Linked Data for Production and the Program for Cooperative Cataloging

Philip E. Schreur
Stanford University
10/9/2017

Background

Linked Data for Production is focused on the transition of basic Technical Services workflows to ones rooted in linked open data through active engagement with the semantic web. This transition will be a critical one for libraries. The current means of encoding data, the MARC communication formats, was developed in the 1960s to transform metadata on library catalog cards to a machine readable format. The importance of this transition can be elucidated by two questions.

First, why not MARC? MARC was a revolution in its day. It allowed data from library card catalogs to be encoded in machine readable form, enabling the catalog cards to be reproducible on the computer screen and the data to be exchanged freely among libraries. It is a fifty year old technology, however, only understood by library systems. In addition, the MARC formats are semantically inexpressive and have isolated libraries from the development of the Web.

But why linked data? It has been apparent that library patrons have preferred searching for information on the Web for quite some time. By integrating library data into the Web in a semantic way, our patrons can find well-formed library data there as well. But in addition, by taking advantage of the semantic web, library patrons can benefit from other important data sources on the Web. A third advantage is that the Web is an international environment. By shifting to linked data, libraries worldwide can take advantage of the bibliographic and authoritative data many national libraries create and make available now as linked data. And last, the Web is a continually evolving environment. Without a doubt, linked data will evolve into some other standard with time. But in order to move along with this evolution, libraries will need to make that first important step in the transition to a Web environment.

Linked Data for Production Phase 1

In Linked Data for Production Phase 1, the partners proposed the development of a communal work environment based in linked data, the strengthening and expansion of the BIBFRAME ontology to cover the multiple formats (e.g., books, music, maps, etc.) that libraries must catalog, the tools needed to perform the work itself, and the development of light weight workflows (Tracer Bullets) to prove that the transition to linked data was both possible and practical.

> **Communal Work Environment:** The partners were fortunate that Casalini Libri developed a communal work environment in support of the project called the SHARE-VDE, or the SHARE-Virtual Discovery Environment. The environment includes a semantically enhanced MARC to BIBFRAME converter, an advanced reconciliation system, and a discovery environment. The availability of this environment will be key to the success of Linked Data for Production Phase 2.

BIBFRAME Ontology: The Library of Congress has been very open to working with LD4P in the refinement and expansion of the BIBFRAME ontology over the past year. In addition, the partners have made extensions to BIBFRAME in the areas of performed music, art, rare books, and cartographic materials. These will allow libraries to be able to catalog all materials passing through their traditional workflows.

Tooling: Tooling will be key to any linked data transition, and the most important tool will be an editor to both create and edit data. In Phase 1, LD4P is currently experimenting with two BIBFRAME 2.0 editors. The first is the LC BIBFRAME editor. It is currently getting a thorough shake-down as LC prepares to train over seventy staff to use it in linked data creation. The second is a new editor being developed at Stanford called CEDAR. CEDAR is a more flexible tool, capable of creating BIBFRAME 2.0 as well as data in other schemas.

Workflows: In LD4P Phase 1, Stanford focused on the conversion of four key workflows to a linked data strategy: two related to the traditional ILS and two to the digital repository. The ILS workflows have been established in a light weight fashion and we are now ready to expand them both in depth and in participants for the next phase of LD4P. The expansion of these workflows will be the next critical development as they cover the predominate resources libraries must handle in their day-to-day production.

Linked Data for Production Phase 2

In Linked Data for Production Phase 2, Stanford intends to build upon the foundational work in the first grant to move us onto the critical path to implementation. The next phase is planned as a three year development, allowing us to establish the environment in which to work and transition to implementation at first-adopter institutions. There will be four foci in the next stage: expansion of Tracer Bullet 1 (use of preexisting metadata); expansion of Tracer Bullet 2 (creation of new data); enhancing discovery; and community collaboration. Together, these four projects will allow us to realize seven key aspects of implementation: the publication of a collection of high-quality library metadata to the Web; participation in a communal work environment among a group of peers; the implementation of production level workflows and tools; the creation of a cloud environment for the training of new participants; support for the creation and maintenance of new identifiers in metadata development; visible benefits to a semantically enriched discovery environment; and the strong engagement with a developing, worldwide LOD library community. Of special interest to PCC will be Tracer Bullets 1 and 2.

Tracer Bullet 1 (use of preexisting data): Tracer Bullet 1 is the heart of Technical Services workflows. Most resources (in the case of monographs, approximately 85%) come in with some form of copy. In a MARC record-based economy, this pathway is well understood. As we transition to a recordless, communal RDF environment, simple concepts such as editing and holdings are no longer clear. We will need an environment such as the SHARE-VDE environment in which to define these concepts and test implementation.

Tracer Bullet 2 (new metadata creation): Tracer Bullet 2 is the complement of Tracer Bullet 1. It is the creation of metadata for resources for which none exist. In Tracer Bullet 2, we would like to establish a sandbox for experimentation in the creation of RDF data according to PCC guidelines. This experimental component would be open to all PCC members. In addition, there would be formalized training for up to 10 libraries in the implementation of RDF cataloging as a replacement for current MARC cataloging.

Program for Cooperative Cataloging

In 1992, Sarah Thomas helped to form the Program for Cooperative Cataloging in an effort to modernize and bring efficiencies to the cataloging process. In effect, though, she created a decentralized national library. By forming a group to develop standards for high-level metadata production and sharing, she replaced what is done in some countries by a national library with a cooperative network of peer institutions.

PCC has always had a critical function in helping guide its members in the transition to new standards or technologies. The transition to RDA is an excellent case in point. The PCC's many task groups and training efforts made the shift from AACR2 to RDA possible in the United States. The shift from MARC to linked data will need no less from the PCC.

The shift to linked data is far more than a shift in technology. As PCC looks to its future, it will be redefining what it means to create and share data in an international, open Web environment. How do we move forward with what is most important for our users and yet fully integrate ourselves into the richness of the Web? The questions will be many in this exciting and turbulent transition.

But in the meantime, there will be critical issues that PCC will need to resolve quickly in order that the transition to the implementation of linked data can begin. The issues naturally group themselves into three areas: technology-related, policy, and training.

Technology-related: RDF is a Web standard developed and maintained internationally. However, there will be multiple issues to resolve in how it applies to the work of the PCC. For instance:

- The BSR and CSR are defined independently of the current communication format (MARC). They still have MARC elements embedded in then, however. How must they be reinterpreted in a linked data context? The recent report issued by the PCC moves this issue forward but still needs to be finalized.
- Will there be a single, encouraged ontology to be used as a common language for the PCC (BIBFRAME for instance)?
- How can the idea of PCC quality be reinterpreted in a recordless environment?
- Can PCC support an environment in which members can do their work, especially those that cannot develop one locally?
- Does PCC have a role in the community maintenance of BIBFRAME and supporting domain extensions?

Policy: Just as with the transition to RDA, we will be working in a mixed MARC/linked data environment for a number of years. Not only will there be policy decisions to be made concerning the transition to linked data but also how members should work in this mixed environment. For instance:

- Does an RDF representation of RDA cataloging meet PCC standards for program participation?
- Is there an obligation to provide a MARC parallel for members who cannot make use of the RDF?
- If there is a MARC and RDF version of the data, how do they relate to each other over time?
- Is an identifier sufficient to support an access point in PCC cataloging?
- May PCC members make use of international bibliographic and authority data?

Training: One of PCC's many wonderful functions is that of supplying training for its members. How will this manifest itself in the linked data transition?

- Training in the use and maintenance of ontologies?
- Training in the use of tools?
- Training in the reuse of data?

LD4P is anxious to partner with PCC in the first steps of the linked data transition. We would like to develop a cloud based sandbox in partnership with PCC in which PCC members could experiment with the use of BIBFRAME editors for their cataloging. We would also like to develop a more intensive program for those interested in early implementation at the end of the next grant. In addition, we'd like to work with PCC in the testing of what "copy-cataloging" might mean in a recordless environment through the use of the SHARE-VDE platform. It is our hope that minimal policies will be in effect by the end of the next grant so that members who wish to transition to the use of RDF instead of MARC can do so with the clear understanding and support of the PCC.

This is one of the truly pivotal moments for libraries in general and technical services in particular. PCC has been the guiding force in the creation and exchange of library data in the United States for the past 25 years. They have the authority and means to lead us through this transition as well.