Ontology Team Overview

Workshop Ontology Overview Presentation (PDF)

Introduction

The LD4L ontology team formed very early in the project and has met on a weekly basis for discussions on a wide range of topics, from proposing possible use cases to reviewing the ontology aspects of use cases proposed by other teams to discussing the specifics of how best to represent the data coming from our three library catalogs and from other internal and external sources.

Team members are listed on the LD4L Working Groups (LD4L 2014) page, and the team has benefited from the addition of new members including strong representation from the technical services and metadata departments of Cornell, Harvard, and Stanford.

Several principles have guided our discussions and influenced the projects selected by ontology team members. The group early on confirmed the intention stated in the proposal to reuse appropriate parts of currently available ontologies rather than building a new, self-contained ontology for LD4L. While there are advantages to working from a blank slate, we believe 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.

At the outset the ontology team recognized the existence of a great deal of prior art in the form of published ontologies and significant ongoing ontology initiatives addressing the representation of bibliographic information in RDF. Elements of the Bibliographic Ontology and FaBIO had already been incorporated into the VIVO-ISF Ontology (GitHub) and were familiar to team members from previous work – and Paolo Ciccarese from Harvard was a principal FaBIO contributor. The BIBFRAME initiative at the Library of Congress addresses the representation of MARC metadata in RDF, while OCLC has worked to extend the Schema.org ontology as a bridge between the library community and the Web. The Collections ontology and ORE address collections of digital objects; the Open Annotation Data Model annotations, and PROV-O and PAV provenance.

From the 2013 LD4L proposal

SRSIS Ontology

Because no existing ontology supports the range of entities and relationship that SRSIS will encompass, we will use the Protégé ontology editor to develop a SRSIS ontology framework that reuses appropriate parts of currently available ontologies while introducing extensions and additions where necessary. The framework will be based on and remain compatible with the existing VIVO and emerging research dataset and research resource ontology work. It will be sufficiently expressive to encompass traditional catalog metadata from Cornell, Stanford, and Harvard; the basic linked data elements described in the Stanford Linked Data Workshop Technology Plan; and the usage and other contextual elements from StackLife. The ontology will capture a series of basic concepts and be structured as modules that draw inspiration from and reuse existing ontology classes and properties where appropriate, such as the Semantic Publishing and Referencing ontologies, and that also support arbitrary system-wide refinement, including local extensions.

Ontology Team Activities, 2016

Engaging with BIBFRAME

The LD4L ontology team spent much of 2015 engaged in the evolution of BIBFRAME, namely through the BIBFRAME discussion list and directly with the Office of Network Standards at the Library of Congress. In April Rob Sanderson submitted Analysis of the BIBFRAME Ontology for Linked Data Best Practices to the Library of Congress; the document recommends changes to BIBFRAME to better reflect best practices in the Linked Data domain so as not to marginalize the RDF data libraries contribute to the semantic web.

Following on Rob Sanderson's and LD4L's recommended changes to BIBFRAME, the team wrote a bibliographic ontology with the hope that derivations from BIBFRAME found in this ontology will be folded into the official BIBFRAME namespace. The proposed changes include, but are not limited to, incorporation of the following conventions:

1. Reuse stable pre-existing classes and properties from external ontologies rather than declaring new ones within the BIBFRAME namespace.
2. Use URIs rather than strings to identify resources.
3. Replace the bf:Authority classes with Real World Entity classes for people, places, things, etc.
4. Define only one pattern to model one feature of the knowledge domain.
5. Clarify the directionality of properties via naming, definitions, and, where applicable, domain and range constraints, and add inverse properties where appropriate.
6. Name terms consistently, and make the distinction between classes, object properties, and datatype properties clear through standard naming conventions.

Efforts have been made to consider each class and property, but this ontology is largely untested. LD4L may provide revised/expanded versions in future as we identify new use cases, we begin to test the ontology with instance data, and BIBFRAME 2.0 revisions solidify. The RDF generated as an output of the project will be based on the LD4L ontology and will be made available for testing. For future consideration, there is a need to not only align the LD4L ontology with BIBFRAME 2.0, but also Schema.org (we've had early conversations with OCLC about this), the Doremus Project (using FRBRoo), Europeana's Data Model (EDM), and other RDF models within the bibliographic and cultural heritage domains.

Preprocessing Metadata for Richer RDF Conversions

Early experience in the project suggested preprocessing MARC data to include URIs for entities referred to records would significantly increase the value of the RDF output from the BIBFRAME converter. To date, in both the MARC specification and in cataloging practices, the use of URIs to identify entities is uneven. To this end, URIs in MARC: A Call for Best Practices was submitted to the Program for Cooperative Cataloging (PCC) discussion list in May of to spur conversation on the importance of URIs in library data and standardizing how they are stored in MARC. A number of outcomes have occurred out the document:

- The MARC Advisory Committee (MAC) discussed the paper with members of LD4L and recommended formal proposals be submitted for areas where MARC could better support the storage of URIs with consistent semantics.
- The PCC Task Group on URIs in MARC was formed to address immediate policy issues and develop a work plan with related guidelines in collaboration with MAC, library platform vendors and open source developers, and linked data experts (some from LD4L).
- In parallel with the PCC URI Task Group, an number of other organizations (e.g. the British Library) have begun to make MAC proposals related to URIs in MARC.

Because such a large portion of MARC records in libraries are provided by vendors, Nancy Lorimer has created a preliminary set of vendor specifications for enhancing MARC records. The specifications include both recommendations for the addition of URIs and for adjusting cataloging procedures to include specific fields and vocabularies that enhance the transformation into BIBFRAME. Adjustments to these specifications will be ongoing as the BIBFRAME converter is improved and modified to accommodate BIBFRAME 2.0. She and Phil Schreur will also be meeting with Casalini Libri at ALA Midwinter 2016 to discuss implementing these specifications.

Continued Work with Global URIs

In the fall of 2015 Linked Data for Libraries (LD4L) hosted a conversation with members of the CONSER, VIVO, and ISSN communities to discuss global URIs for serials. As the provider of unique identifiers that underpin the systems currently driving much of the continuing resources ecosystem, the ISSN network brings together a wealth of domain expertise and an established network of stakeholders. ISSN is well positioned to have a significant role in providing global URIs and linked data for serials.

The LD4L community is heartened to see the growing interest in modeling serial works as RDF with permanent, stable URIs. ISSN's current practice of creating identifiers for different serial entities (e.g., ISSN-L, e-ISSN, and p-ISSN, and ISSN for language versions) is very much in parallel with the task of minting URIs for different entity types. These practices provide a sound foundation for further progress. For example, a URI and modeling that explicitly collates all the various versions of a serial would be valuable to connect different languages and formats via a single identifier while preserving the full granularity of current ISSN practice.

We understand that there are number of initiatives underway or scheduled to happen that may have an effect on ISSN's work in this arena, including the CONSER Operations Committee's involvement with BIBFRAME and the 2016 review of ISO-3297. We hope there is a convergence on these fronts to ensure the most viable solution for establishing global URIs at the earliest and most authoritative moments in the lifecycle of a serial.

LD4L is particularly happy to hear that publishing URIs and linked data for serials is under consideration at ISSN. We have a strong interest in adopting and promoting ISSN linked data and would endorse further investigation of options for making complete, current, and open RDF descriptions of entities available to data consumers. We are confident that stable identifiers for ISSNs provided in linked data format will encourage widespread adoption of any URIs that ISSN creates. Ideally ISSN would adhere to Linked Data Best Practices, including:

- Provide persistent links.
- Provide useful data through standard protocols (content negotiation, SPARQL, etc.).
- Provide useful links to other entities. ISSN tracks how different serial entities relate to each other and their respective agents; without access to these relationships the data is far less useful.
- Provide the data according to model/s that meet data consumer needs; consulting the CONSER BIBFRAME Task Force and experienced linked data practitioners would be advisable for determining the model/s.

If ISSN continues to pursue a linked data agenda, LD4L is ready and willing to contribute to the discussion and testing. We look forward to ISSN's efforts in this space.

Ontology team activities to 2015

Local vs. global identifiers

The ability to directly link resources in our three libraries and to extend that linking arbitrarily in the future is a central premise of the LD4L project. Local resources and local authorities will continue to need stable identifiers, with the increasing expectation that these identifiers will be URIs directly dereferenceable from anywhere on the Web. These resources may be directly interlinked across institutions as special relationships are discovered, as for example between members of similar special collections across two or more libraries. However, we see OCLC's linked data initiatives in general and stable global identifiers for works in particular as an essential enabling resource that bring together multiple manifestations of a work into one entity. When local library resources share relationships to these global work identifiers, querying these relationships will reveal many further cross-library linkages that can significantly enrich local searches and collections, either on the fly or through deeper analysis.

However, realizing these goals will require scalable, publicly-accessible services for discovering works identifiers from locally-held WorldCat bibliographic identifiers. These services must also support mining the full content accessible through works records including their embedded linkages to other entities maintained at OCLC or through external authorities.

- OCLC Linked Data. OCLC Developer Network; accessed 2/8/2015.
- OCLC Releases WorldCat Works as Linked Data. News release, 28 April 2014.

Strings to things

Connecting library metadata with linked data 'in the wild' is a central goal of the LD4L project. To that end much of the ontology team's work has focused on identifying external authorities, stable identifiers (preferably URIs), and sources and services capable of linking the people, places, organizations, events, and subject headings in library metadata to real world entities. In some cases existing metadata in both MARC and non-MARC metadata includes references to local or external authorities, but the vast majority of potentially identifiable entities are represented only as strings of characters. Some of our catalog records have been linked to Library of Congress, OCLC (including the VIAF international authority file), or ISNI identifiers through contracts or internal record enhancement projects. A need to extend from authority file links or a registry of named entities to resolvable URIs compatible with linked data has motivated several LD4L investigations, with some focusing on quality and others more on the efficacy of existing services.

- International Standard Name Identifier (ISNI)
- Library of Congress Linked Data Service
- Virtual International Authority File
- ORCID
- Encoded Archival Description and EAD Linking Elements

Converting MARC to RDF

For MARC metadata, the team has worked with the Library of Congress BIBFRAME converter as a central component in a workflow that may include pre-processing to address variations in local MARC cataloging practice and in most cases will also require post-processing to produce data ready for consumption and interoperability with other linked data on the Web. While the conversions to BIBFRAME of a range of some 30 record types have been explored in concert with technical services staff at our three libraries, the ontology team has focused primarily on the availability and representation of data pertinent to the LD4L use cases rather than analyzing converter output across the board to ascertain completeness and correctness.

The classes and properties themselves in BIBFRAME, as well as some of their definitions, remain under active discussion on the BIBFRAME mailing list (archives) and in other venues. With our project's strong focus on linking through to real world entities, we remain flexible in our interpretation and application of the BIBFRAME ontology, in some cases electing to use properties and/or classes in an LD4L namespace until such time as consensus has been reached in later releases or through pilot projects scheduled for 2015 and/or community practice. Fundamental questions will continue about distinctions between information and real world entities and conflicts between a desire to retain all the information encoded in MARC records vs. allowing bibliographic metadata to more freely inter-operate with other Web data.

- *Common Ground: Exploring Compatibilities Between the Linked Data Models of the Library of Congress and OCLC*. Jean Godby and Ray Denenberg. This provides a high-level comparison between the LoC BIBFRAME approach and the work that OCLC has been doing on expressing bibliographic metadata in Schema.org.
- *The Relationship between BIBFRAME and the OCLC's Linked-Data Model of Bibliographic Description: A Working Paper*. Jean Godby, Senior Research Scientist, OCLC Research, September, 2013.
- Bibliographic Framework Initiative
- Technical site for the Bibliographic Framework Initiative (bibframe.org)
- BIBFRAME primer document (PDF)
- BIBFRAME master RDF file (December 10, 2014)

Addressing complexity

Several levels of complexity may legitimately exist in parallel and be utilized based on the availability of data or the goals of an application. This choice can be seen in PROV-O ontology where direct object properties have been paired with more complex options involving intermediate nodes that add additional temporal or role information. The related PAV (Provenance, Attribution, and Versioning) ontology offers a simpler set of classes and properties sufficient for many applications requiring only simple attribution. Application software can also often mask a more complex underlying data model, and in many cases it may be preferable in production contexts to separate logging and provenance information from user-facing applications entirely.

- PROV Ontology
- Prov-O Ontology on the VIVO wiki
- PAV (Provenance, Authoring, and Versioning) (on Google Code)

Working with non-MARC metadata

While our library catalogs are very likely the largest single sources of metadata, each partner university maintains a large number of digital collections representing a diversity of subject domains, size, and complexity. Several of our use cases involve connecting catalog data with these non-MARC sources, not only to provide a more unified search interface, but to be able to interconnect and cross-references sources that for now remain almost entirely separate. The benefits go both ways, and the addition of sources outside the traditional library domain brings in yet more possibilities for value-added services enhanced by entity recognition and external links. Prime examples of these non-library sources are Stanford's Profiles (CAP Network software), Harvard's Faculty Finder (Harvard Catalyst Profiles software), and Cornell's VIVO (VIVO software and VIVO-ISF Ontology).

Two LD4L investigations are focusing on identifying external identities and resource types in visual resource metadata during a process of conversion of that metadata to BIBFRAME for compatibility with catalog records expressed in RDF. The metadata for Cornell's collection of Hip Hop flyers, encoded using VRA Core, was selected for a pilot because of the number and range of external references, including musicians, illustrators, events, geographic locations, and other works by the artists involved. The Karma data integration tool from USC's Information Sciences Institute provides a graphical user interface for constructing conversions of VRA to RDF. Harvard has an extensive Visual Image Access system with a metadata schema similar to VRA Core, and Paolo Ciccarese has developed a suite of workflow tools (on GitHub) to serve as a pipeline for both entity resolution and conversion to RDF, including selection of types for visual resources drawn from the Getty Art and Architecture Thesaurus.

- VRA Core data standard
- Harvard's Visual Image Access (VIA) system (VIA Schema information)
- Getty Vocabularies as Linked Open Data
- The DBpedia Ontology (2014)

- Linked Data in VRA Core 4.0: Converting VRA XML Records into RDF/XML (Jeff Mixter)
- PBcore – Public Broadcasting Metadata Dictionary Project

Annotations and virtual collections

The first two use cases address user tagging and the ability of librarians or others to curate potentially very large collections of library resources through annotations external but linked to the bibliographic metadata. Existing ontologies were identified that support annotations and the assembly and ordering of individual resources into collections.

- W3C Open Annotation Data Model
- Open Archives Initiative Object Reuse and Exchange | ORE Specification - Vocabulary
- LD4L Use Case 1.1 - Triples Examples
- LD4L Use Case 1.2 - Triples Examples

Usage data

The fifth group of use cases explore including usage data to supplement library discovery interfaces and to inform collection review and additions. Here the team first explored a very granular model for capturing usage information from circulation-related events and other direct user interactions with library resources. On further investigation, however, this data proved not only to be difficult to come by but fraught with concerns about privacy, even when stripped of any directly identifying information. Later discussions have focused on the compilation and use of a simple 'stack score' usage metric on a percentage scale potentially more comparable across institutions despite differences in size, discipline, population makeup, and other factors.

- David Weinberger: "A Good, Dumb Way to Learn from Libraries" from the Chronicle of Higher Education, October 7, 2014. This provides motivation and explanation for a simple usage metric that protects user privacy.

Expressing bibliographic metadata to support discovery

The LD4L Use Cases largely target ways to supplement traditional library catalog metadata, whether through linkages to external identities and resources, connecting catalog records with other digital collections, or adding usage metrics and annotations. Some use cases suggest new functionality by leveraging this new "library graph" in services that go beyond text-based matching of search terms to suggest deeper connections and even loop back from externally linked entities to additional local resources or related resources in other libraries.

In parallel with these efforts, the LD4L project also seeks to more reliably populate facets such as genre, format, uniform title, subject, and online availability for library resources. Indexing MARC metadata for search now typically requires complex case statements reflecting variations in local cataloging practice by discipline or over time, and resources from other digital collections may not even include the appropriate information. We see conversion to RDF using ontologies as an opportunity to express these and other user-facing facet values explicitly, thereby reducing reliance on 'black box' processing and increasing the interoperability of metadata and re-usability of tools across libraries despite historical differences in local cataloging practice.