Project Description (LD4L Labs)

Table of Contents

C. Project Description

1. Overview

The specific proposed work falls into four broad areas:

- Developing linked data creation and editing tools based on the Hydra framework and other proven implementation approaches such as the Vitro[1] editor, which is part of the widely used VIVO faculty profiling system[2], and eagle-i[3], a tool developed to support shared description of open science resources (section 2).
- Exploring strategies to use linked data relationships and analysis of the graph to directly improve discovery and understanding of relevant scholarly information resources. The project team will test these strategies with specific sets of collections and users (section 3).
- Providing feedback for BIBFRAME ontology development and piloting efforts in reconciliation and URI persistence (section 4).
- Supporting metadata conversion tool development for the LD4P Partners and the broader library community (section 5).

The project will continue the effort that the LD4L project has begun on providing constructive and influential feedback to the Library of Congress on the BIBFRAME initiative, supporting their goal that BIBFRAME continue to evolve over time. The project will also work with other libraries, archives, and museums in the LD4P community and beyond to standardize on shared, compatible ontologies beyond BIBFRAME and to demonstrate how the additional Linked Open Data that this approach makes available can improve description, discovery, and understanding of scholarly information resources.

The project will also seek to collaborate with linked data efforts that are making use of Schema.org[4], such as those at OCLC[5] and UIUC[6], to promote alignment between the two standards. Quoting the description of the relationship between BIBFRAME and Schema.org from Stanford's LD4P proposal:

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.

All tools and ontologies from the project will be made freely available with permissive licenses. The project partners will provide the linked data produced by the project as Linked Open Data on the Web, with resolvable URIs. Acknowledging URIs lose persistence through simple inattention to server, repository, application, and/or website changes that render online addresses unreachable, the project team will explore approaches to making these URIs persistent and reusable for the long term. Institutional commitments will be required together with appropriate policies and procedures to persist URIs for organizations that have merged or otherwise disappeared, and to maintain individual URIs for people who have left the institution. Including links to global identifier initiatives including Library of Congress name authorities, VIAF, ISNI (International Standard Name Identifier)[7], and ORCID (Open Researcher and Contributor ID)[8] will boost confidence in local records and reduce maintenance costs.

Solving the question of persistence may involve partnering with existing organizations, such as OCLC and the Library of Congress, that are committed to the long-term maintenance of identifiers; however, the project team is not yet certain of either organization's interest in providing services around persistent URIs to individual institutions working outside of traditional MARC workflows, and business models for these activities do not yet appear to exist at either OCLC or LC. Partnership with these organizations might simply mean drawing on their experience to ensure persistence locally, with similar patterns and institutional commitments by the partners themselves to persist and support these URIs indefinitely

The project will also ensure URIs are made fully linked data compatible through HTTP content negotiation, a protocol for returning content not as formatted web pages (HTML) but in structured data formats such as RDF/XML or JSON-LD. The project partners will also incorporate the results of entity resolution and reconciliation to reduce duplication within catalog data and link with established identifiers already available and becoming available over the life of the project.

The project will continue to build on the work of the LD4L project in exploring the issues of scalability, distribution, and location of triples and triplestores[9] through adoption and testing of both open source and proprietary triplestores offered with commercial support. The former offer greater immediate promise for wide adoption in the library community, but the rapidly expanding marketplace for triplestore, reasoning, and search services has not yet settled on a single de facto solution to address all aspects of performance at web scale. Centralized models present easier service models for query and update, but distributed models fit library needs to continue to manage local knowledge assertions locally. Commodity search tools such as Apache Solr[10] and Elasticsearch[11] are examples of technologies that will likely bridge between triplestores and web discovery and retrieval front-end applications while more robust, web-scale semantic technologies emerge in the next decade.

LD4L Labs will continue to leverage the best of breed tools available, including commercial tools where appropriate academic licensing can be arranged, in order to focus on what can be accomplished and shared widely in the library world in the next 3-5 years. The issue of rapid technology evolution will be a particular concern for the work on improving discovery and understanding, but it will also be relevant to the work on tool creation and entity resolution /reconciliation.

While the focus of the LD4L Labs project is not on ontology development, many of the tools and processes being developed require and are dependent on library-relevant ontologies. To address the project's ontology needs, the project team will seek to reuse existing ontologies and vocabularies (e.g. Open Annotation, BIBFRAME, FOAF, Event Ontology), building on the model from the LD4L project of both adopting external standards and providing guidance and feedback to other linked data projects to make their work more linkable and reusable.

Ontology reuse will in most cases contribute to the adoption of linked data and ease the development of applications and services relying on interoperability of data. Care must be taken, however, to review the definitions, assumptions, and logical entailments embedded in existing ontologies to avoid the assignment of unintended meaning in new situations. As a corollary, the practice of ontology reuse must also avoid ontology "hijacking," where the original definitions or logical implications of an existing ontology are altered in the new context. The project team made significant progress on ontology reuse and development during the initial LD4L project; this work will be leveraged by LD4L Labs, which will minimize the investment of new effort required for ontology development. Much of the ontology reuse and development effort will occur as part of the LD4P community efforts with the Rare Materials and Music communities and will target specific issues of concern to those domains, such as item-level data.

If the outputs from LD4L Labs are to be used and useful, it is important to work with existing collections of scholarly information resources to support grounded, concrete work on tool development, piloting metadata production practices, and experiments in improving discovery and understanding of those resources using linked data. Some of this work can be carried out on the aggregated catalog data developed as part of the original LD4L project, which is continuing to be enhanced and maintained by the original partners. However, other development work will be easier within focused collections with specific characteristics.

The complementary nature of the LD4P linked data collections and the LD4L Labs tools and services is the clearest opportunity for synergy between the two projects. As LD4L Labs develops tools and services for linked data creation and editing; metadata conversion; visualization, discovery, and understanding; annotation; and much more, the LD4L Labs team will make use of the linked data collections and projects being developed by the LD4P Partners. In some cases, the project description below references specific LD4P collections as examples that will be used to ensure that the tools and services being developed by LD4L Labs address the concrete needs of the libraries that are working to pilot linked data production efforts.

---

[1] https://github.com/vivo-project/Vitro
[2] http://vivoweb.org
[3] https://www.eagle-i.net/get-involved/for-developers/
[4] http://schema.org/
[5] https://www.oclc.org/news/releases/2012/201238.en.html
[6] https://www.lis.illinois.edu/articles/2015/12/mellon-foundation-supports-cirss-research-linked-open-data-digitized-special
[7] http://www.isni.org/
[8] http://orcid.org/
[9] https://en.wikipedia.org/wiki/Triplestore
[10] http://lucene.apache.org/solr/

[11] https://www.elastic.co/products/elasticsearch

---

2. Linked Data Creation and Editing

To support creating new linked data beyond the conversion of legacy MARC metadata to BIBFRAME, the project team needs to develop tools and approaches that can be used by librarians, scholarly end-users, and software developers. This section describes several different proposed research and development efforts to enable the creation and editing of new linked data about scholarly information resources.

2.1 Leveraging annotations to support discussion, tagging, organization, and crowdsourcing

The LD4L project explored the use of annotations for the support of virtual collections, commentary, and tagging. These LD4L annotations follow the W3C Open Annotation[12] standard; as semantic entities in their own right, they maintain a separate identity from the described resource for clarity in filtering or aggregation. This project will extend this work both in terms of the annotation purpose and the target systems. The project partners will:

- Extend work on the ActiveTriples[13] gem, part of the Hydra framework[14], to support the creation of new organization, curation, annotation, and usage data as linked data compatible with the ontologies adopted by LD4L (e.g. Open Annotation, BIBFRAME/LD4L, FOAF, Event Ontology). ActiveTriples will be used as the basis for tools to support annotation of scholarly resources by librarians, catalogers, and other information professionals. The tools will support a range of annotation types (e.g. tagging, supplemental description, and discussion and review) and we will demonstrate integration of these annotations into discovery systems;
- Go beyond providing tools for information professionals and extend Hydra-framework tools to support BIBFRAME/LD4L-ontology-based crowdsourced curation and annotation by scholars, students, and subject matter experts. Such curation and annotation will include appropriate provenance to identify the person or group making a selection, enhancement, or annotation; the time and date of the activity reflected; and the

specific resource targeted. The partners will create Hydra-based widgets that support annotations using restricted vocabulary tags, folksonomy tags, and free text. The prototype Triannon[15] annotation store developed by Stanford will be used as the storage engine, demonstrating integration of W3C Linked Data Platform[16] (LDP) and Open Annotation compliant storage with local library systems. This work extends previous work in LD4L on Use Case 1.2 (Tag Scholarly Information Resources to Support Reuse).

2.2 Conversion of non-MARC data

The LD4L project team created a conversion pipeline that included pre- and post-processors wrapped around the existing Library of Congress MARC to BIBFRAME converter[17]. This pipeline drew on MARC and other data sources to create linked data that represented basic bibliographic information about those resources, and that did a limited amount of entity reconciliation both within the conversion process and to external Linked Open Data entities. An early deliverable from the LD4L Labs project is a production-quality MARC to BIBFRAME/LD4L converter and extensible conversion framework (see section 5.3 for more on that project). Later in this project, the partners propose extending this new converter and pipeline, and performing some level of reconciliation for entities referenced in non-MARC metadata, including holdings, item, and locally defined metadata, as well as other subject-focused metadata standards.

Though there are similarities between this portion of the LD4L Labs proposal and the Mellon-funded project at UIUC on developing LOD for special collections materials, there are also some notable differences. The UIUC project aims to transform legacy metadata into Schema.org rather than BIBFRAME. Over the coming years ontology alignment between BIBFRAME and Schema.org will go a long way toward making this data interoperable. The two projects are also focusing on very different types of special collections and source metadata for conversion, presenting different conversion and modeling challenges. The LD4L Labs and UIUC projects complement each other, and the projects will establish lines of communication to ensure that there is alignment between the two ontology approaches wherever possible. Chew Chiat Naun, LD4L Labs and LD4P participant, serves on the UIUC project advisory board and will be one connection point between the two projects. Further, Tim Cole (UIUC) has been involved in LD4P planning meetings and LD4L Labs team members plan to invite UIUC participants into LD4L Labs discussion as relevant.

2.3 Linked data creation and editing environments

The creation of new linked-data descriptions for unique collections across a wide range of disciplines will require user-friendly editing environments that help catalogers adopt the appropriate ontologies and find appropriate repositories of entity resources. While tools such as Protégé[18] provide sophisticated ontology creation and editing environments, they are complex and not well suited to the entry of instance data. The project team has experience with the Vitro and eagle-i tools, which were developed at Cornell and Harvard, respectively. These are easily configurable and customizable semantic web content editing tools that adapt to a set of ontologies installed in them and focus primarily on adding content (instance data). This feature makes them good candidates for editing linked data according to collection-specific ontologies, while still allowing for experimentation and extension of the installed ontologies.

Vitro is a general-purpose web-based ontology and instance editor with customizable public browsing. Vitro was originally developed at Cornell University, and is used as the core of the popular research and scholarship portal, VIVO. Vitro is an integrated ontology editor and semantic web application implemented as a Java web application that runs in a Tomcat servlet container. With Vitro, you can: create or load ontologies in the Web Ontology Language (OWL)[19] format; edit instances and relationships; build a public web site to display your data; and search your data with Apache Solr.

eagle-i is a distributed system for creating and sharing semantically rich data. It is built around semantic web technologies and follows Linked Open Data principles. eagle-i focuses on biomedical research resources. However, thanks to its ontology-centric architecture, the platform can be adapted to other domains. The eagle-i platform utilizes a variety of tools for repository data ingest, broadly referred to as data tools. The different tools share a common layer for interacting with the eagle-i repository:

- The SWEET (for Semantic Web Entry and Editing Tool) is a web application for manual data entry and curation developed using the GWT (Google Web Toolkit). Its core component is a dynamic forms generation module that translates ontology axioms into UI data entry widgets. The SWEET generates a form per ontology class and thus allows users to create instances of an ontology class. Navigational elements of the SWEET include workflow controls and instance listing and filtering.
- The SWIFT ETL (Extract, Transform, and Load) toolkit provides command line tools for transforming tabular data into eagle-i linked data instances, and for loading them into an eagle-i repository. The data mappings necessary for performing the transformations are captured in an RDF map, in an approach heavily influenced by the RDF123 tool.
- A data management toolkit provides command line tools for performing bulk modifications on instances in a repository, and in particular for migrating existing instance data to conform to a newly released ontology. A data management front-end to be used by data curators is under active development.

The LD4L Labs project team will provide library technical services and digital collections catalogers with Vitro and/or eagle-i based linked data editing, display, and dissemination environments that will support the creation and incorporation of subject and collection-specific ontologies to describe the unique aspects of the collection in a structured, extensible, and shareable manner. The tools will also support easy linking to existing external linked data vocabularies and published globally-resolvable entities (e.g., Getty[20], FAST[21], WorldCat Entities[22], LinkedBrainz[23], and Digital Science GRID[24]).

The project team will evaluate and compare the two tools for this use. The Cornell team will support the use of the Vitro tool by LD4P Partners in their metadata production pilots. These pilots will specifically include the Afrika Bambaataa Collection at Cornell and the Columbia Art Properties Collection, but may include other LD4P efforts as well. The Harvard team will explore the use of eagle-i for bibliographic records converted to BIBFRAME from their current catalog and new BIBFRAME records created as part of the Harvard Film Archive (section 5.1) and Geospatial Metadata (section 5.2) projects. Both the Vitro and eagle-i metadata production approaches will be informed by the use cases developed by other LD4P Partner projects, and by the production workflow requirements being developed by the Stanford Tracer Bullets project.

2.4 Extending Hydra Tools to Support Linked Data

While using Hydra-based tool sets to gather full sets of linked data would require extensive manual configuration and coding, there is an opportunity to extend existing Hydra tools to add or expose specific organizational and curation linked data related to digital objects. The project team will:

- Extend the Spotlight[25] exhibit tool to include remote resources with stable identifiers and described by ontology-compatible descriptions into exhibits. This effort builds on LD4L Use Case 1.1 (Build a Virtual Collection) previously demonstrated within the Hydra framework.
- Extend Sufia[26], a Hydra framework gem providing self-deposit institutional repository features, to use and publish linked data. We will work with the very active Sufia community to build consensus on approaches to extend Sufia's current resource-centric metadata model to leverage richer

linked data descriptions and to better use external data sources. Specifically, we will develop Ruby gems that use external data sources in order to allow users to find and record identities (linked data URIs) where strings have traditionally been used. For example, geographic places have established external URIs in linked data sources such as DBpedia and contributors have external URIs in profile systems such as VIVO, but current repository systems lack convenient ways to connect to these identities and the data behind them. Work on ActiveTriples[27] will be extended for use with Sufia and the project team will consider linked data storage strategies, including use of a triplestore as an alternative to Fedora.

---

[12] http://www.openannotation.org/spec/core/
[13] https://github.com/ActiveTriples/ActiveTriples
[14] http://projecthydra.org/technical-2/
[15] https://github.com/sul-dlss/triannon
[16] http://www.w3.org/TR/2015/REC-ldp-20150226/
[17] https://github.com/lcnetdev/marc2bibframe
[18] http://protege.stanford.edu/
[19] https://en.wikipedia.org/wiki/Web_Ontology_Language
[20] http://www.getty.edu/research/tools/vocabularies/lod/
[21] http://fast.oclc.org/
[22] https://www.oclc.org/developer/develop/linked-data/worldcat-entities.en.html
[23] http://linkedbrainz.org/
[24] https://www.digital-science.com/products/grid/
[25] https://github.com/sul-dlss/spotlight
[26] https://github.com/projecthydra/sufia
[27] https://github.com/ActiveTriples/ActiveTriples

---

3. Improving Discovery and Understanding

The creation and publication of rich linked data opens up new opportunities for connections from library sources to external entities, but further analysis and innovative new discovery services will be required to fully realize the potential value. In presenting linked data to end users, the project will take a use-case driven approach to deciding how much linked data to follow and expose, which may require exploratory analysis as well as presentation of options that allow users to explore on their own. The project team will experiment with pulling in data from authorities connected to reconciled entities; following temporal and place-based associations; exploring different levels of aggregation; and working with annotations or suggested relationships added directly by end users. The project team will explore options on how to present a range of available relationships, from direct connections such as author affiliations on to indirect connections including common linkages to subject concepts, places, or time periods, as may be present in MARC name authorities and other sources such as Wikidata[28].

The project partners further propose to evaluate the frequency of adoption of new discovery affordances as an at least partial indicator of their effectiveness, through A/B testing and iterative, small-scale user studies. While this project will not have resources to undertake comprehensive enrichment of catalog discovery systems in all domains, targeted implementation and testing will help identify promising approaches and their key characteristics.

3.1 Visualization and exploration of related entities

- The project will explore using standard linked data APIs, specifically the W3C Linked Data Platform (LDP), to enable tools and frameworks to access and create LD4L linked data. LDP compatible tools will constitute the architecture for access by both the Hydra framework and discovery platforms such as Omeka[29] and Blacklight[30].
- The project will support browseable visualizations of the related entities around a scholarly resource by developing standardized building blocks, potentially based on existing Javascript graph libraries, that utilize LDP APIs to retrieve the linked data graph. Visualization modules compatible with both Blacklight/Spotlight and Omeka will be developed.
- The initial LD4L effort piloted linked data connections to real-world objects and their metadata in the broader Linked Open Data cloud (e.g., DBPedia[31], Wikidata[32], MusicBrainz[33]/LinkedBrainz[34]). This follow on will build on LD4L Use Case Cluster 3 (Leveraging External Data Including Authorities) by leveraging external URI connections and their associated linked data contexts to improve discovery and understanding via new tools helping users intuitively explore via these additional contextual relationships. Project team members will conduct small-scale user studies to evaluate how well these approaches work in practice.
- Toward the end of this project, project team members will begin initial explorations of how to use the crowdsourced description, recommendation, and annotation data available through Triannon and other sources to improve discovery and understanding of scholarly information resources. This work will build on the LD4L project's Use Case Cluster 1 (Bibliographic + Curation Data) as well as recent work by the Shakeosphere team[35].

3.2 User interfaces exploiting the complex graph

This project will demonstrate how to use LD4L-ontology-based linked data to describe, annotate, and organize collections. Work will focus on The Afrika Bambaataa Collection, an LD4P Partners project, a highly-structured collection that combines many annotations found on physical items (often related to provenance and other notations provided by the artist) and links between items with links to resources (images of LPs and related materials) in a repository system, as well as links to external musical resources, person, event, and location information. The Afrika Bambaataa Collection was selected as it comprises a Cornell LD4P native linked data production project and will have both RDF and MARC data created for similar materials; the MARC data will not be created under the LD4L Labs or LD4P funding streams. Having natively-created RDF and MARC for a set of materials provides a good test case for understanding the impact of linked data on user discovery. The interfaces developed will be evaluated with user experience studies and will focus on use cases described by scholars in this area.

3.3 Network analysis

This project will explore using network analysis of the bibliographic, user, and usage information in the linked data graph within and across institutions to include graph-derived information about the resources as augmentations to a search index, and provide better context for scholarly information resources in results lists, particularly Use Case Clusters 2 (Bibliographic + Person Data) and 4 (Leveraging the Deeper Graph via Queries and Patterns). In some cases this analysis of the graph may result in presentation of additional facets such as geography, while in others it may primarily augment search result snippets and item record displays with additional links.

The project team will evaluate the resulting user experience when data is added to Cornell's Blacklight-based search, and share both code and results with the broader Hydra/Blacklight community. The project team will also explore other more native linked data tools, such as the D3.js[36] Javascript Library, for network analysis, visualization, and discovery as alternatives to using Blacklight. This effort will make heavy use of the combined Linked Open Data catalogs created by Cornell, Harvard, and Stanford as part of the original LD4L project, especially as later versions with more comprehensive external entity resolution and reconciliation develop over the life of this project.

---

[28] https://www.wikidata.org

[29] http://omeka.org/
[30] http://projectblacklight.org/
[31] http://wiki.dbpedia.org/
[32] https://www.wikidata.org/
[33] https://musicbrainz.org/
[34] http://linkedbrainz.org/
[35] http://shakeosphere.lib.uiowa.edu/annotation.jsp
[36] http://d3js.org

---

4. Ontology Work, Reconciliation, and Persistence

4.1 Revisions to the BIBFRAME ontology

The ontology used by the LD4L project incorporates a number of significant differences from the 2014 BIBFRAME ontology which were necessary to follow linked data best practices, as articulated and reported to the Library of Congress in April 2015 by Rob Sanderson and other LD4L project team members. A major revision of the BIBFRAME ontology is expected in 2016, which will converge with the revised entity (class) and simplified attribute and relationship (property) structure of the ontology used by the LD4L project based on ongoing feedback from discussions with the Library of Congress and on the BIBFRAME list[37], including clear distinctions between real world entities (e.g., people, organizations, events, places) and "authorities" (information describing those people, organizations, events, or places). The clarifications and refinements advocated and modeled by the LD4L work and promoted through the Sanderson report to the Library of Congress do much to promote direct connection and interoperability for library data with Linked Open Data on the Web. The LD4L ontology definitions will be revised to exploit this convergence as much as possible to reduce duplication of terms, while also continuing to assemble and, as needed, develop a standard set of ontologies to meet the needs of the broad scope of resources and relationships under consideration.

As part of the LD4L Labs project, team members plan to use the new release of BIBFRAME in many of the tools and services being developed. Based on previous experience with the LD4L project and the VIVO project before it, the needs of the tools and services and their corresponding use cases may well suggest future changes and extensions to BIBFRAME. LD4L Labs team members will work with the Library of Congress, the LD4P Partners, and the BIBFRAME community to articulate these proposed changes and to vet them with the broader library linked data community.

4.2 Reconciliation and persistence

One important model that emerged from the work with LD4L was that of first curating and describing the resource locally, expressing the unique knowledge of the institution that holds the resource, and then linking out to related real-world objects on the Web. In some cases, establishing this connection will involve reconciliation of a local entity with an external one, and in others, it will simply be the assertion of a relationship. For this model to be useful, it is important to define the ontology and create relationships that actually add value for users. This ontology-driven approach will also be used in the creation of linked data for LD4P partner projects.

A key benefit to globally accessible LOD is the unambiguous reference to an entity through a persistent URI. Iowa's CTSAsearch[38] has already demonstrated the utility of supporting disambiguation of author identity in publications by cross-matching VIVO-compatible data from multiple research profiling platforms and providing a web service indicating the home URI for known coauthors of a publication. Iowa will work with other team members to map these tools into the LD4L framework and provide a pilot web service for project participants and others to query for both global persistent URIs for resources and for URIs based on local extensions to the BIBFRAME ontology for additional granularity in describing types and relationships, expressing knowledge present at the institution itself through unique bibliographic resources or subject-based knowledge. The project team will also work to extend this service to develop an infrastructure to register "sameAs" and "closely related" relationship assertions for both LD4L Labs and LD4P partners. This shared registry should be managed locally in distributed but coordinated workflows that would rely on regular automated harvesting to one or more shared indexes of "sameAs" and "closely related" relationship assertions tied to lookup services across large numbers of organizations, as pioneered by the CTSAsearch at the University of Iowa.

A central focus on this work will be the exploration of the architectural trade-offs between centralized approaches (where data from source sites are harvested and analyzed by the service site) and distributed approaches (where data remain at sources sites and are queried on demand by the service site). The overall goal will be to support and simplify the processes of both resolution (mapping metadata strings to URI-based entities) and reconciliation (tracking assertions of equivalence among URI-based entities). In most cases, resolution is a more local task depending on local knowledge and expertise, with reconciliation functioning more globally through automated services that suggest appropriate matches and reflect consensus in ways that support user-facing services without losing track of antecedent sources as a means of correcting errors and resolving discrepancies over time. Hence these processes are roughly equivalent to the local formulation and formalization of authority records (resolution) and merger of local authority records into union catalog entries (reconciliation). CTSAsearch currently offers similar services in support of research profiling "sameAs" assertions among institutional profiling systems.

Ontologies are extended and updated as the needs and understanding of the community changes, optimally as distinct versions indicative of the scope of change[39]. As much as possible, it is important to maintain the semantics of linked data over time even when ontologies and entities themselves are updated. It is equally important to communicate to users of the data or those building applications when and to what extent a new version of an ontology or an application dependent on it differs from a previous version. While this problem is a general one, the project team must work with the community to manage shared ontologies and linked data representations of library resources appropriately going forward. This challenge will be addressed in part by the development of new production processes via the proposed LD4P project, but the same concerns are very relevant to non-MARC sources and to the tools and discovery services built to take advantage of linked data.

---

[37] http://listserv.loc.gov/archives/bibframe.html

[38] http://research.icts.uiowa.edu/polyglot/

[39] http://semver.org/

---

5. Tools for Metadata Conversion

In many cases, the best opportunity for creating library Linked Open Data is the conversion of existing metadata. The projects in this section cover the development of tools for metadata conversion both for specialized collections and for standard MARC records. This work continues the efforts of the first LD4L project in developing a conversion pipeline that combined pre-processing and post-processing steps with the existing Library of Congress MARC->BIBFRAME converter to produce records compatible with the ontology adopted by the LD4L project. The experience gained in engineering that pipeline and scaling it up to convert some 23 million MARC records from Cornell, Harvard, and Stanford to Linked Open Data will serve as a foundation for the work proposed here.

The LD4L Labs team will work closely with the LD4P Partners on the development of these conversion tools. It will be important that these tools fit smoothly into the broader metadata production workflows and environments being developed as part of the LD4P project.

5.1 Geospatial Datasets and Geospatial Images

OpenGeoMetadata[40] is a collaborative, shared GitHub repository hosting several thousand XML-based metadata records describing geospatially-enabled data resources held by participating members. This project will focus on converting a subset of OpenGeoMetadata metadata records, from the Harvard Geospatial Library[41], Stanford EarthWorks[42], and the Cornell University Geospatial Information Repository[43], into linked data descriptions using BIBFRAME/LD4L as a base ontology. Deliverables for the project would include: a BIBFRAME/LD4L profile for geospatial datasets; a set of mapping rules for Federal Geographic Data Committee (FGDC) geospatial metadata standards[44] to the BIBFRAME/LD4L profile; reconciled linked data entities in the source metadata for Originators, Place and Theme keywords, and series works; a linked data triplestore with published descriptions; and a user interface for searching and visualizing geospatial dataset descriptions.

5.2 Harvard Film Archive (HFA)

The HFA project within LD4L Labs will explore and assess the issues in converting legacy metadata for moving image resources to linked data. The project will additionally explore the issues in making that linked data useful for research and discovery. Metadata conversion tools will be developed that create linked data descriptions for a variety of formats (film prints, negatives, DVDs, VHS, Super 8, and others) and content (feature films, trailers, home movies, ethnographic films, propaganda) and related archival materials (including production elements, artwork, film stills, and promotional ephemera) held by the Harvard Film Archive (HFA)[45]. As the tools are developed, the project will assess BIBFRAME/LD4L's effectiveness as a data model for describing moving image materials for research needs, and identify specific vocabularies for description of these materials in a linked data environment. The HFA project will create mappings for records from the HFA's film print database, focusing on a subset of moving image materials by women directors (work that has previously been underexposed and in many cases is unique to this collection). Wherever possible, entities will be reconciled to linked data URIs, including personal and corporate names (ISNI, LCNAF), place names (GeoNames), genres (LC genre/form, Getty AAT), and works.

Specifically, using a subset of the collection aimed at women filmmakers, the HFA project will:

1. Create and pilot a software tool for converting existing HFA metadata into BIBFRAME RDF, including entity resolution as described below
2. Identify BIBFRAME extensions, if any, and propose and implement these in the converter
3. Load the HFA linked data into the Harvard linked data infrastructure to support SPARQL queries defining relevant subgraphs
4. Create or repurpose a linked data visualization tool for displaying relevant portions of the HFA RDF graph for each described item
5. Assess the end-user value for discovery and research of the HFA linked data, including a written summary of project findings disseminated to appropriate moving image and linked data communities with a set of recommendations for future research and development.

5.3 MARC -> BIBFRAME Converter

One key conversion tool for many linked data projects will be the implementation, in collaboration with the Library of Congress and the BIBFRAME community, of a new, community-sourced and supported MARC to BIBFRAME converter based on the revised BIBFRAME 2.0 specification, which is expected to be released in 2016. In collaboration with the LD4P Partners, LD4L Labs developers will create the initial version of this new converter. The goal is to produce a robust, efficient, well-documented, well-tested, open-source converter that can be easily used and adapted by many different libraries for their conversion needs. The LD4L Labs team will engage with the BIBFRAME community on how best to address the ongoing development and support for this converter beyond the initial implementation.

5.4 BIBFRAME -> MARC Converter

The project team will explore the development of a BIBFRAME to MARC converter. The goal here would not be to capture all that can be expressed in either BIBFRAME or MARC, but instead to produce a “core” or “operational” MARC record that can be used for workflow with legacy systems. Preliminary analysis of production workflows indicates that this will be an essential utility for as long as libraries may be rooted in both MARC and BIBFRAME-based description (i.e., for the foreseeable future). Team members will coordinate this effort with the LD4P Partners, the Program for Cooperative Cataloging (PCC), and OCLC, which has expressed an interest in building such a tool, but has provided no target date.

---

[40] https://github.com/OpenGeoMetadata
[41] http://hgl.harvard.edu
[42] https://earthworks.stanford.edu/
[43] http://cugir.mannlib.cornell.edu/
[44] https://www.fgdc.gov/metadata/geospatial-metadata-standards
[45] http://hcl.harvard.edu/hfa/
