Background and Rationale

Background

Libraries worldwide rely upon Machine-Readable Cataloging (MARC)-based systems for the communication, storage, and expression of the majority of their bibliographic data. MARC, however, is an early communication format developed in the 1960s to enable machine manipulation of the bibliographic data previously recorded on catalog cards. Connections between various data elements within a single catalog record, such as between a subject heading and a specific work or a performer and piece performed, are not easily and therefore not usually expressed as it is assumed that a human being will be examining the record as a whole and making the associations between the elements for themselves. MARC itself was a great achievement, eliminating libraries dependence on card catalogs and moving them into a much needed online environment. It allowed for the development of the Integrated Library System, or ILS, and great economy in the acquisition, cataloging, and discovery of library resources. But as libraries transition to a linked-data based architecture that derives its power from extensive machine linking of individual data elements, this former reliance on human interpretation at the record level to make correct associations between individual data elements becomes a critical issue. And although MARC metadata can be converted to linked data, many human-inferred relationships are left unexpressed in the new environment. It is functional, but incomplete. With each day of routine processing, libraries add to the backlog of MARC data that they will want to convert and enhance as linked data. In the last ten years, computer science has embraced the LOD pathway that demands more semantic expression of data (that supports machine inferencing). It has developed approaches to data and international standards that support the new environment in the form of the use of identifiers to link data and the international standard, Resource Description Framework, or RDF, for recording it. Redevelopment of the platform for expressing and communicating bibliographic data is needed to move libraries more firmly into the internet and web environment.

The development of the digital library, often based upon a digital repository, has further complicated the library environment. In addition to their MARC data, libraries have become curators of rapidly expanding collections of digital objects, data sets, and metadata in other schemas such as the Metadata Object Description Schema (MODS). These resources and their metadata are typically stored in digital repositories and become a parallel, yet separate, database of record. This lack of integration has caused great difficulties in consistency and maintenance as the concept of a single database of record has broken down. And even beyond these two repositories (the ILS and the Digital Repository), as libraries look to the future, they will be asked to step outside these more traditional materials to become the curators of the vast knowledge the university creates, in all its richness and diversity. Interactive scholarly works, unpublished data sets, information about faculty contained in profiling systems, metadata about learning objects, once integrated with more traditional library resources, will allow our faculty and students to explore our information resources and make associations that are impossible today.

In 2012, the Library of Congress (LC) began a project to end libraries' isolation from the semantic web through the creation of a new communication format, called BIBFRAME, as a replacement for the MARC formats. The development of BIBFRAME has been a complex one as its creators try to balance the need to capture the data encoded in MARC, the constraints of RDF, and input from the community it hopes to serve. In addition, there are other schemas available for libraries' use, such as Schema.org, the CIDOC Conceptual Reference Model (CIDOC-CRM), and the Europeana Data Model (EDM). Although not designed as replacements for MARC, these other schemas are used by important information communities, such as Europeana9 or Museums, with which libraries interact. The resultant metadata ecosystem has created a very complex environment.

Schema.org itself deserves a special mention in this complex environment. Sponsored by Google, Microsoft, Yahoo, and Yandex, "Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond." It has been designed for the broadest possible use and focuses upon the semantic understanding of Web search engines. Because of this focus, it is of great interest to libraries and library-related organizations, such as OCLC, for embedding library data into the semantic web. It was never designed, however, to capture even the full richness of the data contained in MARC. Rather, its focus is on broad integration into the Web. BIBFRAME has been designed to fill that gap so that, as libraries move to the semantic web, the richness and detail of their metadata can be reflected there.

Likewise, the CIDOC-CRM has a special place in this project. Accepted as an ISO standard since 2006, CIDOC-CRM has been designed to encompass the full description of cultural heritage information: the objects themselves, their digital surrogates, and the metadata describing them, using either an objectcentric or event-centric modeling. The schema is extremely complex and tailored to the world of museums and cultural heritage organizations. Often, libraries may need to describe some of these materials but it is not the focus of their collections. They do, however, need to encompass the description of these objects in their discovery systems. The LD4P projects focusing on these materials will experiment with expanding BIBFRAME to include necessary concepts from CIDOC-CRM to produce a simpler but functional extension to BIBFRAME that can meet the basic needs of describing these materials in a common discovery interface.

Libraries have survived in their current environment by adhering to structural and data quality standards to facilitate the easy exchange of metadata for commonly held resources. These standards also allowed metadata from various institutions to be quickly combined into large discovery interfaces. As libraries transition from their current environment to a much more complex one based in LOD, these standards must be rethought and re-envisioned. Their need is still as strong but their expression is unclear. Since its inception, BIBFRAME has been used in a number of individual projects both within the United States and internationally. For instance, the University College London Department of Information Studies has been awarded a grant to develop a Linked Open Data bibliographic dataset based on BIBFRAME. The Library of Alexandria will focus on the conversion process for data in the Arabic language. The National Library of Medicine has developed a more modular approach to the BIBFRAME vocabulary by paring down the existing vocabulary to its core concepts (BIBFRAME-Lite). We now have arrived at the point where these individual efforts should be drawn together to create the common environment, standards, and protocols that have allowed libraries to interact so strongly in the past. And by expressing relationships in a standard way so that machines can understand the meaning inherent in them, the heart of the semantic web, library's data will finally be able to be embedded into the Web.

Rationale

In order to address these issues, Stanford University proposed a planning grant to the Mellon Foundation in 2014 called Linked Data for Production (LD4P). The planning grant proposed two meetings to define and organize a series of projects that would begin the transition to the native creation of linked data in a library's production environment. The core members of LD4P are Columbia, Cornell, Harvard, the Library of Congress, Princeton, and Stanford. The outcome of those meetings was a report submitted to the Mellon Foundation in July of 2015. The group had a final meeting recently at the Library of Congress to formalize its plans.

This group of six libraries is particularly well suited to pursue this transition in technical services. Cornell, Harvard, and Stanford are founding members of Linked Data for Libraries and will be building upon collaborative efforts already well underway. The Library of Congress is the originator of BIBFRAME and is engaged in a project to explore the use of BIBFRAME in its current workflows. Columbia and Cornell's Technical Services Departments are already allied through another Mellon-supported project called 2CUL. And Princeton was one of the early BIBFRAME experimenters in the United States. Beyond this, however, these institutions are deeply enmeshed in the current technical services ecosystem. The transition to LOD cannot be accomplished exclusively by libraries. Libraries have become dependent upon vendor services (cataloging, authority control), the ILS, standards organizations (the Program for Cooperative Cataloging (PCC), and domain experts. As part of LD4P, these institutions can influence the vendor community as a group to encourage them to make the transition to LOD. They can work with their own ILS (SIRSI, Ex Libris, OLE) to incorporate LOD into future plans. SIRSI/Dynix has already expressed interest in working with Stanford on its linked data workflows through the use of their new product, Blue Cloud. OLE has already been actively engaged with UC Davis and the BIBFLOW project and plans on enhancing their linked data capabilities. Cornell has recently announced that they will be moving to OLE for their ILS. If they make the transition early enough in this grant cycle, they may be able to take advantage of OLE's capabilities as well.

This new communal, distributed model based on web architecture will change how we communicate and share our data. Centralized data stores, such as the OCLC database, will be joined by alternative data pools as the marketplace shifts in support of this new environment. Traditional authority control will be supplemented by identity management. Cataloging standards may have to evolve from their focus on transcription of data as it appears on an item, which a human can easily read and interpret at a computer screen, to data that a machine can understand and link semantically.