Linked Data as Transformation

Philip E. Schreur
Stanford University
Coalition for Networked Information
April 3, 2012

Abstract: Linked data has the potential to transform every aspect of how we create, acquire, and discover information. By creating simple assertions in Resource Description Framework (RDF) and linking them together, a semantic web of data is created. Current library metadata encoded in Machine Readable Cataloging (MARC) is an ideal place to begin this transformation. Its consistency and quality will immediately enrich the Semantic Web and position our data where people are now searching for it.

A revolution is on the horizon, one that is potentially as world-altering as the development of the Web. And, as most truly transformative revolutions, it is driven by a simple concept: linked data. Linked data has the potential to change every aspect of our world of information creation and exchange, and as primary purveyors of information, the Library should be at the nexus of this revolution. Every aspect of our world will be dramatically altered as basic tenets of what we collect, how we collect, how we organize, and how we provide information are questioned and rethought. Much has been said about linked data, its ties to the Semantic Web, and its application for libraries, but what is it exactly and how does it work?

Linked data has so much potential because it is designed to work with the Web. And as more of our professional and private lives move to the cloud, the way in which information is stored and linked on the Web becomes crucial. The four tenets of linked

data are quite simple: use URIs (Uniform Resource Identifiers) to name things on the Web, use HTTP URIs so that someone can look them up, have the information provided by the link be useful, and provide links to other URIs so that people can discover related information.

Linked data is expressed using the Resource Description Framework, or RDF. The structure of any expression in RDF is composed of a collection of triples, each triple having a subject, a predicate, and an object. This simple structure allows anyone to make simple assertions about anything, for instance, The Raven (Subject) has author (Predicate) Edgar Allan Poe (Object). Ideally, both the subject and object would be represented by URIs and the statement itself expressed using an XML-based syntax. The advantage of using URIs is that much more accurate matching can be made. There may be many variations in the spelling of Edgar Allan Poe: Poe, Edgar Allan, 1809-1849, Edgar Allan Poe, E.A. Poe, etc., not to mention all the typos, and so any machine matching by character string is problematic. By linking to a URI, however, for Poe's authority record in the Library of Congress Name Authority File, the link is explicit. And by recording this information in RDF, applications can exchange information on the Web without loss of meaning. As RDF is a common language, information expressed in it can be used by many applications and applications can be developed to take advantage of this growing pool of data.

The strength of this model is that it allows anyone to make assertions about anything. What is equally as powerful is that any two expressions may be linked together and

through this process an immensely rich web of data is created. Although it is true that there is no requirement that these statements are true (e.g. The Raven has author Philip Schreur is equally as valid a statement in RDF), it is equally as true that anyone may correct invalid statements. In this way, through an iterative process of data use and correction, the web of data becomes more rich and more reliable; crowd-sourcing at a truly international level.

Since the days of the card catalog, our focus has been on bibliographic records. These discreet bundles of information supply metadata about resources in our collections. Their record structure is carefully controlled and access points such as names, subjects, or series come from recognized thesauri and carefully curated authority files. With our transition to online catalogs made possible through the development of MARC, our focus remains on bibliographic records. The information they contain is fractured into various fields and subfields and stored in relational databases where they can be associated and maintained.

This fixation on bibliographic records, though, has drawbacks. First, many institutions prefer their own particular version of a bibliographic record. Even though OCLC might espouse the use of the master record in their database, libraries are free to alter and enhance the copy of that record in their local database. Corrections to perceived errors in other's cataloging, missing data elements, and local practices can all be incorporated into a local version of the record to meet local users' needs. Large numbers of staff are

dedicated to this work at enormous cost. As the number of records grow, so does the cost of attempting to maintain them.

Second, these bibliographic records are stored in relational databases which are by definition closed systems. In order for a patron to discover a resource in the online catalog, a bibliographic record for it must be present in the system. The downside to this arrangement was driven home to me by Michael Keller, University Librarian at Stanford University, in the 1990s. At that time, I was Head of the Catalog Department and Mike asked me about cataloging the resources on the Web. I've puzzled over this question for more than a decade now. The question itself was very perceptive but far ahead of its time. Our patrons were increasingly interested in what was on the Web and it was natural for us to provide access to it; however, our mechanism for providing access was both too expensive and too restrictive to provide access to a nearly infinite number of resources. Within a world of limited staffing and records in relational databases, consistent access to the web of data is an impossibility.

Linked data, however, is not focused on bibliographic records but individual statements of fact. There are no discreet records to be maintained in a local ILS, no master records in a world-wide relational database, simply massive collections of triples in triple stores. By bypassing the a priori need for a record, linked data frees us from the cycle of record creation, maintenance, and deletion. Valuable staff time can be freed from these activities and the confines of the relational database can be broken.

But is linked data the solution?

From June 27 to July 1, 2011, Stanford University hosted a group of librarians and technologists to examine the use of linked data in the academic environment. The hope was that in the short week allotted to us that we could both confront the challenge of planning a multi-national, multi-institutional discovery environment and lay the ground work for its development. One of the most interesting products of the workshop was a series of value statements as to why a linked-data approach was worth pursuing:

1. Linked Open Data (LOD) puts information where people are looking for it – on the Web

2. LOD can expand discoverability of our content

3. LOD opens opportunities for creative innovation in digital scholarship and participation

4. LOD allows for open continuous improvement of data

5. LOD creates a store of machine-actionable data on which improved services can be built

6. Library LOD might facilitate the breakdown of the tyranny of domain silos

7. LOD can provide direct access to data in ways that are not currently possible, and provides unanticipated benefits that will emerge later as the stores of LOD expand exponentially.

As people shift to the Web as their first point of discovery, it's important for library

resources to be represented there. Although it is true that our catalog records may appear on the Web, any semantic meaning embedded in the MARC coding is lost. For the most part, the data in them becomes blocks of text. Through the use of RDF, however, important information encoded by the MARC tags can be translated into triples that carry semantic meaning for machine processing. And each one of the elements in the triple can be recorded as a URI that can link these data points to matching data points within the web of data. By intelligent conversion of our library MARC records to machine resolvable RDF triples, the semantic meaning in the records is preserved. By moving these statements to the Web, the data becomes a vital, structural part of the Semantic Web.

In the world of linked data, these MARC records are a prime, preliminary source of information. All the effort that catalogers have put into controlled subject access, controlled names, classification, and consistent description has made it extremely desirable. As any library foray into linked data must begin with its collection of MARC records, it's worthwhile taking a closer look at that format.

Take, as an example, a typical MARC catalog record for a sound recording:
The record is quite impressive. It gives a description of the medium, the contents, the years of performance, controlled subject headings, analytical entries for all the individual musical works it contains, and displays the information in an easily digestible structure for the eye. It's simple for anyone glancing at this record to see that it represents a recording of Fritz Kreisler performing a selection of violin music. The musical works are

clearly articulated and responsibilities are clear from glancing at the record as a whole. But what would a machine make of this record?

Much of the semantic meaning in this example can only be derived from the bibliographic record as a whole. The human eye can easily see that the main entry is Fritz Kreisler and that he is a violinist, that the piece by Joseph Sulzer is for violin and piano and if they liked this type of music that they could follow the subject heading Violin and piano music, and that the Mozart Violin concerto is accompanied by the London Symphony Orchestra conducted by Sir Landon Ronald.

This dependence on a complete bibliographic record for semantic meaning is a holdover from the card catalog days. The MARC format allowed these records to be transformed into electronic documents and shared internationally, but they are still bibliographic records and to be understood must be evaluated as a whole. Individual statements such as the Participant Note "Fritz Kreisler, violin, with various accompaniments" or the Event Note "Recorded 1904-1924" are meaningless taken out of context. RDF, however, is a series of independent statements meant to be understood by a machine. The conversion of MARC to RDF has to overcome two great obstacles, the first is the concept of the bibliographic record and the second is the inability of the MARC communication format to clearly convey semantic meaning.

It's often difficult for us to realize how much information our minds supply. From the author field we see Fritz Kreisler is listed as a creator, from the Participant Note we see

that he is a violinist. From the Contributor fields we see the recording includes Efrem Zimbalist, from the Contents note we see that he is also a violinist. From the Contents Note we see that Kreisler performs a piece by Tchaikovsky (Chant sans paroles) that was originally for piano, from the Included Works Note we see that this piece is from Tchaikovsky's work "Souvenir de Hapsal", from the Subject Notes we see that the correct LCSH subject term for this work is Violin and piano music, Arranged. There is nothing in the bibliographic record itself, though, that links these bits of information together. It is the human mind that makes these logical associations

The MARC format itself was created to clearly communicate the information encoded in our card catalogs, and in this it has been very successful. Although perpetuating the concept of the bibliographic record, it very clearly articulates and differentiates all the elements in the record. The MARC format is used, however, almost exclusively by the library community and much of its semantic meaning is lost to machine understanding. In the semantic world of linked data, these MARC records themselves are inarticulate. The shift to the Web as a primary source of information is unarguable. And as it is impossible for us to encompass the entirety of the Web in our library catalogs, our catalogs must move intelligently to the Web. Our millions of bibliographic records and the resources they represent are one of the truly great treasures we have to offer the web of data. The care with which we have created, maintained and enhanced them have made them a primary focus of the Semantic Web, but the way in which the data has been recorded in MARC prevents any intelligent, automated manipulation or linking. Although a daunting challenge, this conversion of our bibliographic records from MARC

to linked open data will become one of the most powerful drivers in the transformation to the semantic web, placing our data and resources where people are searching, and tying them intelligently to the wealth of the Web.