TOPIC: LINKED DATA FOR LIBRARIES, ARCHIVES, AND MUSEUMS

LINKED DATA VOCABULARY MANAGEMENT
EUROPEANA MOVING TO LINKED OPEN DATA
LINKING LIVES END-USER INTERFACE
JOINING THE LINKED DATA CLOUD
OCLC’S LINKED DATA INITIATIVE
LC’S BIBLIOGRAPHIC FRAMEWORK INITIATIVE
KEVIN M. FORD

LC’s Bibliographic Framework Initiative and the Attractiveness of Linked Data

With the Bibliographic Framework Initiative—a community effort led by the Library of Congress (LC) and first announced in 2011—the library world has begun its transition from the MARC 21 communication formats.

The MARC format is one of the oldest data format standards still in wide use today. Indeed, the format permeates everything in the library community: it is embedded in library technology and it is embedded in the minds of most librarians, especially catalogers, who know MARC and only MARC. It is undeniably part of the library family—it is the butt of jokes; it is the topic of conversations; it is worried about; it is cared for; it is loved; it is hated—and it is hard to envision life without MARC. It is, after all, forty-five years old. Most librarians working today began their careers after MARC was born, though they may have spent the first decade or two of their careers at a safe distance from the format. Some have never known life without MARC.

In 2011, LC started the initiative to phase out this library-technology stalwart and explore replacing it with a Linked Data model. The data model would, therefore, be grounded in Resource Description Framework (RDF), about which more is said below, and, in conjunction with an RDF model, the new framework would embrace the Linked Data practices and methods with respect to sharing and publishing library data. In this way, RDF provides a means to represent the data and the Linked Data methods and practices provide a means to communicate the data, the two core and historical functions of MARC.

A brief history of MARC

The acronym stands for MAchine Readable Cataloging. The first tangible MARC project began at LC in January of 1966. The format—known as MARC I—was complete by April 1966, at which time testing to determine feasibility began. The fact that the basic format was established in a four-month period is nothing short of astonishing. Around April 1967, work was underway revising the format, now called MARC II. It was formally published the following year. The LC MARC Distribution Service also began operation in 1968. In short, the MARC format was designed, tested, and implemented in little more than two years. LC led an aggressive development cycle that included a number of instrumental partners—Indiana University, the University of Chicago, Harvard University, and the National Agricultural Library, to name but a few—working in concert, testing and reporting their results back to LC. Incredibly, the MARC format, the second version, remained essentially unchanged for thirty years. It was in 1998 that the “21”—a nod to the rapidly approaching 21st century—was appended to “MARC,” marking the occasion when LC and the National Library of Canada merged their respective formats, the USMARC format and CAN/MARC. And so, today, we speak of MARC 21.

When working with MARC, one typically refers to a MARC record, which is a set of attributes and values that together independently describe a resource. Initially that resource was a book, but the MARC format was soon extended to describe other format types such as serials, maps, music, still images, and many more. In the mid-1970s these different format types went through some consolidation from which was born the more encompassing MARC Bibliographic Format. Following that, the MARC Authority format was formally published in 1981 (though LC had its authority data in an LC INTERNAL MARC format since about 1973); MARC Holdings format followed...
a few years later. MARC Classification has existed since 1991. These various formats are collectively referred to as the MARC communication formats. Although structurally the formats adhere to the ISO 2709 standard (also standardized as ANSI/NISO Z39.2), official since 1973, each communication format employs its own codes and conventions to identify and describe the data elements within a record.

During the past forty-five year period, systems have been developed that permit catalogers to create bibliographic, authority, and classification records directly in the MARC format. In other cases, systems are at least capable of constructing and deconstructing MARC records even if the internal storage structure does not itself reflect the MARC 21 format. Programmers have written software to manipulate one or a group of MARC records, often by transforming the data from one format to another. Additional programs have been written that perform endless quality checks and other statistical analyses of MARC data. Libraries have all types of personnel who can look at and understand a MARC record as readily as they can read this sentence.

All of this is to say that the MARC format has had a long and productive life for the library community. By every measure, MARC has been a success, but it is antiquated when compared to our ability to model and store data in the second decade of the 21st century.

The attractiveness of Linked Data
MARC was designed for the representation and communication of bibliographic and related information in machine-readable form. Any replacement to MARC must be capable of performing those two functions: representation and communication. The knowledge, principles, practices, and technologies that have been developed for and that exist in support of Linked Data, and the accompanying movement, have made Linked Data a promising avenue of exploration.

In its barest form, Linked Data is about publishing structured data over the same protocol used by the World Wide Web and linking that data to other data to enhance discoverability of more information.

Though not an absolute requirement, it is expected that the data conforming to Linked Data principles will be described using a very simple, but powerful, data model called the Resource Description Framework (RDF). Borrowing an analogy from English grammar, the parts of an RDF statement can be equated to those found in a basic linguistic sentence, which must contain a subject, verb, and, optionally, an object. In the case of RDF, the subject is a uniquely identified concept or thing (preferably with an HTTP URI, a uniform resource identifier), about which the statement is made. The other two parts are called the predicate (like a verb) and object. The predicate—also, identified with a URI—records the relationship between the subject and object. The object of the statement may be identified with a URI or it may be a string. It is possible to describe a thing or concept fully by making a number of RDF statements about it (see Figure 1).

From Figure 1 we learn that the thing identified as “ex:Book12345” is a book by Giuseppe Moretti (died in 1945) about the Ara Pacis in Rome, Italy. Each of the lexical strings that are the objects of the above sentences could also be
identifiers, which, when queried, would return a series of RDF statements where each identifier is the subject. Indeed, in a future bibliographic environment, the names and subjects, minimally, will be identifiers, thereby eliminating the current practice of embedding the lexical string into library records. If this alone were the only outcome of a new bibliographic environment (rest assured, there will be many more outcomes), then libraries and librarians would likely consider the enterprise a success.

RDF is a World Wide Web Consortium (W3C) recommendation (i.e., standard). The W3C is an international body that manages standards applicable to the Internet, and specifically the World Wide Web. HTML is the most well-known standard W3C has managed. RDF, which was formally published in 1999, comes under the umbrella of the W3C Semantic Web initiative. Although RDF itself is celebrating its thirteenth birthday, it has only been in the last five years that storage software, software tools, and general knowledge about RDF have matured beyond the “innovation” stage and penetrated safely into the relatively more confident “early adopters” stage, per the technology adoption lifecycle developed by Bohlen, Beal, and Rogers, shown in Figure 2.

Libraries, librarians, and developers have been active innovators throughout the thirteen year period of RDF’s ascendancy, during which time much has been tried and much has been learned—all of which directly informs current thinking about the Bibliographic Framework Initiative. One of the first instances of the Library of Congress publishing RDF came in 2005 when the MARC Relators list (a list of codes used to identify the role an individual had in the lifecycle of a resource, such as a book) was mapped to Dublin Core properties and published online. Although not the first of its kind, the LC Linked Data Service—id.loc.gov—went online in early 2009 and featured the LC Subject Headings file in RDF.

Libraries have legacy systems, legacy practices, and legacy data that must be carefully managed through any and all transitions. As mentioned in earlier announcements and articles about the Bibliographic Framework Initiative, the transition away from MARC will not be revolutionary, but a gradual process that ensures data integrity, system stability, and that no group is unintentionally left behind, in so far as is manageable. That RDF technologies may be on the cusp of entering the early majority stage means the library community is moving at the right time. It has only been in the last five years, or so, that RDF technologies have matured, including the uptake of those technologies, such that there is sufficient confidence in their continued support, development, and improvement. These technologies range from RDF vocabularies to RDF software libraries and to robust and scalable RDF triplestores (databases.
specifically designed to store and query RDF data). The library community needs to migrate to a new framework in a conscientious and responsible manner that accounts for the current state of library technology, but we also desire to enjoy a technological future that does not require costly, complete redevelopment every decade.

The fact that RDF is published by and managed by a standards organization like the W3C means that many developers and technologists beyond the library sector will more easily understand library data formats and technology. This is not the case today. Any RDF vocabulary or ontology developed in support of library data will, of course, still require subject matter expertise to fully understand its semantics, but developers and programmers will not need to understand the structural design of an ISO 2709 MARC record. Moreover, because the Bibliographic Framework Initiative is grounded in well-known and well-understood standards and technology that are widely used beyond the library sector, more individuals and companies will be competing in this space. Libraries will have a greater selection of services and solutions from which to choose.

Beyond the technology surrounding and supporting RDF, Linked Data methods and principles coincide perfectly with the mores and practices of libraries. Linked Data is about sharing data (i.e., publishing data). Developers have identified, promoted, and coalesced around an entire set of technical procedures—all grounded in the HTTP protocol—that have become commonly accepted (and expected) practice to facilitate access to structured RDF data via the World Wide Web. Dereferenceable URIs and content negotiation are two such procedures (neither belonging exclusively to the domain of Linked Data). Not only do these methods help to expose library data widely and make it more accessible, but also the Linked Data movement provides a strong and well-defined means to communicate library data, one of the main functions requiring attention in the community’s migration from MARC. Perhaps most importantly, by pursuing the Linked Data model for information sharing, the future Bibliographic Framework will embrace the notion of “The Network.”

Instead of library information having to be specially extracted from backend databases, packaged as groups of records (or singly), and then made intentionally available via some kind of transfer protocol, The Network will be the center of the model. Today, library records are independently understandable; an author’s name is in the record, as are the subjects, for example. In the future, there may be only an opaque identifier—a reference, a link, an HTTP URI—

CONTINUED »
to a resource that further describes the author or subject. It will be by performing content negotiation on the dereferenceable HTTP URI that a requesting agent (a human or system) will learn the lexical, human-readable string that is the actual name or subject heading associated with the identifier. More likely, systems will maintain a local copy of authority data, such as names and subjects, not only for indexing purposes but also because this is more efficient than requesting the lexical value over the network millions of times a day. But the maintenance, and the sharing, of this type of information will be fluid and infinitely easier in a Linked Data model than it is today.

Where to now?
The library community will need to further refine, customize, and possibly standardize (at least for faithful operation within the library community) the technology methods surrounding the exchange of and (potentially) the representation of library data in order to fully realize the Linked Data approach, with accompanying RDF model, in a new Bibliographic Framework. Work to date has revealed that technically conformant linked data service installations such as LC’s Linked Data Service will require expansion and refinement to serve the greater requirements of the new Bibliographic Framework. Although the present Linked Data methods are technically satisfactory, additional services can be implemented to ameliorate server loads, client processing energy, and a host of other small issues that singly amount to little but in the aggregate quickly become issues of scale. For example, a simple URI-to-string service would permit a client to request the lexical, human-readable value for a URI without the programmatic drudgery of sifting through every statement about a resource (especially when only one is needed), which is the current result based on Linked Data practices.

As characterized above, this is a transition from one bibliographic framework to a new one. Our legacy systems today deserve careful consideration and a measured approach. The timeline pursued by LC and its partners from January 1966 to the creation of the MARC Distribution Service in 1968 will not, therefore, be nearly as aggressive. Nevertheless, work has begun on developing a model for community discussion and identifying the technology needs to support the model. In the end, however, LC will still require—and, more importantly, wants—partners for this effort. There will be much for the community to contribute. LC is taking the lead with the Bibliographic Framework Initiative by coordinating and managing it, but the Initiative’s success rests on the valuable contributions (already received and still to come) from the wider community in the forms of discussion, feedback, testing, and, above all, participation during this process.

KEVIN M. FORD (kefo@loc.gov) works in the Network Development and MARC Standards Office, Library of Congress, and is the project manager for the LC Linked Open Data service.