Linked Data for Libraries: A Project Update

Dean B. Krafft

Cornell University Library, Ithaca, NY
dean.krafft@cornell.edu

Abstract. This poster reports on the first eighteen months of the Mellon-funded two-year Linked Data for Libraries (LD4L) project [1], a partnership of Cornell University Library, Stanford University Libraries, and the Harvard Library Innovation Lab. The goal of the project is to use Linked Open Data to leverage the intellectual value that librarians and other domain experts and scholars add to information resources when they describe, annotate, organize, select, and use those resources, together with the social value evident from patterns of usage. The project is producing an ontology, architecture, and set of tools that work both within and across individual institutions in an extensible network.

Keywords: Ontology, Linked Data, Libraries, VIVO, Hydra Framework

1 Project Summary

The Cornell University Library, the Harvard Library Innovation Lab, and the Stanford University Libraries have all been exploring new approaches to dramatically improve the discovery experience for users seeking 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. All three institutions have been looking at ways to gather context and relationships about these resources that go far beyond traditional metadata approaches. The goal of this project is to create a Linked Data for Libraries (LD4L) model that works both within individual institutions and through a coordinated, extensible network of Linked Open Data (LOD). This LOD will capture the intellectual value that librarians and other domain experts add to information resources when they describe, annotate, organize, select, and use those resources, together with the social value evident from patterns of usage.

To achieve this goal, the project team will:

- Create a set of use cases that specify how LD4L information can enhance user discovery and understanding of scholarly information resources
- Assemble and, where necessary, create an LD4L ontology to represent the required bibliographic, person, curation, and usage information as linked data
- Hold a two-day workshop with library, archive, and museum linked data experts from a variety of institutions to gather feedback on the LD4L use cases, ontology, and work plan

- Create linked open data sources at each institution providing bibliographic, person, curation, and usage data for the scholarly information resources of the institution using the LD4L ontology
- Create and release open-source software for creating institutional LD4L instances and using LD4L data as part of the Hydra Framework [2]
- Create a demonstration search across the combined LD4L linked data from all three institutions

2 Progress Report

The poster to be presented will summarize progress on three focus areas for the overall project: use case development, the LD4L ontology, and outcomes from the LD4L workshop. The poster should be of interest to those who:

1. Are interested in understanding use cases for applying linked data techniques to describing, discovering, and understanding scholarly information resources;
2. Want to understand the specific ontology choices that the project has made to address these library use cases; and
3. Want to hear about the feedback from linked data experts on the use cases, ontology, and demonstration systems that were presented at the LD4L workshop.

The sections below briefly summarize the material to be presented in each of these areas.

2.1 LD4L Use Cases

The work of LD4L has been heavily influenced by use cases; if the LD4L ontology, any consuming applications, or linked data in general are going to be fit for purpose, the purpose and criteria for success need to be defined. For the first half of Year 1 of the LD4L project, partners invested heavily in an extensive process of articulating what they wanted to accomplish via linked data, for whom, and why it would be beneficial to realizing the mission of a library. This multi-stage effort used a classic approach borrowed from agile software development methodologies to articulate functional requirements or use cases in the form of "stories": "As a <type of user>, I want to <perform an action>, so that I can <realize a benefit>".

Partners at Harvard, Cornell and Stanford generated a total of 42 raw use cases in this form. After reviewing for overlap, applicability to linked data, feasibility for engineering, and availability of data, the project team reduced this suite of use cases to 12 use cases in 6 distinct clusters [3]. The 6 clusters are links between: 1) bibliographic and curation data; 2) bibliographic and person data; 3) leveraging external data including authorities; 4) leveraging the deeper graph (via queries or patterns); 5) leveraging usage data; and 6) cross-site services.

As an example, Use Case 2.1 (in cluster 2, bibliographic and person data) is: See and search on works by people to discover more works, and better understand people. An example story in this use case is: “As a researcher, I'd like to see / search on works <by,

about, cited by, collected, taught> by University faculty <in an OPAC, profiles system>, to discover works of interest based on connection to people, and to understand people based on their relation to works."

The project team continues to consult the use cases as an ongoing guide. In Year 1, the work around Use Case Cluster 1, for example, focused on using linked data allow users to build virtual collections of scholarly information resources drawn from a variety of source (e.g., items described by library catalog records or items cited in faculty research profiles). Cornell and Stanford both developed and designed systems to exercise these capabilities, and demonstrated them in versions of their Blacklight-based catalogs.

2.2 LD4L Ontology

One of the major outputs of the project is the LD4L Ontology, which is used to share information about scholarly resources among the project participants and to interconnect those resources with the broader web of Linked Open Data.

The group early on confirmed that it makes eminent sense for a project focused on linked data to draw as much as possible on existing ontologies that have already achieved significant adoption or show promise for doing so, rather than creating a new self-contained ontology. Elements of the Bibliographic Ontology [4] and FaBiO [5] had already been incorporated into the VIVO-ISF Ontology [6] and were familiar to team members from previous work. The BIBFRAME initiative [7] at the Library of Congress addresses the representation of MARC metadata in RDF, while OCLC has worked to extend the Schema.org ontology [8,9] as a bridge between the library community and the Web. Use cases 1.1. and 1.2 use the Open Annotation Data Model [10] to represent collection annotations and the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) [11] for ordering. For provenance, the Provenance, Authoring and Versioning (PAV) ontology [12] provides a solid starting point while the W3C Provenance Ontology (PROV-O) [13] offers additional granularity when data are available to support more nuanced attribution.

The team has also prioritized the ability to convert references within library metadata records from "strings" to "things," reducing reliance on the lexical form of a name by adopting URI-based identifiers as the primary means of disambiguation. Whenever possible we seek out persistent global identifiers for the entities being represented – identifiers from established international efforts such as the Open Researcher and Contributor ID [14], the Virtual International Authority File [15], or the International Standard Name Identifier [16] for people; global identifier systems are also emerging for organizations (VIAF, the Ringgold Identify Database [17], and others).

This approach does not preclude the creation or reuse of URIs in a local institutional namespace as identifiers as part of publishing linked data. When a metadata record only references an external standard – a subject term in the Getty Art and Architecture Thesaurus, for example – no local URI is necessary, but when the metadata includes additional original statements about that entity, a local URI supplemented by an owl:sameAs assertion to the external entity will be necessary to allow those locally-asserted statements to be retrievable as linked data.

Following these principles, the project has now assembled an LD4L Ontology, which will be used to implement demonstration systems for the use cases during the final phase of the project. The poster will present a summary of this ontology.

2.3 LD4L Workshop

The poster will also summarize outcomes of the LD4L workshop, which brought together fifty linked data experts at Stanford in late February 2015, who provided extensive feedback on the use cases, ontology design, and engineering work to date. Input from the workshop informed both the ontology and the specific demonstration systems to be built in the final phase of the project. A full agenda for the workshop, as well as session notes and slides from the presentations, is available at [18].

Acknowledgements. The work described in this poster includes contributions from the entire LD4L project team, including from Cornell: Dean B. Krafft, Jon Corson-Rikert, Brian J. Lowe, E. Lynette Rayle, Rebecca Younes, Simeon Warner, Chew Chiat Naun, Steven Folsom, Jason Kovari, and Jim Blake; from Harvard: David Weinberger, Paul Deschner, Paolo Ciccarese, Jonathan Kennedy, and Randy Stern; and from Stanford: Tom Cramer, Philip Schreur, Rob Sanderson, Lynn McRae, Naomi Dushay, Nancy Lorimer, Darren Weber, and Joshua Greben. The project would like to thank the Andrew W. Mellon Foundation for its generous support of this research.

References

1. Linked Data for Libraries (LD4L), http://ld4l.org
2. Project Hydra, http://projecthydra.org
3. https://wiki.duraspace.org/display/ld4l/LD4L+Use+Cases
4. Bibliographic Ontology, http://bibliontology.com/
5. FaBiO, http://vocab.ox.ac.uk/fabio
6. VIVO-ISF Ontology, https://wiki.duraspace.org/display/VIVO/VIVO-ISF+Ontology
7. BIBFRAME, http://www.loc.gov/bibframe/
8. http://blog.schema.org/2014/09/schemaorg-support-for-bibliographic_2.html
9. Schema.org, https://schema.org/
10. Open Annotation Data Model, http://www.openannotation.org/spec/core/
11. OAI-ORE, https://www.openarchives.org/ore/
12. PAV, http://www.jbiomedsem.com/content/4/1/37
13. PROV-O, http://www.w3.org/TR/prov-o/
14. ORCID, http://orcid.org/
15. VIAF, http://viaf.org/
16. ISNI, http://www.isni.org
17. Ringgold Identify Database, http://www.ringgold.com/identify
18. https://wiki.duraspace.org/display/ld4l/LD4L+Workshop+Overview