Previous Partner LD Work

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.

Cornell University

The Cornell University Library began exploring Semantic Web ideas, tools, and data in 2003 through Metadata Working Group forums [1] addressing ontologies, RDF, the National Science Digital Library (NSDL) [2] , and the Resource Description and Access (RDA) guidelines [3] . The NSF-funded National Science Digital Library project in the Department of Computer Science at Cornell provided technical leadership, including system design and software development, in defining metadata for educational resources and developing tools for large-scale metadata harvesting and integration. Dean Krafft co-authored an influential paper [4] calling for a resource-centric overlay network capable of " represent[ing] the diversity of relationships among the content, agents, services, standards, and the other entities in the NSDL context. The relationships must be expressed in an extensible ontology, so that other systems are able to use and reuse the relationship structure." Krafft has also served as Principal Investigator on the NSDL Technical Network Services project for two intervals since 2005.

Another member of the project team, Jon Corson-Rikert, developed the first version of the VIVO (not an acronym) software for the Life Sciences Working Group [5] in the Cornell Library in 2003. VIVO was created to address needs for improved cross-disciplinary interchange and awareness expressed by faculty members serving on a Cornell Genomics Initiative [6] task force on student and faculty recruitment. VIVO [7] created a network of Cornell life science researchers across departments and campuses presenting a single search and browse interface, and it emulated the ontology structure and relationships of the AKT (Advanced Knowledge Technologies) Project [8] in the United Kingdom while initially still using a relational database back end.

VIVO's novel structure and presentation led to expansion beyond the life sciences to all major disciplines at the University in 2007-2008, and VIVO was converted to full compatibility with the OWL (Web Ontology Language) ontology [9] and RDF data Semantic Web standards that same year. In 2009, Cornell applied to the National Institutes of Health (NIH) with 6 other universities and medical schools and was awarded a two-year, $12.2M project to expand the scope of VIVO from a tool used independently at a handful of universities into an open-source platform capable of supporting a coherent national discovery network of researchers and research activities [10] . During the *VIVO: Enabling National Networking of Researchers* project, adoption extended the reach of the software internationally and to additional types of institutions including the United States Department of Agriculture (USDA), the U.S. Environmental Protection Agency (EPA), the American Psychological Association (APA), the Food and Agriculture Organization of the United Nations (FAO), and the Inter-American Institute for Cooperation on Agriculture (IICA). All VIVO sites in full production [11] provide Linked Open Data, and three additional software platforms [12] have adopted the VIVO ontology as a standard for exchange of research information. Dr. David Eichmann of the University of Iowa has created a search application encompassing VIVO data including over 118,000 researchers from 14 institutions. [13]

In 2012, the Cornell University Library began implementation of a new catalog discovery tool leveraging previous work at Stanford [14] , Columbia [15] , and other member institutions on the Hydra Project [16] . The Hydra Project is a partnership of a number of research libraries and related organizations to create a shared software framework to support creating, managing, and accessing digital repositories (see Appendix B for a fuller description of Hydra). The new Cornell Library search has now been launched in beta form [17] and leverages an intermediate RDF model converted from Machine Readable Cataloging (MARC) [18] records in the Voyager Online Public Access Catalog (OPAC) [19] for construction of an Apache Solr [20] search index supporting a Blacklight [21] discovery interface (see Appendix B for a description of Blacklight). Creation and refinement of an RDF version of Voyager bibliographic and holdings records opens the door to new work focusing on publishing information about Cornell Library holdings as Linked Data, as well as linking Cornell's holdings with other information not currently held in the Voyager OPAC.

Stanford University

The Stanford University Libraries have been pursuing Linked Data as a means to transcend the traditional silos of library-based information stores since the Stanford Linked Data Workshop [22] in June of 2011. With the support of the Mellon Foundation, Stanford University Libraries hosted a weeklong workshop of librarians and technologists from across the world with the objective of identifying the most promising areas of work to establish a global Linked Data network of scholarly information resources. The report from this workshop [23] details the value proposition of Linked Data, the overall lifecycle and methods for publishing and linking resources as Linked Data, and identifies several potential next steps and open issues in fostering the creation of this global network.

Since then, Stanford University has undertaken numerous efforts in line with the vision and roadmap articulated at the workshop. These efforts include multiple trials to transform traditional MARC bibliographic records to Linked Open Data, using a variety of algorithms. In the most notable experiment, over three million MARC catalog records were transformed to RDF and published in FreeBase, a Linked-Data-powered search engine (http://www.freebase.com/). Metaweb, the company that produced and maintained FreeBase, has since been acquired by Google, and incorporated into efforts to enhance Google's searching with semantic methods. Stanford University Libraries is currently exploring linking these MARC-based triples with Linked Data from other sources, including data on people, and metadata from published journal articles. The overall objective of these efforts is to transcend traditional library information silos (books in a catalog, articles in a database or abstract and indexing services, people in a directory), and enable seamless discovery of all relevant works (books, articles, and more) on any defined topics, by any authors.

Harvard University

Two years ago, the Library Innovation Lab prototyped exposing catalog metadata about Harvard Library's collection of 12.3M items as Linked Data. They did this by translating that data from its traditional non-Linked Data database (MySQL) into RDF (the Linked Data data format), using a tool called D2RQ. This data could then be retrieved as RDF through a standard query interface called SPARQL (SPARQL Protocol and RDF Query Language). In another project, the Lab set up an RDF data store (using the 4Store software), pulling in the full set of information about authors maintained by VIAF [24] (Virtual International Authority File) as well as the catalog records of the 1 million most widely held items in the WorldCat global catalog maintained by the OCLC consortium.

The Lab also engaged in a six-month project with Dan Brickley to try to gather Linked Data about particular high-value Web sites – initially the TED Talks – to see if Library of Congress Subject Headings could be automatically applied to those pages. The project entailed "cross-walking" multiple Linked Open Data datasets to try to generate enough text relevant to a Web page so that it could be statistically distinguished from the content of non-relevant LCSH's. The team working on the project was unable to gather enough data to produce reliable results, but the Library Innovation Lab was able to learn about and explore the Linked Open Data datasets available at the time. The Lab is also involved in the scaling up of its LibraryCloud metadata server for use by the entire Harvard Library system; the Harvard-wide version will include a Linked Data strategy that is currently under development, and it will integrate Linked Data applications under development by the Harvard Library.

---

[1] https://metadata-wg.mannlib.cornell.edu/forum/
[2] http://nsdl.org
[3] http://www.rda-jsc.org/rda.html
[4] An Information Network Overlay Architecture for the NSDL, http://arxiv.org/pdf/cs/0501080.pdf
[5] http://ecommons.cornell.edu/handle/1813/11679
[6] The Genomics Initiative began at Cornell in 1997 and was later subsumed in the Cornell New Life Sciences Initiative, described at http://bmcb.cornell.edu/cornell/index.html
[7] http://vivo.cornell.edu
[8] http://www.aktors.org/akt/
[9] http://www.w3.org/standards/techs/owl#w3c_all
[10] http://www.nih.gov/news/health/nov2009/ncrr-02.htm
[11] https://wiki.duraspace.org/display/VIVO/VIVO+Main+Page#VIVOMainPage-publicvivos
[12] Harvard Profiles (http://catalyst.harvard.edu/spotlights/profiles.html), SciVal Experts (http://info.scival.com/experts/vivo), and the University of Iowa's Loki (https://www.icts.uiowa.edu/Loki/)
[13] http://research.icts.uiowa.edu/polyglot/
[14] http://searchworks.stanford.edu
[15] http://clio.columbia.edu
[16] http://projecthydra.org
[17] http://search.library.cornell.edu
[18] http://www.loc.gov/marc/
[19] http://www.exlibrisgroup.com/category/Voyager
[20] http://lucene.apache.org/solr/
[21] http://projectblacklight.org
[22] http://www.clir.org/pubs/abstract/reports/pub152
[23] http://lib.stanford.edu/files/Stanford_Linked_Data_Workshop_Report_FINAL.pdf
[24] http://www.oclc.org/viaf.en.html