Statements of Work (LD4L Labs)

Table of Contents

6.2 Cornell Statement of Work

Cornell University Library will take the lead on the *Linked Data for Libraries: LD4L Labs* project, providing overall project management and guidance, scheduling and convening project meetings, managing the project mailing lists, website, and wiki space, and, with the assistance of all project participants, reporting to Mellon and the library community on project activities and results. In addition, Cornell staff will undertake the following specific activities:

1. Leveraging annotations to support discussion, tagging, organization, and crowdsourcing. Extend work on the ActiveTriples gem, part of the Hydra framework, to support the creation of new organization, curation, annotation, and usage data as linked data compatible with the ontologies adopted by LD4L. Extend Hydra-framework tools to support BIBFRAME/LD4L-ontology-based crowdsourced curation and annotation by scholars, students, and subject matter experts (section 2.1).

Budget:

- 25% of Lynette Rayle

2. Conversion of non-MARC data. Extend the MARC->BIBFRAME converter and pipeline to support conversion of and reconciliation for entities referenced in non-MARC metadata, including holdings, item, and locally defined metadata, as well as other subject-focused metadata standards (section 2.2).

Budget:

- 5% of Chew Chiat Naun
- 10% of Rebecca Younes
- 10% of Muhammed Javed

3. Develop specifications for and support the Vitro linked data editing system. Cornell staff will provide library technical services and digital collections catalogers with Vitro 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 (section 2.3).

Budget:

- 10% of Muhammed Javed
- 5% of Chew Chiat Naun
- 20% of Rebecca Younes

4. Extending Hydra Tools to Support Linked Data. Extend the Spotlight exhibit tool to include remote resources with stable identifiers and described by ontology-compatible descriptions into exhibits. Extend Sufia, a Hydra framework gem providing self-deposit institutional repository features, to use and publish linked data (section 2.4).

Budget:

- 25% of Lynette Rayle

5. Improving Discovery and Understanding. 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. Explore using 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. 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. Demonstrate how to use LD4L-ontology-based linked data to describe, annotate, and organize collections. 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 (section 3).

Budget:

- 35% of Cornell LD4L Labs Developer
- 5% of Chew Chiat Naun
- 10% of Muhammed Javed

6. Ontology Work, Reconciliation, and Persistence. Revise LD4L ontology specifications to address changes in BIBFRAME 2.0. Work with the Library of Congress and the LD4P Partners to propose future changes to BIBFRAME suggested by work on the LD4L Labs tools and services, and vet these changes with the broader library linked data community. In close collaboration with David Eichmann and his colleagues at Iowa, advise on tools and approaches for resolution and reconciliation. Work with the community to manage shared ontologies and linked data representations of library resources appropriately going forward (section 4).

Budget:

- 20% of Muhammed Javed
- 5% of Chew Chiat Naun
- 20% of Rebecca Younes

7. Metadata Conversion. In close collaboration with Stanford, develop a new, robust, efficient, well-documented, well-tested, open-source MARC to BIBFRAME converter to support the revised BIBFRAME ontology. Explore the development of a BIBFRAME to MARC converter to support the integration of BIBFRAME catalog records into existing Integrated Library System (ILS) workflows (sections 5.3 and 5.4).

Budget:

- 5% of Chew Chiat Naun
- 50% of Rebecca Younes

6.3 Harvard Statement of Work

Harvard Library is a participant in the research grant *Linked Data for Libraries: LD4L Labs* that will build on and extend the initial work of the LD4L project. Harvard will undertake the following activities:

1. **Deploy a pilot linked data conversion infrastructure.** The linked data infrastructure piloted by Harvard is the foundation for Harvard LD4L Labs tool development and testing. The infrastructure will support the conversion, ingest, hosting, and update of Harvard linked data created under the grant, and the establishment of a Linked Open Data endpoint to make that data accessible as linked data on the Web. It will most likely leverage the open source, Harvard Catalyst funded eagle-i platform[1], an ontology independent platform for hosting, editing, and searching linked data resources with any ontology, however Harvard will also assess the Cornell Vitro platform. Building on the work Cornell, Harvard, and Stanford completed during the LD4L project, Harvard will deploy a triplestore that scales to handle Harvard's BIBFRAME RDF, approximately 1 billion triples. The pilot infrastructure will enable the linked data to be updated from both legacy records in the Harvard ILS as well as from Harvard Geospatial Library and Harvard Film Archive linked data created as part of this grant. To link existing MARC metadata to the triplestore, Harvard will integrate and deploy the legacy record converters created under LD4L and LD4L Labs by Cornell and Harvard with Harvard's Library Cloud metadata pipeline, enabling Harvard to automatically and regularly update the triplestore with fresh data from the ILS. This work will include deploying revisions of the Cornell and Stanford developed MARC->BIBFRAME converter as a conversion step in Harvard's Library Cloud metadata processing pipeline.

Budget:

- 25% of Senior Software Engineer (see the Appendix for job description)
- 5% of Michael Vandermillen
- AWS web services to run the infrastructure as a pilot for 2 years

2. **Pilot a hosting environment for BIBFRAME linked data.** Assuming a good assessment from LD4L Phase 1 work, Harvard plans to pilot eagle-i within its infrastructure as a platform to provide catalogers with linked data creation, editing, display, and dissemination. If the LD4L Phase one eagle-i assessment is negative, Harvard would plan to deploy the Vitro platform. The environment 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. Both eagle-i and Vitro linked data platforms provide an ontology driven RDF creation and access environment. Since they are configurable based on any ontology, Harvard will configure one with the BIBFRAME ontology, with extensions for geospatial and Harvard Film Archive requirements. Harvard catalogers will evaluate the suitability of the chosen platform as an easily extensible production platform, and will collaborate with Cornell and Stanford in comparing and contrasting this environment with environments that they may deploy, as well as other BIBFRAME creation and editing environments. Development will include revisions to the native metadata editing functionality and user interface to reflect cataloger feedback and implement efficiency improvements.

Budget:

- 20% of Senior Software Engineer
- 5% of Metadata Technologies Program Manager (see the Appendix for job description)
- 10% of Marc McGee
- 10% of Christine Eslao

3. **Pilot linked data conversion, publication, and visualization of Harvard Geospatial Library metadata.** Working with the other project partners, Harvard will develop a BIBFRAME/LD4L profile; develop metadata conversion software to convert existing geospatial metadata records from the Harvard Geospatial Library and from Stanford (see the Stanford SOW in section 6.5) describing raster maps and vector map data layers into BIBFRAME; publish the RDF to the Harvard linked data endpoint; and integrate a beta of graph visualization software into the Harvard Geospatial Library or an Omeka virtual collection to assess end user value. This project will focus on converting a subset of OpenGeoMetadata metadata records from the Harvard Geospatial Library and Stanford (where they are now represented using the geospatial community standard Federal Geographic Data Committee (FGDC) schema, ISO 19139) 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 FGDC geospatial metadata standards 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.

Budget:

- 25% of Senior Software Engineer
- 5% of Metadata Technologies Program Manager
- 15% of Marc McGee

4. **Pilot linked data conversion, publication, and visualization of Harvard's Harvard Film Archive metadata.** The project will explore best practices for creating linked data descriptions for moving image resources including 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. The project will evaluate BIBFRAME/LD4L's effectiveness as a data model for describing moving image materials for research needs and the lifecycle of moving image materials, and identify vocabularies for description of these materials in a linked data environment. The project will create mappings for records from the HFA's film print database, focusing on a subset of moving image materials by women directors. 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. The project deliverables will include: a BIBFRAME/LD4L profile for moving image resources; a set of published descriptions for moving image materials and related archival collections; deployment of descriptions as linked data in the triplestore; a user interface and visualization for film researchers based on an Omeka or Harvard Geospatial Library on-line collection; and a written evaluation of the project and set of recommendations for future research and development.

Budget:

- 25% Senior Software Engineer
- 5% of Metadata Technologies Program Manager
- 15% of Christine Eslao over two years

5. **Collaborate with Cornell and Stanford on LD4L Labs and LD4P projects**. Participate in biweekly phone meetings, semi-annual LD4L Labs face to face meetings, and discussions of project related issues as they arise.

Budget

- 5% of Senior Software Engineer
- 5% of Marc McGee
- 5% of Christine Eslao

6.4 Iowa Statement of Work

The University of Iowa team will build upon previous work in disambiguation, data integration and visualization to provide both end-user interfaces and infrastructure integration mechanisms. A number of the existing capabilities in CTSAsearch and Shakeosphere serve as proofs of concepts for the planned LD4L Labs work, but virtually all require adaptation or extensive enhancement to function in the much broader LD4L semantic domain. Please see the expanded descriptions below for details on each of the following tasks:

1. Develop a disambiguation framework and pilot service for LOD "sameAs" assertions between LOD sites.
2. Pilot integration services for external data resources.
3. Develop and pilot a set of reusable visualization building blocks that libraries can incorporate into their LOD interfaces.

Expanding on each area:

1. Develop a disambiguation framework and pilot service for LOD sameAs assertions between LOD sites. A key benefit to globally accessible LOD is the unambiguous reference to a resource through a persistent URI. Iowa's CTSAsearch 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. The team will first extend CTSAsearch's limited framework of disambiguation of a single specific class (person) through the use of a single specific resource class (authorship) into an extended mechanism which explicitly models the disambiguation heuristics, allows for multiple resources, and that can be tuned for a particular context. Iowa will then 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. A central focus on this work will be the exploration of the architectural trade-offs between the centralized approach used by CTSAsearch (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).

Budget:

- 3% of David Eichmann
- 20% of graduate research assistant

2. Pilot integration services for external data resources. Iowa's work to date on Shakeosphere and CTSAsearch has drawn heavily on external data resources to extend core information resources and expand the utility of related user interfaces. In the case of CTSAsearch, VIVO-based research profiling data is expanded with externally curated resources (e.g., citation data and grant award data), filling in the gaps where source data are only partial (e.g., a missing abstract for a publication). In the case of Shakeosphere, the collaboration with the Map of Early Modern London (MoEML) project at the University of Victoria[2] has jointly benefited both projects, with gazetteer data from MoEML providing vocabulary to Shakeosphere for extraction of place names from publisher entries in catalog data and relative spatial references from Shakeosphere aiding MoEML in extending map annotations. The Shakeophere-MoEML cross-linkage also serves as an example of the potential of LOD integration. Since each external resource is distinct in value, structure and point of attachment, each requires custom development and integration into the larger architecture. Iowa will transform a number of external data resources (e.g., GRID) into LOD and pilot a resource similar to that for the sameAs assertions. Guidelines for community development of such resource integration services will also be created.

Budget:

- 3% of David Eichmann
- 10% of graduate research assistant

3. Develop and pilot a set of reusable visualization building blocks. Graphical presentation of richly-structured data has been a key component of the success of both CTSAsearch and Shakeosphere. Much of the visualization work to date in CTSAsearch has been in the aggregation of discrete relationship links between two entities (e.g., two researchers' history of coauthorship) into formation of constructs corresponding to a group of collaborators or a research community. The true power of a modular, architecture-driven approach to such direct-manipulation interfaces lies in the fact that much of the user interface of Shakeosphere, a project focusing on publication between 1540 and 1800, uses precisely the same visualization building blocks. Histograms, scatter plots, force-directed graphs, etc. form a *grammar of graphics*[3] that can use used to create significantly intuitional user interfaces, require little training for effective use, and substantially enhance viewer comprehension. Iowa will explore the applicability of their current D3[4] approach to LD4L data, and evaluate alternative technologies - looking to establish clean, agile support for visualization that is approachable by the library community. Iowa's current visualization architecture is comprised of two layers – the (browser-side) end user visualization layer in D3 JavaScript and a (server-side) connector layer in JSP/JDBC/SQL that feeds data from an underlying database to the D3 script in the user's browser. Iowa's existing library of D3 scripts will serve as an initial visualization capability, but will require substantial extension to support the temporal and representational requirements of LD4L LOD. The entire connector layer will also need to be recoded to use SPARQL for querying and to provide compatible modules for Blacklight /Spotlight, Omeka, and other platforms as need arises over the course of the project.

Budget:

- 5% of David Eichmann
- 20% of graduate research assistant

6.5 Stanford Statement of Work

Stanford University Libraries will be an integral partner in the *Linked Data for Libraries: LD4L Labs* project, continuing on the work that it has contributed in the LD4L project, and its ongoing and productive collaboration with Cornell and Harvard. It will undertake this both as a funded contributor to the LD4L Labs project, and also as the lead institution on the LD4P Program grant, which seeks to coordinate diverse but inter-related linked data efforts among LD4P partners into a coherent program. Specific Stanford activities as part of the LD4L Labs project include:

1. Collaboration and Coordination on Annotations, Linked Data Tools, and Discovery Analysis
2. Ontology work, especially ongoing refinement of the BIBFRAME / LD4L ontology
3. Metadata converter development
4. Geospatial metadata conversion

1. Collaboration and Coordination on Annotations, Linked Data Tools and Discovery Analysis. Stanford will maintain an ongoing programmatic effort in leveraging linked data and building an environment to support its digital library environment, as well as for the specific deliverables described in the LD4P proposal. Through its involvement in the LD4L Labs team, it will be an integral partner in coordinating these efforts with those described in this, the LD4L Labs, proposal. Stanford specifically will be invested and conducting ongoing development in adding annotation support to its production systems, especially SearchWorks, the Stanford library catalog. This was first prototyped and proven in concept in the LD4L grant, and is described further in section 2.1 of this proposal.

Stanford is also doing ongoing exploration of how to best apply LDP, the Linked Data Platform specification, to its digital library, especially via the Fedora and Hydra repository platforms. As its digital repository architecture shifts to linked data over the course of the next two years, it will provide another reference point and operational environment to compare and contrast, and complement, the activities described in 2.4. This cross-pollination of similar technical environments, and ongoing in depth conversations among linked data technologists among the LD4L partners, will continue to be a key part and advantage of the LD4L Labs structure.

Budget:

- 10% of the LD4L Labs Technologist

2. Ontology work described in section 4, particularly the revision of the BIBFRAME/LD4L ontology (4.1). Stanford will also be an active architect and implementer of strategies for reconciliation and persistence (4.2). The LD4L Labs team will provide a forum and project structure for exploring high level approaches to these issues; Stanford's work on in the LD4P Program, and specifically its Tracer Bullet projects (described in the LD4P proposal) will provide a proving ground with test implementations.

Budget:

- 5% of the LD4L Labs Technologist

3. Metadata converter development. Stanford's LD4L Labs technologist will be a lead developer on a new MARC->BIBFRAME converter, described in 5.3 This utility will not only apply the revised BIBFRAME ontology (and through adaptation, support any extensions further described in the LD4L Ontology) but also provide a foundation for a reusable converter framework that can be applied to other metadata transformations by establishing a generic pipeline. Extensions may include BIBFRAME->MARC (described in 5.4), as well as converting MODS to BIBFRAME.

Budget:

- 20% of the LD4L Labs Technologist

4. Geospatial metadata conversion. Stanford Libraries has a robust and growing program around geospatial resources, including amalgamation of geospatial records via OpenGeoMetadata, ongoing exploration of leveraging geospatial authority data and gazetteers, and leading development of GeoBlacklight, a Blacklight plugin that optimizes the popular open source discovery application for GIS data and maps. It will contribute to the LD4L Labs activities described in 5.1 through contribution of records, consultation with other GIS and linked data experts in the LD4L Labs project, and analysis of the feasibility and advantages of expressing Linked Data descriptions of geospatial resources in the GeoBlacklight discovery platform. The LD4L Labs Technologist will serve as the key liaison between the LD4L Labs project team members (especially those at Harvard) with Stanford's geospatial development team.

Budget:

- 5% of the LD4L Labs Technologist

---

[1] https://www.eagle-i.net/about/
[2] http://mapoflondon.uvic.ca
[3] https://www.cs.uic.edu/~wilkinson/TheGrammarOfGraphics/GOG.html
[4] http://d3js.org