Project Description

Go to LD4L Wiki Gateway

Archived

LD4L 2014, which was the Linked Data for Libraries original grant running from 2014-2016, has been completed. This page is part of the archive for that grant.

Overview

In creating the SRSIS platform and ontology, our focus is on discovery and access – helping users identify scholarly information resources that will meet their needs and getting the users to those resources. The goal is to create a SRSIS model that works both within individual institutions and through a coordinated, extensible network of Linked Open Data to 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.

The project to create a SRSIS platform and ontology shares some of its goals with the Stanford Linked Data Workshop Technology Plan[1], a document that stimulated our thinking by articulating a vision whereby the semantic models developed at Cornell for VIVO could fruitfully be applied to the management and discovery of "the intellectual activity embodied in all of a research university's academic endeavors and in all the academy's use [of] resources and programs of its research libraries." (p. 9). The report speaks of "useful enhancements to the descriptive aspects of metadata for resource description and management," including "records decomposed into statements of fact with strong identifiers" and "reconciliation of connections among such facts." In this context, a strong identifier is an identifier defined and maintained as unique (i.e., not recycled) by an appropriate authority (e.g., an employer) expressed as a uniform resource identifier (URI)[2] resolvable as a uniform resource locator (URL)[3] on the Internet. Reconciliation of connections among such facts occurs in Linked Data through accessing RDF statements referencing the same identifier(s), often from more than one source located anywhere on the Internet. Statements in RDF gathered from multiple sources have identical format and may be aligned on the basis of common subject, predicates, or objects – and statements may of course be found to be inconsistent. Ontologies referenced in RDF statements can be mapped to each other through assertions that two classes (types) or properties (relationships) are equivalent, and two different URIs can be declared to refer to the same person, organization, or other entity via a *sameAs* relationship[4].

The report suggests new types of links including scholarly annotations, citation maps, pointers into content, activities of service programs, paths that bridge vocabularies, and links crossing institutional boundaries (p.9). The report continues by describing data about Mark Twain displayed in a web browser configured to request, consume and display RDF data, called LinkSailor[5], and adds that "When the density of the graph's fabric and scope of growing coverage from academic institutions comes into play, the capabilities of discovery, navigation, and access tools that are the children and grandchildren of LinkSailor and its siblings will need to provide all manner of personalization alternatives, *e.g.,* capabilities allowing one to filter and select for relevant resources and information from a wealth of alternatives."

By creating the SRSIS platform and ontology across the three partner institutions, we will be able to assemble large bodies of linked data that describe library resources of many diverse types. Each institution will contribute a unique perspective on how to describe these resources, the types of context and relationships that can be used to organize and characterize them, and the scholarly use cases for discovery and access. By melding these perspectives together into a common ontology, we will be able to relate and share context and resources across our own institutions and with new external partners. Our VIVO experience across multiple institutions leads us to believe this approach will provide benefits even as techniques continue to mature, and that driving ontology development from the available data and the uses people can make of that data will result in better and more generalizable ways to describe, discover, and access scholarly information resources.

We do propose in this effort to go beyond the “Phase One” proposal in the Stanford Linked Data Workshop Technology Plan, which proposes implementing the British Library Data Model[6], an RDF schema[7] developed by the British Library from 13 ontologies and vocabularies for publishing Linked Data about 2.8 million book and serial records. The proposed Phase One focuses on library metadata and other bibliographic/citation data as the source to create an institution-wide set of linked data statements. We will seek in our initial efforts to characterize a distinctly broader set of contexts for library information resources, drawing on the previous work at Harvard, Cornell, and Stanford, as well as many others.

The project will explore mapping to the evolving Bibliographic Framework (BIBFRAME)[8] effort that Zepheira is leading for the Library of Congress[9]. BIBFRAME is intended to serve as a linked-data-based successor to the MARC 21 standard, providing structured descriptions of four fundamental resources: works, instances, authorities, and annotations, and the relationships among these resources (see Appendix B for a fuller description of BIBFRAME). The SRSIS ontology will go beyond BIBFRAME by representing a much broader set of contexts for information resources than the MARC data that is the focus of BIBFRAME. Such contexts could include usage information across multiple libraries, the fact that the book was referenced in a librarian-authored Library Guide[10], or the fact that an article had been on the reserve list for a specific course. In particular, we can leverage the annotation frameworks of BIBFRAME and the Open Annotation Specification[11] to integrate the rich SRSIS data and its web of resource relationships with the basic resource description in BIBFRAME.

The work of the Linked Data for Libraries project falls into two separate but strongly interrelated parts. The first part is the creation of the SRSIS ontology, both by building on and integrating existing ontologies and by creating new ontology elements that capture broader contexts about how scholarly information resources are used, described, referenced, and organized. The second part is the implementation of a platform that ingests data about these resources from multiple sources, including MARC-based catalogs, anonymized circulation records, finding aids, and Library Guides; translates that data into RDF triples in the SRSIS ontology; stores those triples in a triplestore; and then serves them up on the web in both human and machine readable forms.

Much of the early effort on the project will focus on the first part of the work, the creation of a set of ontology modules that capture the full range of contextual information about resources embodied in a wide range of sources: traditional catalogs, finding aids, StackLife, VIVO, LibraryCloud, the Digital Medieval Manuscript project, Cornell's recently-developed Curated List of Library Resources (CuLLR), and many others. This ontology will primarily describe types of information resources and their relationships to each other and to people, organizations, places, dates, events, and other entities rather than encompassing all possible subject terms for classifying these entities. The goal will be to use this ontology to capture "all that we know" about information resources and to link those resources to independently maintained subject headings and internationally-managed unique identifier systems for authors[12], institutions[13], journals[14], events, and geographic entities. The team working collaboratively on the ontology will use Protégé, a Stanford-developed ontology editor, to create the SRSIS ontology modules and test their logical consistency for support of simple reasoning.

The second part of the work of the project, the creation of a SRSIS platform, requires different capabilities than those provided by Protégé, which focuses on manipulating ontologies, rather than actual metadata and data about the scholarly information resources. To provide the primary infrastructure supporting management of this data and metadata in the SRSIS triplestore, the SRSIS developers will use the VIVO software in its more basic form, stripped of customizations specific to the VIVO application and its ontology, called simply Vitro[15]. Vitro provides an ontology importer/editor/annotation tool that can read the SRSIS ontology modules developed in Protégé and derive from them all the types and relationships for Vitro's fully-featured, web-based content editor, storage, and display system. Vitro uses an Apache Jena[16] triplestore as a data persistence layer by default but is configurable to work with any triplestore supporting create, read, update, and delete operations through a standard query interface called SPARQL (SPARQL Protocol and RDF Query Language)[17].

A Linked Data approach moves away from focusing on the freestanding bibliographic record to a structure supporting relationships among the people, organizations, places, and events that may be the creators, players in, or subjects of the works represented. Using Linked Data will not only allow more nuanced treatment of units of information as "independent entities" but will expand the range of relationships among these entities and provide an improved structure for managing disambiguation. By "independent entities" we simply mean treating any identifiable entity as something that can have a name and attributes in its own right rather than simply referring to it as a string value or describing it in narrative text. Documentation of topically relevant information resources uncovered via a reference consultation, for example, may involve a contributor, topic, and reference to a publication. By representing all of these referenced entities as semantic resources with their own unique identifiers – even if "authoritative" identifiers may not yet exist – we will improve our ability not only to build well-structured local data in the short term but also to link to authoritative data sources as they are identified or become available.

Locally managed identifiers in the form of URIs serve useful purposes and aid in cross-linking information known or asserted locally to refer to the same entity. Since the Linked Data world is not a closed platform or "walled garden,"[18] any later authoritative external URI discovered to be referencing the same entity may be associated with the local identifier by asserting a *sameAs* relationship[19] between the local URI and the authoritative external URI. Should the assertion of sameness prove incorrect, the entire body of local assertions can be disassociated from the external identifier simply by removing the *sameAs* statement.

We will recommend linking local entities to external vocabularies as data are first loaded into SRSIS, and we will build interfaces to support authoritative external tagging of information resources as well as organizations, people, events, and places. Increasingly, taxonomies and thesauri are being made available as Linked Data to facilitate direct referencing by URI in other Linked Data – notably Library of Congress Subject Headings[20] and the Getty vocabularies[21]. Referencing external authoritative URIs directly or via *sameAs* assertions avoids duplicating effort by re-inventing domain specific controlled vocabularies within our ontologies. Any common external references embedded in SRSIS as the repositories are loaded will simplify reconciliation of data from multiple sources.

The links to contextual information, both local and external, that are made possible by our proposed ontology are a key part of the challenge of disambiguating authors and other entities. We will address disambiguation by using both human and automated means (moving inevitably to the latter) to analyze the full contextual information before producing new assertions that something is the same as something else. For authors without unique identifiers, for example, we will accumulate name parts, email addresses, affiliations, subject expertise, and lists of co-authors; for other types of entities we will have to rely more on common linkages and string comparisons, as well as online services such as VIAF[22] and ISNI[23]. Because these techniques will never be error free, initiatives such as ORCID, which show promise for capturing identifiers as data are created and will link people to organization identifiers[24], offer the best long-term solution.

In the absence of authoritative external sources, SRSIS records in the namespace of each partner institution will provide stable, accessible, context-rich URIs documenting what is known in the local setting. Our ability to link resources across the Cornell, Harvard, and Stanford SRSIS resources through curation, automated suggestion, and patron involvement will help suggest best practices and open research questions going forward, and we will follow advances in the wider Semantic Web and identity management communities closely.

In mapping the domain of research activities in VIVO, we have not just modeled outputs but also the activities from which they resulted, such as teaching, bench research investigations, analysis, workshops, community service and professional service. The VIVO work has been informed by earlier efforts including the AKT project[25] in the United Kingdom and incorporates elements of 12 other ontologies; new classes (types) and properties (relationships) have been created where no suitable pre-existing alternatives were found. In addition, we are working on modeling the tools and instruments that are part of the creation of those outputs, for example cell lines, spacecraft, laboratory instruments, and clinical trials through collaboration with ontologists and bioinformaticists on the CTSAconnect project, an NIH-funded grant to unify the VIVO and eagle-i[26] ontologies and add clinical expertise[27]. We have also sought to discriminate among the roles people perform while participating in these activities in order to distinguish leadership from attendance, gather evidence on subject expertise, and document creative contribution. As datasets gain credence as outputs of research, and as funding agencies such as the NSF are now encouraging researchers to list all types of research resources on biosketch forms in proposals, we believe such semantic data will be necessary to show how datasets and other resources derive from one another and relate to traditional scholarly publications. This richer context for research outputs will also put us in a better position to understand the trail of scholarship.

While VIVO describes scholars and scholarship in all disciplinary areas (for example, the Cornell Classics Department web site uses VIVO data to create their faculty profile pages[28]), the primary focus of the NIH funding was on developing additional research and researcher context in biomedicine and related areas. With the growing adoption of VIVO by many different institutions, Cornell is now working with the VIVO implementation teams at Brown University and Duke University to extend the VIVO ontology to support better measures for accomplishment across a wide range of disciplines, especially in the humanities and creative arts[29]. While accomplishment in the sciences is typically measured through articles in peer-reviewed journals, accomplishment in the humanities is reflected by very different measures, including the creation of monographs, book chapters, and reviews, as well as evidence of ongoing or derivative scholarly value. For example, the performance of a work by a prestigious orchestra; the translation of a monograph into many different languages; or the inclusion of a poem in an anthology may all reflect high levels of accomplishment. This type of context is not necessarily captured by typical cataloging and metadata, and yet it may be a very useful addition to information resources in the SRSIS. We will draw on both our VIVO-related efforts and the collaborations within this grant to improve the context for information resources in the humanities and creative arts.

One advantage for the project in working with the basic Vitro framework is that we can leverage tools that are built for other projects on top of that framework. In particular, we propose to use the VIVO Search software developed at Cornell to harvest VIVO-ontology-compatible RDF from multiple VIVO sites to create a multi-institutional VIVO search[30]. To use this software for the Linked Data for Libraries project, we will need to specialize the faceting and user interface to support the SRSIS ontology, but the basic software framework to harvest SRSIS-ontology-compatible RDF, build a Solr index for search, and display the combined search results already exists. As part of the Linked Data for Libraries project, we will build a demonstration SRSIS Search application that searches the combined SRSIS RDF from all three partner institutions.

Finally, we will work to make the SRSIS more broadly usable beyond the basic Vitro, triple-store/SPARQL, and Linked Open Data environments by integrating SRSIS with the Hydra project's[31] Repository/Discovery framework. The Hydra project, an extension of the Blacklight effort, has created an impressive framework and flexible programming environment that uses Fedora for storage and Solr for indexing and search. This integration will allow Ruby-on-Rails developers the same ease of use in working with SRSIS data as they currently get working with data and metadata in Fedora. The integration will also allow us to easily provide the Blacklight discovery interface on top of SRSIS data.

A key technology component that has led to the success of the Hydra project is ActiveFedora, which implements the Ruby-on-Rails Active Model[32] interface pattern on top of Fedora objects instead of an SQL database. The ActiveFedora component enables programmers to work with Fedora objects as easily as they would with a native Ruby object. We propose creating an ActiveTriples component that will provide the same simplicity of use when working with linked data in the SRSIS triple-store, or in other triple-stores.

There has been work to create "object triple mapper" (OTM) software that provides interfaces similar to those that object-relation mappers do for relational databases, and several efforts are described in a recent review article[33]. Of a number of existing projects[34], one of the most advanced is ActiveRDF [35], written in Ruby. However, ActiveRDF has not followed development of Ruby-on-Rails and, in particular, does not implement the Active Model API. We will build on this body of work to create ActiveTriples. By implementing the Active Model API we will make the creation and reuse of linked data accessible to Ruby-on-Rails developers using a design pattern and interface with which they are already familiar.

In addition to providing tools for integration of Linked Open Data into objects within Hydra-based components, we recognize the value of providing simpler tools enabling web content managers to query a triple store directly and display results within common off-the-shelf content management systems including Drupal[36], WordPress[37], and Joomla[38]. Drupal has native support for mapping its internal data representation to and from RDF, and we propose to leverage the JSON-LD[39] tools and/or existing implementations of an open-source linked data API[40] in Java[41] and PHP[42] to make the ontology structure and the content of the SRSIS accessible to a much wider audience of developers and libraries worldwide.

Stanford Contributions

To contribute to the overall effort, Stanford University Libraries (SUL) would both stand up a Stanford-specific SRSIS, and provide practical input on how to refine it via an iterative process of “plan-do-assess” by integrating SRSIS-powered services into user-facing parts of Stanford’s information ecosystem (described below), and then fine tuning the SRSIS data models and mechanics based on evaluation of their utility in these environments.

The first package of work will leverage the SRSIS models and tooling to transform metadata for "everything scholarly" and available at Stanford—bibliographic records, person records, profiles data, authority files, local authority files, etc.—into linked data and loading these into a Stanford SRSIS, linked to the Cornell and Harvard stores directly and indirectly (e.g., through such standard identifiers as ORCID, VIAF, and LCSH). Stanford will also pilot the use of a variety of algorithms to transform traditional metadata to Linked Data, such as those from Library of Congress[43], OCLC[44] and those based on the Europeana Data Model[45]), and triplestores other than the Jena triplestore used by default in Vitro (such as Sesame[46] and OWLIM-SE[47]) as the heart of its SRSIS. Along with the other project partners, Stanford will use Protégé[48], a Stanford-originated ontology editor stemming from bioinformatics, as the primary tool to create the SRSIS modular ontology. Vitro’s own internal ontology editor will then be used for minor adjustments and annotations to the resulting ontology modules in the SRSIS; both Protégé and Vitro use the Web Ontology Language (OWL)[49] for two-way interchange of ontologies between them and with other semantic web tools.

The second package of work will focus on leveraging the linked data in SRSIS, and determining if the information created is useful and usable. With this in mind, Stanford would explore enriching specific, existing services with functions or links powered by SRSIS. These might include:

- SearchWorks, Stanford’s Blacklight-powered next generation catalog
- SUL's citation database—an amalgamation of publications information that feeds CAP (Community Academic Profiles) Network[50], Stanford’s profiles application (i.e., Stanford’s “VIVO”, though not powered by VIVO)
- The Stanford Digital Repository (SDR), including persistent views upon published digital objects and collections, as well as linked-data-based input forms for faculty depositors
- SDR-based description tools and services for both crowd-sourced and curated metadata

As part of this work package, Stanford will also test and provide input on the semantically-enriched Hydra or Blacklight-based code produced by the Cornell team, to assess the code's broader utility for these open source communities.

Finally, Stanford will actively participate in exploring the use of linked data and community-based ontologies to represent current metadata schemas (MARC, MODS, DC, EAD) for library and archival assets; this is an area of tremendous promise, as we search for more flexible and robust ways of cross-walking descriptive metadata for scholarly objects that might be coming from non-traditional sources (e.g., looking at "tags", geocoordinate data, or annotations that would typically have to be mapped into MARC or MODS). The ontology work that Cornell will lead is also of particular interest to Stanford as SUL explores modeling digital objects and their component parts in Fedora using a primarily RDF-based data model. SRSIS, and cooperative development with Cornell and Harvard in general, holds the promise of "atomizing" and integrating disparate sources of metadata (regardless of schema), with heterogeneous digital objects and collections (potentially stored across disparate repositories), and then producing standard forms of output (e.g., MODS or Dublin Core for metadata, "books" or geolocated maps for data) on demand.

Harvard Contributions

The Harvard Library Innovation Lab (LiL) will provide two packages of work:

First, LiL will engage with Harvard librarians and metadata experts to discover the properties and relations among some of Harvard's unique data sources—such as topical collections, international law works, and historic archives—within Harvard's community of users and uses. Because of LiL's current experience with various types of usage data (or, more broadly, event data, such as a work being checked in, put on reserve, called back early from a loan, etc.), LiL will focus on how that data is currently gathered and represented, and how it might usefully be exposed. LiL will provide this information to the project as feedback to be incorporated in the design of the SRSIS ontologies and tools.

Second, LiL will stand up an instance of SRSIS as a prototype fork of LiL's metadata server, LibraryCloud. LiL expects doing so will involve moving LibraryCloud's 12.3M bibliographic and holdings records into a triple store, and integrating it with RDF-based event data, and quite likely with VIVO-based faculty data (Harvard Faculty Finder).

---

[1] http://www.clir.org/pubs/reports/pub152/LDWTechDraft_ver1.0final_111230.pdf
[2] http://tools.ietf.org/html/rfc3986
[3] http://tools.ietf.org/html/rfc3986#page-7
[4] http://www.w3.org/TR/owl-ref/#sameAs-def
[5] http://linksailor.com/nav
[6] http://www.bl.uk/bibliographic/datafree.html
[7] http://www.bl.uk/schemas/
[8] http://bibframe.org/
[9] http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf
[10] http://guides.library.cornell.edu/
[11] http://www.openannotation.org/spec/core/
[12] http://orcid.org
[13] http://viaf.org
[14] http://agris.fao.org has a disambiguated list of 40,000 journals with non-proprietary, stable URIs
[15] https://github.com/vivo-project/Vitro
[16] http://jena.apache.org
[17] http://www.w3.org/TR/rdf-sparql-query/
[18] http://en.wikipedia.org/wiki/Closed_platform
[19] http://www.w3.org/TR/owl-ref/#sameAs-def
[20] http://id.loc.gov
[21] http://www.getty.edu/research/tools/vocabularies/lod/index.html
[22] http://viaf.org
[23] http://www.isni.org
[24] http://orcid.org/blog/2013/06/27/orcid-plans-launch-affiliation-module-using-isni-and-ringgold-organization
[25] http://www.aktors.org/akt/
[26] https://www.eagle-i.net
[27] http://ctsaconnect.org
[28] http://classics.cornell.edu/people/faculty.cfm
[29] "Modeling Humanities Scholarship in VIVO", presented at the 2013 VIVO Conference by Steven McCauley and Ted Lawless, Brown University Library; http://www.vivoweb.org/files/vivo2013/friday_pm/VIVO-Humanities_McCauley.pdf
[30] http://vivosearch.org
[31] Project Hydra, http://projecthydra.org/
[32] ActiveModel API, http://rubydoc.info/gems/activemodel/frames
[33] Supporting Object-Oriented Programming on Semantic-Web Software, Matthias Quasthoff and Christoph Meinel, IEEE Trans. Systems, Man. and Cybernetics Part C, 42 (1) 2012.
[34] Other object-triple mapper projects:

http://code.google.com/p/object-triple/ (Java, read-only), https://github.com/mhgrove/Empire#readme (Java, read-only), http://packages.python.org/SuRF/ (python, like ruby ActiveRDF)

[35] ActiveRDF: Object-Oriented Semantic Web Programming, Eyal Oren et al., WWW2007. http://www2007.org/htmlpapers/paper272/and ActiveRDF current documentation at http://activerdf.org/ , code on github https://github.com/ActiveRDF/ActiveRDF (little activity in past 3 years)
[36] http://drupal.org/
[37] http://wordpress.org/
[38] http://www.joomla.org/
[39] http://json-ld.org
[40] http://code.google.com/p/linked-data-api/
[41] http://code.google.com/p/elda/
[42] http://code.google.com/p/puelia-php/
[43] https://github.com/lcnetdev/marc2bibframe
[44] http://www.oclc.org/en-US/home.html
[45] http://pro.europeana.eu/edm-documentation
[46] http://sourceforge.net/projects/sesame/
[47] http://owlim.ontotext.com/display/OWLIMv43/OWLIM-SE
[48] http://protege.stanford.edu
[49] http://www.w3.org/TR/owl-features/
[50] http://cap.stanford.edu/