requirements.)

Requirements for a New Bibliographic Framework Environment

Although the MARC-based infrastructure is extensive, and MARC has been adapted to changing technologies, a major effort to create a comparable exchange vehicle that is grounded in the current and expected future shape of data interchange is needed. To assure a new environment will allow reuse of valuable data and remain supportive of the current one, in addition to advancing it, the following requirements provide a basis for this work. Discussion with colleagues in the community has informed these requirements for beginning the transition to a "new bibliographic framework". Bibliographic framework is intended to indicate an environment rather than a "format".

- *Broad accommodation of content rules and data models*. The new environment should be agnostic to cataloging rules, in recognition that different rules are used by different communities, for different aspects of a description, and for descriptions created in different eras, and that some metadata are not rule based. The accommodation of RDA (Resource Description and Access) will be a key factor in the development of elements, as will other mainstream library, archive, and cultural community rules such as Anglo-American Cataloguing Rules, 2nd edition (AACR2) and its predecessors, as well as DACS (Describing Archives, a Content Standard), VRA (Visual Resources Association) Core, CCO (Cataloging Cultural Objects).
- *Provision for types of data that logically accompany or support bibliographic description,* such as holdings, authority, classification, preservation, technical, rights, and archival metadata. These may be accommodated through linking technological components in a modular way, standard extensions, and other techniques.
- *Accommodation of textual data, linked data with URIs instead of text, and both*. It is recognized that a variety of environments and systems will exist with different capabilities for communicating and receiving and using textual data and links.
- *Consideration of the relationships between and recommendations for communications format tagging, record input conventions, and system storage/manipulation*. While these environments tend to blur with today's technology, a future bibliographic framework is

likely to be seen less by catalogers than the current MARC format. Internal storage, displays from communicated data, and input screens are unlikely to have the close relationship to a communication format that they have had in the past.

- *Consideration of the needs of all sizes and types of libraries, from small public to large research*. The library community is not homogeneous in the functionality needed to support its users in spite of the central role of bibliographic description of resources within cultural institutions. Although the MARC format became a key factor in the development of systems and services, libraries implement services according to the needs of their users and their available resources. The new bibliographic framework will continue to support simpler needs in addition to those of large research libraries.
- *Continuation of maintenance of MARC until no longer necessary*. It is recognized that systems and services based on the MARC 21 communications record will be an important part of the infrastructure for many years. With library budgets already stretched to cover resource purchases, large system changes are difficult to implement because of the associated costs. With the migration in the near term of a large segment of the library community from AACR to RDA, we will need to have RDA-adapted MARC available. While that need is already being addressed, it is recognized that RDA is still evolving and additional changes may be required. Changes to MARC not associated with RDA should be minimal as the energy of the community focuses on the implementation of RDA and on this initiative.
- *Compatibility with MARC-based records*. While a new schema for communications could be radically different, it will need to enable use of data currently found in MARC, since redescribing resources will not be feasible. Ideally there would be an option to preserve all data from a MARC record.
- *Provision of transformation from MARC 21 to a new bibliographic environment*. A key requirement will be software that converts data to be moved from MARC to the new bibliographic framework and back, if possible, in order to enable experimentation, testing, and other activities related to evolution of the environment.

The Library of Congress (LC) and its MARC partners are interested in a deliberate change that allows the community to move into the future with a more robust, open, and extensible carrier for our rich bibliographic data, and one that better accommodates the library community's new cataloging rules, RDA. The effort will take place in parallel with the maintenance of MARC 21 as new models are tested. It is expected that new systems and services will be developed to help libraries and provide the same cost savings they do today. Sensitivity to the effect of rapid change enables gradual implementation by systems and infrastructures, and preserves compatibility with existing data.

Approach

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. The protocols and ideas behind Linked Data are natural exchange mechanisms for the Web that have found substantial resonance even beyond the cultural heritage sector. Likewise, it is expected that the use of RDF and other W3C (World Wide Web Consortium) developments will enable the integration of library data and other cultural heritage data on the Web for more expansive user access to information.

Regarding a general data model, developments in web modeling are currently centered on RDF, which is a W3C recommended method for conceptual description or modeling of information. RDF data can be "serialized" or "written out" in several different syntax formats, one of which is XML (eXtensible Markup Language). RDF data may be used in relational databases, which underlie most library catalogs today, just as MARC 21 records are used in most library catalogs. Triplestores, which are databases designed specifically for storing and querying RDF data, are widely and readily available, and promise to provide the library community with more options about how to store and retrieve its data in the future.

Embracing common exchange techniques (the Web and Linked Data) and broadly adopted data models (RDF) will move the current library-technological environment away from being a niche market unto itself to one more readily understandable by present and future data creators, data modelers, and software developers. It is anticipated that all of these considerations, taken together, will result in greater cost savings for libraries. For example, libraries will be able to take advantage of a broader selection of technological solutions and leverage the knowledge and skills of current and future professionals. Those professionals are, or will be, deeply conversant with more contemporary data creation, data modeling, and software development practices.

The following investigations are projected but do not preclude others that are identified in the exploration of this initiative.

Developing possible interaction scenarios within the information community.

Some modeling of interaction of information community entities and services is needed to help guide the initial development of the types of services and types of requisite metadata models for the community. Development of use cases will help to scope the boundaries of the new bibliographic framework initiative development and its interdependence with other data initiatives, for example PREMIS (Preservation Metadata Implementation Strategies) or METS (Metadata Encoding and Transfer Standard). This will focus, in particular, on scenarios in which RDF and linked data play a central role.

Developing domain ontologies for the description of resources and related data in scope.

An RDF data model addresses a particular domain, such as description metadata used by the cultural heritage institutions, through development of domain ontologies (roughly comparable to the tagging of MARC plus the interrelationship of tagged elements, such as which elements can occur together and how many times). The community has created and experimented with ontologies in several cultural heritage areas and they will be considered in this activity. For example, the Dublin Core Metadata Initiative (DCMI) abstract model is an example of a foundation, upper level model that would be important to review. Other ontologies are being used or developed by the Library of Congress, International Federation of Library Associations and Institutions (IFLA) groups, and experimenters in the bibliographic community.

Organizing prototyping and reference implementations.

This work is already underway in projects by the Library of Congress, OCLC (Online Computer Library Center), British Library (BL), and Deutsche Nationalbibliothek (DNB), among others. Experimentation with new models that are supportive of an RDF data exchange environment is needed. This activity will be informed by use case requirements and other studies noted above.

A key component of this work will be collaboration. Close collaboration with the principal MARC partner institutions -- Libraries and Archive Canada (LAC), BL, and DNB -- will be instrumental. Forums such as the MARC advisory bodies (e.g., American Library Association's Machine-Readable Bibliographic Information Committee (MARBI), Canadian Committee on MARC (CCM)) will be called upon for review and asked for input. Networks like OCLC and the vendor community whose valuable and valued services are built around MARC are needed for advice, prototyping, and development of reference implementations, which will not only help determine the practicality of specific solutions and assumed specifications but also provide sample implementations for the community. Institutions currently pioneering RDF services with library data and others will be involved and will provide valuable experience. Input from the resource description community that is involved in modeling and in development and refinement of RDA will also be important. Mechanisms will be developed to identify other active participation and prototyping as this initiative progresses.

Project timetable (preliminary)

The Library of Congress will be developing a grant application over the next few months to support this initiative. The two-year grant will provide funding for the Library of Congress to organize consultative groups (national and international) and to support development and prototyping activities. Some of the supported activities will be those described above: developing models and scenarios for interaction within the information community, assembling and reviewing ontologies currently used or under development, developing domain ontologies for the description of resources and related data in scope, organizing prototypes and reference implementations.

Appendix: A Brief History of MARC 21

The MARC 21 format was developed in the late 1960s to encode the data recorded on catalog cards into machine-readable form and thus enable the Library of Congress to communicate its cataloging data electronically to other institutions. Initially it was viewed as a carrier for cataloging data that was specified by the Anglo-American Cataloguing Rules (AACR) and earlier library rule sets.

A MARC 21 record has three basic components: the *communication format structure*, the *data element set*, and the *data*, structured according to content standards. The *communication format structure* was designed to carry data for the exchange of information between systems; data that could be used in a variety of processing environments. The communications format structure was approved as a National Information Standards Organization (NISO) (Z39.2) and then International Organization for Standardization (ISO) (2709) standard, promoting its widespread adoption and use by libraries.

The *data element set* (MARC fields and tags) identifies and characterizes the specific pieces of data within a record to support its use and manipulation. The data element set that became MARC 21, used the ISO 2709 format structure for its carrier.

The *data* is primarily defined outside of the format, both through content standards or general rule sets (e.g., AACR2 (AACR, 2nd Edition), RDA: Resource Description and Access, and others) and as the content of specific data elements (e.g., terms in a thesauri of subjects). The MARC format documentation has provided usage guidelines that reflect the International Standard Bibliographic Description (ISBD), aspects of the AACR rules, and Library of Congress (LC) practices. In a few cases the MARC 21 documentation stipulates the structure of data to better support machine processing.

While the initial focus of MARC was bibliographic data for <u>books</u>, over the following 20 years the original format developed into a family of formats and added functionality in response to the library community to enable broader applications. These include:

- The ability to describe other forms of material, incorporating serials, sound recordings, still and moving images, maps, archival material, computer software, etc.
- Accommodation of bibliographic content rules other than AACR (and its predecessors), as well as DACS (Describing Archives, a Content Standard) for archival descriptions and CCO (Cataloging Cultural Objects) for cultural objects. There are currently 39 sets of content rules that have been registered in the MARC source code list to specify a rule set used for the content of a MARC record, and others are used without registration. While some of these may be obsolete in the future and not all of these will require support, the number of registered content rules underscores the importance, adoption, and versatility of the MARC 21 format.
- Metadata beyond descriptive data that allowed institutions to support a larger range of its activities, including the Authority format, Holdings format, Classification format, and Community Information format. This development allowed MARC 21 to communicate data that supported many functions in a library or other cultural heritage institution:

search and discovery, acquisitions (including automated check-in), inventory, and circulation, and it enabled the communication of important support data such as name and subject thesauri and classification schemes. The content rules for such data vary greatly, from the NISO standard for holdings data to detailed rules that support the Dewey Decimal Classification (DDC) to local conventions.

- Coordination of data elements across functions to enable the development of integrated library systems where bibliographic, authority, and holdings data are highly interrelated.

Over the last 20 years there were several developments that made MARC 21 what it is today:

- *Format integration and simplification.* This brought the data elements for different forms of material together, eliminating special fields defined only for a particular form of material.
- *Expansions for digital resources*. Changes were made to better describe digital resources and to allow for linking of records to other resources on the Internet.
- *Integration with other national formats*. The MARC 21 formats were adjusted to enable countries with national formats to move to MARC 21 in order to take advantage of the active MARC 21 record sharing environment and the development of highly standardized integrated library systems and bibliographic networks.
- *Accommodating RDA.* Over the last few years proposals and discussion papers were prepared and changes adopted so that the MARC 21 formats could better carry data cataloged using the new RDA content rules. This work is ongoing.

Current Environment

There is widespread and worldwide adoption of MARC 21, resulting in thousands of highly developed systems that work with and/or expect MARC, including integrated library systems, networks, auxiliary services (e.g., distributors sending MARC records with books and services that massage or reformat data content to make it more consistent), etc. Extensive open source tools and systems have been built around MARC 21 to enable enhanced search, discovery, and display of MARC records. Many countries have retooled to move to MARC 21 in the recent past so they could take advantage of the record sharing in the MARC 21 environment. As a result there are over a billion MARC 21 compatible records in large and small network and local systems. The MARC tools and systems support are highly evolved and heavily used to reduce costs. In addition, users of non-AACR/RDA cataloging rules (e.g., DACS and CCO) use MARC to enable integration of their data with library data. Coordinated MARC formats such as authorities, holdings, and classification are being used for communication of specialized data.

In short, the MARC format was the most important piece of the infrastructure that developed between the 1970s and the present as it enabled development of a highly successful record sharing environment that resulted in large cost savings for libraries. Initially the environment focused around LC providing its catalog records as a service to other libraries and later included wide sharing across institutions, with networks like OCLC (Online Computer Library Center) playing a key role.

Issues Today

The MARC format has been in use for over 40 years and, while that longevity has enabled development of a very rich environment that depends on the format for data exchange, it has also created issues that make change difficult.

- A widely used format with open maintenance procedures like MARC builds up redundancies over time as cataloging rules and conventions change, and types of data are accommodated, such as transcribed data, cataloger-supplied elements, authorized forms, coded data, and textual data. Because MARC has always been maintained with sensitivity to the fact that most systems will have older records containing currently obsolete data, the longevity of the format complicates simplification efforts and reuse of data tags.
- While the format was seldom used by systems as an internal format, over time it influenced the format for record creation activities, causing thousands of catalogers to learn the format tags and subfields names.
- New ideas for the organization of data are difficult to accommodate in the format in a desirable manner because they would make data incompatible with earlier data.
- Other limitations for the format are a result of choices made in the initial development of the format.

In addition, the format is "tuned" to the data specificity of library cataloging traditions that are evolving toward newer models. These include simpler models such as Dublin Core with the expectation that less specific, textual data can be understood and used by computers, obviating the need for the detailed data identification and coded data found in MARC 21. On the other hand, there is a call for more data identification specificity in newer content rules such as RDA in order to support rich linking of data. Both point to different requirements for data exchange format standards.

Recent Initiatives

Developing XML formats for bibliographic and related data.

Newer structures such as the widely implemented XML (eXtensible Markup Language) have flexibility, almost unlimited tagging ability, and can provide user-friendly tags. In addition, XML tools are not only very mature but also continue to be developed at a rapid rate, as are auxiliary XML standards that support the XML environment such as eXtensible Stylesheet Language Transformations (XSLT) (which facilitate conversion between different XML data schemas) and XML Query (XQuery) (which provides an efficient method for finding information in XML data). Thus an alternative format structure for MARC 21 was established in recent years: MARCXML, which is simply a transformation of the ISO 2709 structure of MARC 21 to an XML structure leaving the MARC 21 tagging and coding of data intact.

Another initiative has been the development of MODS (Metadata Object Description Schema) and MADS (Metadata Authority Description Schema), which are XML schema that are highly compatible with MARC bibliographic and authority data but have somewhat simpler word-based tagging. They also address some of the organizational, extensibility, and linking needs of bibliographic data today. This effort has dealt with the limitations of MARC tagging and organization of data, while retaining much of its ability to carry rich bibliographic data that is used in our current technical environment. Partially in response to library administrator pressure for decades, MODS intentionally avoids the detailed tagging and redundancies found in MARC.

Both MARCXML and MODS were developed partly to enable MARC-related bibliographic data to be in synch with communication protocols such as SRU (Search and Retrieve via URL) and OAI (Open Archive Initiative) and formats like METS (Metadata Encoding and Transfer Standard) that prefer or require XML data records. These have given the community experience with using a different structure. MODS in RDF (Resource Description Framework) is another area where development is enabling experimentation with a different structure.

LC has also maintained transformations for the MARCXML, MODS, MADS, and Dublin Core data formats and shared them with the community, yielding valuable experience in the ease and difficulty of transformation of data between formats that may have different granularity in tagging.

Accommodating RDA. The Network Development and MARC Standards Office (NDMSO) at LC has worked with the community to integrate RDA data in the MARC 21 format. This still ongoing effort has enabled the MARC format to better accommodate data formulated according to RDA instructions. Mappings between RDA and MARC and RDA and MODS are published in the RDA Toolkit.

Making LC vocabularies available as Linked Data. LC has developed and implemented a web service for stable and trusted vocabularies which has contributed to experimentation with linked data in the community (id.loc.gov). As part of this initiative, a data model for MADS as RDF was developed that enables this linked data service to expose the rich MARC data in an RDF form. RDF is currently the preferred linked data publishing model.

Increased interest in and experimentation with RDF-based models. While still in a developmental and experimental stage, the theories and technologies of the Linked Data initiative being explored by the W3C (World Wide Web Consortium) may provide the potential to widely share our rich bibliographic and related data with communities both within and outside of the library environment. Over the last year an international group sponsored by the W3C studied the potential for library data in the Linked Data space. The conclusion was positive, and a great deal of thought went in to identifying possible use cases and recommendations.

LIBRARY OF CONGRESS
OCTOBER 31, 2011