Summary, Scope, Background, and Rationale (LD4L Labs)

Table of Contents

Linked Data for Libraries: LD4L Labs – Creating, Discovering, and Understanding Library Linked Data

January 15, 2016

Summary

The *Linked Data for Libraries: LD4L Labs* proposal is a collaboration of Cornell, Harvard, Iowa, and Stanford to continue to advance the use and usefulness of linked data in libraries. Project team members will create and assemble tools, ontologies, services, and approaches that use linked data to improve the discovery, use, and understanding of scholarly information resources. The goal is to pilot tools and services and to create solutions that can be implemented in production at research libraries within the next three to five years. The project team will develop tools and provide direct support for projects within the related LD4P Program as described in a grant proposal from Stanford University Libraries. This proposal seeks $1,500,000 from the Andrew W. Mellon Foundation for the period from April 1, 2016 through March 31, 2018 to support this work.

Scope

The LD4L Labs project is focused on developing new tools and approaches that will make it easier for libraries to create and use Linked Open Data that describes their scholarly information resources. The project will develop and support tools for linked data creation and editing, the bulk conversion of existing metadata to linked data, and a common system to support initial work in entity resolution and reconciliation. The project will also explore strategies to use linked data relationships and analysis of the Linked Open Data graph to directly improve discovery and understanding of relevant scholarly information resources. Finally, the project will provide feedback to the library ontology community about the use of BIBFRAME and other relevant ontologies within the tools being developed and in support of discovery and understanding.

In contrast to LD4L Labs, the scope of the Stanford-led LD4P project the establishment of the foundations of the transition of library technical services operations to ones based in linked open data. As a whole, the six institutions will focus on four main areas of development. First will be the establishment of the ability to create linked open data communally. Second, in collaboration with external standards organizations such as the Program for Cooperative Cataloging and linked data projects such as BIBFLOW, will be the establishment of common procedures and protocols for the creation of library metadata as linked data. Third will be the expansion of the BIBFRAME ontology to better encompass subject domains such as art and music. And last will be the transition of a selection of current library workflows to ones based in linked open data. The projects will make use of a collection of preliminary tools (those developed by LC, Vitro, and eagle-i) and adapt them for production work at their individual environments. The feedback from the use of these tools in a production environment will allow their developers to further refine their tools to meet practical demands. As LD4L Labs enhances the suite of tools, the LD4P partners can take advantage of the enhancements in the transition of their workflows.

The projects propose close collaboration, with frequent joint meetings and two shared staff positions that will devote different parts of their effort to the two grants. In addition, tools being developed by LD4L Labs will be directly used in the metadata production work of some of the LD4P partners. Conversely, the LD4P linked data and use cases will also directly inform the development of LD4L tools. Details of these collaborations are discussed in the individual sections below.

A. Background

In the initial Mellon-funded Linked Data for Libraries project[1], the LD4L team gathered data, assembled an ontology, and built some of the basic infrastructure to share linked data about scholarly information resources, such as traditional monograph and journal publications, archival materials, research datasets, images, recordings, cultural artifacts, newspapers and magazines, web archives, and much more. This infrastructure included the creation of a shared processing pipeline that converts existing MARC[2] catalog records at the three partner institutions to linked data, together with pre- and post-processing steps to make this linked data more useful and uniform across the partners.

The linked data created for these scholarly information resources included bibliographic data, curation data (e.g., metadata related to the organization and annotation of the resources), data expressing how the resources relate to people and organizations, and usage data. While bibliographic data is typically directly specified by catalogers, curation data is more commonly the byproduct of a scholarly activity. Curation would happen, for example, when a resource is included by a museum or archive as part of a focused exhibit; when a resource is added to a course reading list by a professor; or when resources show up together on a focused bibliography. The data that is captured as a byproduct of these activities is typically represented as a virtual collection or annotation and will be referred to within this proposal as “curation data”.

The project developed a set of use cases[3] focused on creating relationships from those resources to real world entities (represented as linked data URIs) that provided context for, and enhanced understanding of, those resources. Based on those use cases, the project team assembled an ontology[4], drawing together elements from a variety of existing ontologies, including the Library of Congress (LC) BIBFRAME[5], VIVO-ISF[6], Open Annotation[7], OAI-ORE[8], and several others. In the case of BIBFRAME, the project made a number of modifications to the original proposed ontology and provided those changes and the rationale behind them to LC in a report authored by Rob Sanderson with contributions from other LD4L team members. LC has indicated that most of these recommendations will be included in the next revision of BIBFRAME.

In February 2015, the project held an LD4L Workshop[9], assembling fifty participants to provide feedback on the use cases, ontology, and planned demonstration implementations. That workshop provided immediate guidance to the project and also made significant suggestions for follow-on work, many of which are reflected in this proposal. Results from the workshop and from the overall project have been publicly presented in a number of forums[10], including the DLF Forum in Vancouver, BC in October 2015.

The project is finalizing the process to make available Linked Open Data[11] representing approximately 23 million cataloged scholarly information resources from Cornell, Harvard, and Stanford using the LD4L ontology. This public data will be available in 2016 through the ld4l.org website. The initial release will reconcile works based on common OCLC Work IDs[12] and people based on common VIAF (Virtual International Authority File) identifiers[13] across that linked data. It will also provide Stackscore[14]-based usage data on the resources from Cornell and Harvard, where a Stackscore is simply a number from 1 to 100 that represents how relevant an item is to the library's patrons as measured by how they've used it. The number can reflect data from circulation, number of copies held, browse counts, or other library-specific usage information.

The use of standardized and reconciled linked data representations for scholarly information resources across all the partner institutions allows these resources to be treated as one comprehensive collection. Having such a common collection means that tools, relationships with and analysis of the linked data graph, and new linked data sources can and will build on the scholarly information resources from all the partner institutions, not just on resources from a single institution. This approach also means that the tools and approaches being developed will be reusable, and allow inclusion of resources from any institution that makes information available as compatible Linked Open Data. Through linked data, alignment to shared identifiers by any institution also means that references to external entities, including global identifiers for people, organizations, concepts, events, places, and other types, will enrich much more than that single institution's data.

B. Rationale

The first Linked Data for Libraries project, funded by the Andrew W. Mellon Foundation in January 2014, presented an extensive discussion of the importance of linked data and semantic web technologies for making scholarly information resources at academic libraries more discoverable, understandable, and usable[15]. That discussion will not be repeated here, but it is important to note that since that proposal was funded, there has been growing interest in and use of linked data by the library, archives, and museum community, as reflected in many workshops, conference presentations, and publications.

Cornell University Library (CUL) is itself pursuing several other linked data efforts in addition to the work proposed here. In particular, CUL, in partnership with the Library of Congress, OCLC, the Program for Cooperative Cataloging, and Harvard University Library, has requested funds from IMLS to hold a national forum on issues concerning local authorities in library metadata, with a focus on name identities. CUL is also part of the NSF-funded EarthCollab project[16], a partnership with the National Center for Atmospheric Research and UNAVCO to use semantic web technologies to better describe the interconnected network of research datasets, publications, researchers, experiments, and research instrumentation. Finally, CUL has long been part of the VIVO project[17], focused on using linked data to describe the full academic context (e.g., publications, teaching, grants, departments and programs, and research projects) for researchers and scholars.

CUL has also recently committed to adopting the Open Library Environment (OLE) as its integrated library system. CUL staff will work with LD4L Labs team members to evaluate the potential to integrate the tools and approaches described in the Project Description section below with the OLE environment. In doing so, CUL will seek to collaborate with and build on the work of UC Davis on BIBFLOW[18]. BIBFLOW is an IMLS-funded project focused on "a research agenda and a set of activities" to help the community understand its resource landscape and develop a roadmap for the coming years to move its technical services workflows into ones based in linked data.

The original LD4L project has led to two separate but closely related follow-on proposals. This proposal focuses on building on and extending the research, tool development, experimental pilots, and infrastructure work of the prior LD4L project. The goal of the LD4L Labs work is to develop solutions in the specific areas described in the Project Description section below that could be piloted within the term of the grant and implemented on a production basis in research library environments within the next 3-5 years.

The goal of the companion Linked Data for Production (LD4P) project being submitted by Stanford University Libraries is:

"to begin the transition of technical services production workflows to ones based in Linked Open Data (LOD). This first phase of the transition focuses on the development of the ability to produce metadata as LOD communally, the enhancement of the BIBFRAME ontology to encompass the multiple resource formats that academic libraries must process, and the engagement of the broader academic library community to ensure a sustainable and extensible environment."

While the LD4P and LD4L Labs projects have separate agendas and goals, there is a great deal of synergy between the two efforts. In particular, several of the LD4P efforts make use of tools being developed and supported as part of LD4L Labs, and the linked data created through the LD4P projects will be used by several other efforts within LD4L Labs focused on improving discovery, use, and understanding of scholarly information resources.

The participants invited to the LD4L Workshop in February 2015 made a number of recommendations on how to advance the use and usefulness of linked data in the library community. They also identified some specific challenges that the library community must address if it is to successfully move to a linked data infrastructure. Here is a summary of some of the recommendations and related challenges:

- The goal should be that others outside the library community use the linked data that libraries produce. Instead of just LD4L, the focus should be on LD4E (Linked Data for Everyone). One major challenge related to this goal is that libraries must start to think outside the box of the traditional bibliographic record. Instead libraries must create data that truly links to and is visible on the Web.
- The project must create applications that let people do things they couldn't do before – don't talk about linked data, talk about what people will actually be able to do with it. A related challenge is a tendency to focus on perfecting the linked data infrastructure, such as ontologies and reconciliation, rather than providing infrastructure that is good enough to make possible new applications and solutions.
- To be discoverable as linked data, local original assertions (new vs. copy cataloging) should use local URIs even when global URIs exist. The use of local URIs allows libraries to make local assertions about their resources that will be remain in their namespace, thereby establishing provenance for assertions without having to add layers of explicit attribution, and ensuring that changes or errors elsewhere in the linked data cloud will not break the work done by local librarians. The challenge here is that to be truly useful, those local URIs must also be reconciled with global entities – in the face of issues such as the evolution in the entities being referenced, organizational succession, and differential adoption of global entity URIs in different but overlapping schemes that will necessitate additional cross-referencing.

- Look to linked data to bring together physically/organizationally dispersed but related collections. By providing common infrastructure and relationships across resources at many different institutions, it is now possible to build virtual collections, exhibits, and libraries across highly diverse scholarly information resources. This approach is very similar to that taken by the Shared Canvas[19] approach to assembling distributed manuscript pages and fragments into coherent views and volumes.
- Libraries must create a critical mass of shared linked data to ensure efficiency and to benefit everyone. The work of the LD4L project to make some three billion triples describing 23 million scholarly information resources at Cornell, Harvard, and Stanford openly available as Linked Open Data is a first step toward this goal. A major challenge here is to address the issue of scale both by identifying tools and infrastructure that will work at library scale, and also by making careful technology choices that address specific library use cases, rather than trying to solve every possible need.

The LD4L Labs proposal seeks to make progress in addressing both the recommendations and the challenges identified in the LD4L Workshop, as well as those identified by the project itself and the broader library community. The goal is both to advance the overall linked data agenda and to provide a set of specific tools, solutions, and approaches that can provide production-level value to libraries within three to five years and demonstrate the value of linked data to solve real problems for research libraries, scholars, and students.

---

[1] http://ld4l.org
[2] http://www.loc.gov/marc/
[3] https://wiki.duraspace.org/display/ld4l/LD4L+Use+Cases
[4] https://github.com/ld4l/ontology
[5] http://www.loc.gov/bibframe/
[6] https://wiki.duraspace.org/display/VIVO/VIVO-ISF+Ontology
[7] http://www.openannotation.org/spec/core/
[8] https://www.openarchives.org/ore/
[9] https://wiki.duraspace.org/display/ld4l/LD4L+Workshop+Overview
[10] https://wiki.duraspace.org/display/ld4l/Communications+and+Outreach
[11] http://linkeddata.org/
[12] https://www.oclc.org/developer/develop/linked-data/worldcat-entities/worldcat-work-entity.en.html
[13] https://viaf.org/
[14] http://stacklife.harvard.edu/explainer.php
[15] “Why Linked Data?” at https://www.ld4l.org/linked-data
[16] http://earthcube.org/group/earthcollab
[17] http://vivoweb.org/
[18] https://www.lib.ucdavis.edu/bibflow/
[19] http://iiif.io/model/shared-canvas/1.0/index.html