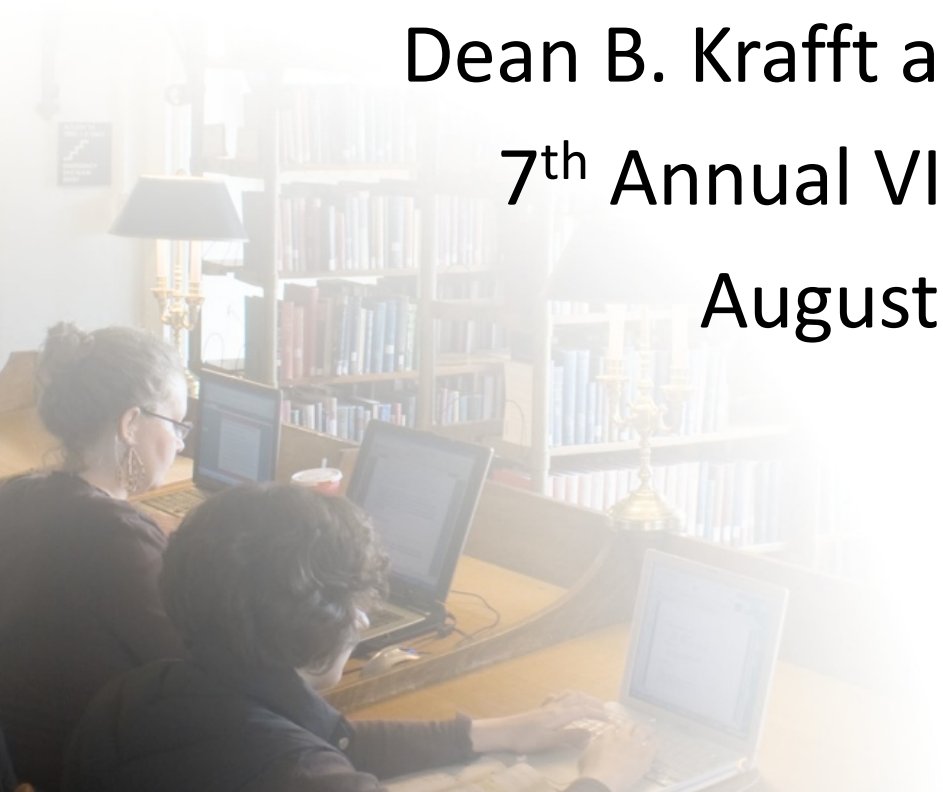
Linked Data for Libraries (LD4L): Expanding the Linked Data Ecosystem

Dean B. Krafft and the LD4L Team

7th Annual VIVO Conference

August 18, 2016



Cornell University Library

HARVARD LIBRARY

STANFORD UNIVERSITY LIBRARIES

Linked Data for Libraries (LD4L)

- Just completed (3/31/16) a two-year $999K Mellon grant to Cornell, Harvard, and Stanford
- Partners assembled ontologies and created linked data and tools to provide relationships, metadata, and broad context for Scholarly Information Resources
- Leveraged existing work by both the VIVO project and the Hydra Partnership

Vision: Enable libraries and their users to easily create, use, and benefit from LD specifically designed for libraries and scholars, and from broader sources of LD on the web.

Courtesy National Gallery of Art, Washington

LD4L Workshop Recommendations

- Our goal should be that others outside the library community use the linked data that we produce
- We must create applications that let people do things they couldn't do before – don't talk about linked data, talk about what we will be able to do
- Local original assertions (new vs. copy cataloging) should use local URIs even when global URIs exist
- Look to LD to bring together physically/organizationally dispersed but related collections
- Libraries must create a critical mass of shared linked data to ensure efficiency and benefit all of us

LD4L Research Outputs

- LD4L Ontology – modification/extensions to BIBFRAME: github.com/ld4l
- Recommendations to Library of Congress on BIBFRAME, leading to ongoing revisions
- Conversion of 23 million Cornell, Harvard, and Stanford catalog records to ~3 billion triples of resolvable LD available at draft.ld4l.org/downloads/index.html
- Demonstration search over LD at search.ld4l.org
- Contributions to Hydra: ActiveTriples gem; work on LD annotations (available at github.com/ld4l)
- Demonstration of links to external web URIs (e.g. MusicBrainz/LinkedBrainz)

LD4L Recommendations to LoC

- Create a domain model to ensure that BIBFRAME: defines appropriate terms from the domain model; defines only terms from the domain model; defines only one pattern for each feature; and considers dynamic resources carefully.
- Make proper use of Linked Data URIs: use URIs, not strings; URIs must identity one thing; use HTTP URIs; use Natural Keys in URIs; clients should treat URIs as opaque; and avoid dates and hash URIs
- Allow others to make use of BIBFRAME data: provide useful information when a URI is requested; describe your own resources - individually; include links to other resources; and avoid reification, lists, and blank nodes.
- Reuse existing work: reuse existing vocabularies; define terms in your own namespace and relate new terms to appropriate existing ones; name terms consistently; only define what matters; and inverse relationships matter.

https://www.cni.org/topics/assessment/the-future-of-linked-data-in-libraries-assessing-bibframe-against-best-practices

downloads

RDF Files

Converter output

The MARC records from each library were converted to BIBFRAME 1.0 RDF by the Library of Congress mar2bibframe converter. LD4L's bib2lod converter was then used to produce RDF in the LD4L data model. The result is RDF in the N-Triples format.

These dumps are available:

- cornell.ld4l.full-catalog.2016-03-17.tar.gz -- Converted triples for the Cornell catalog -- 13GB
- harvard.ld4l.full-catalog-1.2016-03-24.tar.gz -- Converted triples for the Harvard catalog (1 of 4) -- 5.4GB
- harvard.ld4l.full-catalog-2.2016-03-22.tar.gz -- Converted triples for the Harvard catalog (2 of 4) -- 5.2GB
- harvard.ld4l.full-catalog-3.2016-03-21.tar.gz -- Converted triples for the Harvard catalog (3 of 4) -- 6.3GB
- harvard.ld4l.full-catalog-4.2016-03-22.tar.gz -- Converted triples for the Harvard catalog (4 of 4) -- 5.4GB
- stanford.ld4l.full-catalog-1.2016-03-23.tar.gz -- Converted triples for the Stanford catalog (1 of 4) -- 3.4GB
- stanford.ld4l.full-catalog-2.2016-03-22.tar.gz -- Converted triples for the Stanford catalog (2 of 4) -- 3.4GB
- stanford.ld4l.full-catalog-3.2016-03-21.tar.gz -- Converted triples for the Stanford catalog (3 of 4) -- 4.0GB
- stanford.ld4l.full-catalog-4.2016-03-22.tar.gz -- Converted triples for the Stanford catalog (4 of 4) -- 4.8GB

Usage data

StackScore usage data is available for the Cornell and Harvard holdings. The scores appear as annotations on the individual bib_ids. Each file contains the usage data for the corresponding, similarly named file of converter output. Data is in N-Triples format.

These data files are available:

- 2016-03-17_cornell_ld4l_full_catalog_anno.tar -- Usage data for the Cornell catalog -- 478MB
- harvard.ld4l.full-catalog-1.2016-03-24_anno.tar -- Usage data for the Harvard catalog (1 of 4) -- 286MB
- harvard.ld4l.full-catalog-2.2016-03-22_anno.tar -- Usage data for the Harvard catalog (2 of 4) -- 272MB
- harvard.ld4l.full-catalog-3.2016-03-21_anno.tar -- Usage data for the Harvard catalog (3 of 4) -- 296MB
- harvard.ld4l.full-catalog-4.2016-03-22_anno.tar -- Usage data for the Harvard catalog (4 of 4) -- 239MB

Additional triples

Additional triples were created to supplement the converter output, adding Work IDs to the Works, and creating links across institutions, between corresponding Works and Instances.

A concordance file was created, associating all known OCLC numbers with their corresponding Work IDs. This file was made with data extracted from a recent Research snapshot of WorldCat, and is structured as follows:

- Column 1: every OCLC number found in a record from both 001 and 019
- Column 2: the current OCLC number for the record, from 001
- Column 3: the current Work ID associated with the record

linked open data Search

Limit your search

Topic

- Open source software 208
- Open Source software 142
- United States 128
- Open-source software 97
- Data mining 69
- Electronic data processing--Distributed processing 69
- - Plasma Physics And Fusion Technology 67
- Plasma Physics And Fusion Technology 67
- Data Acquisition Systems 63
- Database management 63
- Open Plasma Devices 61
- Thermonuclear Devices 61
- Data 60
- Magnetic Mirrors 60
- Information 57
- Electronic books 56
- Numerical Data 56
- Alaska 55
- Computer science 54
- Experimental Data 54
- more »

Language

Library

You searched for: linked open data ✖ Start Over

« Previous | **1** - **10** of **6,427** | Next »

Sort by relevance 10 per page

1. VIVO [electronic resource] : a semantic approach to scholarly networking and discovery / Katy Börner ... [et al.] Bookmark

Library: Stanford University LIBRARIES
Class: Monograph, Work, Text
RDF linked data: http://draft.ld4l.org/stanford/ne9dc47329f8d1d57

2. What Can OpenEI Do For You? [electronic resource] Bookmark

Library: Stanford University LIBRARIES
Class: Monograph, Work, Text
RDF linked data: http://draft.ld4l.org/stanford/n9dfc35463fee4d74

3. Energy Informatics Panel (Presentation) [electronic resource] Bookmark

Library: Stanford University LIBRARIES
Class: Monograph, Work, Text
RDF linked data: http://draft.ld4l.org/stanford/nca4c634e4a3389fb

4. Linked data / Antonella Iacono Bookmark

Library: HARVARD LIBRARY
Class: Monograph, Text, Work
RDF linked data: http://draft.ld4l.org/harvard/nf73a065cfe2e006c

```
@prefix : <http://draft.ld4l.org/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix ld4l: <http://bib.ld4l.org/ontology/> .
@prefix ld4lcornell: <http://draft.ld4l.org/cornell/> .
@prefix ld4lstanford: <http://draft.ld4l.org/stanford/> .
@prefix madsrdf: <http://www.loc.gov/mads/rdf/v1#> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix void: <http://rdfs.org/ns/void#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

ld4lstanford:n1193211b3ab8a954 a oa:Annotation;
   oa:hasTarget ld4lstanford:ne9dc47329f8d1d57 .

ld4lstanford:n6507d1d542761974 a ld4l:Instance,
     ld4l:Electronic;
   rdfs:label "Electronic Resource";
   ld4l:isInstanceOf ld4lstanford:ne9dc47329f8d1d57 .

ld4lstanford:n69c8905b91d82e70 a ld4l:Topic;
   skos:prefLabel "linked open data" .

ld4lstanford:n6e81745abb6b60f1 a ld4l:Topic;
   skos:prefLabel "Semantic Web" .

ld4lstanford:n8d143253a539eccf a madsrdf:Title;
   rdfs:label "VIVO [electronic resource] : a semantic approach to scholarly networking and discovery / Katy Börner ... [et al.]"

ld4lstanford:n903e4f77b106a06a a ld4l:Topic,
     madsrdf:Authority;
   rdfs:label "Scholars--Databases";
   skos:prefLabel "Scholars--Databases" .

ld4lstanford:nab60e0d85d7f15b0 a ld4l:Topic;
   skos:prefLabel "scholarship" .

ld4lstanford:ncfa10279e7cef73d a ld4l:Topic;
   skos:prefLabel "research networking" .

ld4lstanford:nd9bdccade2059f40 a madsrdf:Authority,
     ld4l:Topic;
   rdfs:label "Semantic Web",
     "Semantic web";
   skos:prefLabel "Semantic web",
     "Semantic Web" .
```

LD4L Labs and LD4P

- Beginning April 1, 2016, the Mellon Foundation funded two new linked data grants for two years at $1.5 million each
- LD4L Labs is a collaboration of Cornell, Harvard, Stanford, and Iowa focused on building new LD tools and services
- LD4P is a collaboration of Stanford and five partners piloting a range of projects on metadata production using linked data

LD4L Labs

1. Create a modern, efficient, extensible, community-supported, and well-documented MARC to BIBFRAME converter to support the revised BIBFRAME ontology
2. New LD creation and editing tools based on Vitro
3. New Hydra-based LD tools to support the organization, annotation, and description of scholarly information resources
4. Improve discovery and understanding by using LD for better search results, to visualize context, and to link to resources on the web
5. Pilot tools and services to support LD URI resolution and reconciliation (led by Dave Eichmann, UIowa)

LD4P

- Stanford: Convert four cataloging workflows to LD and create a Performed Music Ontology
- Columbia: Catalog a collection of 2D & 3D art objects*
- Cornell: Develop Rare Materials Ontology and catalog an annotated hip-hop LP record collection*
- Harvard: Catalog cartographic materials*
- Library of Congress: Catalog audiovisual and sound recordings, prints, and photographs
- Princeton: Catalog annotated materials from the library of Jacques Derrida*

* Using the Vitro editor

What's Happening with Vitro?

- Creating a version of Vitro configured with pages, menus, class and property groups, custom views, etc., based on LD4L ontology
- Like VIVO, a full application on top of Vitro
- Creating a full set of Vitro documentation independent from VIVO
- Created a fork of both the Vitro and VIVO repositories, will create our modifications on the fork, and create pull requests to the base repositories
- For new features added to Vitro core, we will work with the VIVO community to get input on design and implementation choices

What is VitroLib?

VitroLib is an ontology-based cataloging editor enabling manual cataloging in RDF based on a generalized bibliographic ontology with extensions for specialized cataloging. Built on Vitro, a generalized semantic web ontology and instance editor with customizable browsing, VitroLib is a joint product of the Mellon-funded Linked Data for Libraries Labs and Linked Data for Production projects.

With VitroLib, you can:

- Create or load ontologies in OWL format
- Edit instances and relationships
- Build a public web site to display your data
- Search your data

Log in

Email

Password

Log in

Search VitroLib

limit search → Search

Statistics

75	88	17	103	53	160
Works	Instances	Items	People	Organizations	Agents

Ontologies

Add new ontology

#	Ontology	Namespace	Prefix
1	Content Ontology	http://www.w3.org/2011/content#	cnt
2	DC Elements	http://purl.org/dc/elements/1.1/	dc
3	DC Terms	http://purl.org/dc/terms/	dcterms
4	FaBIO (FRBR-Aligned Bibliographic Ontology)	http://purl.org/spar/fabio/	fabio
5	FOAF (Friend of a Friend)	http://xmlns.com/foaf/0.1/	foaf
6	LD4L-O (LD4L Proposed Bibliographic Ontology)	http://bib.ld4l.org/ontology/	ld4l
7	Lingvoj	http://www.lingvoj.org/ontology#	lingvo
8	MADS/RDF	http://www.loc.gov/mads/rdf/v1#	madsrdf
9	Open Annotation	http://www.w3.org/ns/oa#	oa
10	OWL (Web Ontology Language)	http://www.w3.org/2002/07/owl#	owl
11	PROV-O (Provenance Ontology)	http://www.w3.org/ns/prov#	prov
12	RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	rdf
13	RDFS	http://www.w3.org/2000/01/rdf-schema#	rdfs
14	Schema.org	http://schema.org/	schema
15	Vitro Public Ontology	http://vitro.mannlib.cornell.edu/ns/vitro/public#	vitro-public

hide subclasses

works

Display Rank: 10

Classes:

- Audio
- Cartography
- Collection
- Continuing resource
- Dataset
- Dissertation
- Finite resource
- Index
- Integrating resource
- Journal
- Magazine
- Monograph
- Moving image
- Multimedia
- Multipart monograph
- Newspaper
- Notated movement
- Notated music
- Periodical
- Periodical part
- Serial
- Series
- Still image

Englische Suiten | Audio | Monograph

Titles | Contributors | Topics | Related Instances | Instances | Annotations | Other | View All

Titles

has title

Englische Suiten

Contributors

has contribution

Bach, Johann Sebastian (Creator)

Hewitt, Angela (Performer)

Topics

subject

Suites (Piano)

New Vitro Webinars (Avail 9/15/16)

- Installation
- Ontologies: a demo version using the LD4L ontology
- Instance data: browsing, ingesting, adding, editing, quick overview of custom forms and custom views
- SPARQL queries
- Data export
- Site configuration
- User admin

- Potential advanced topics (if there is sufficient interest):
  - Ontology editing
  - Faux properties
  - Advanced data tools

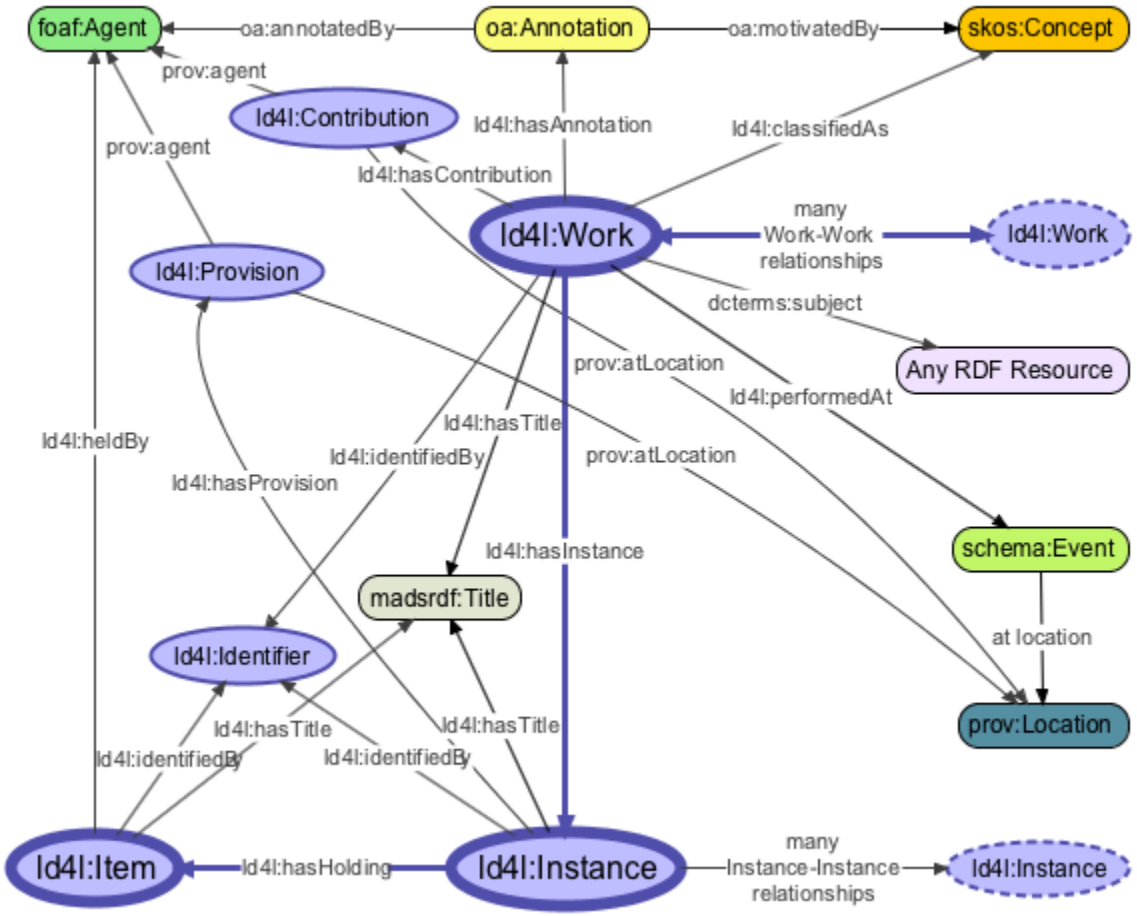
Track the VitroLib project:

https://wiki.duraspace.org/display/ld4l/The+Vitro+Metadata+Editor

Upcoming LD4P Ontology Outputs

- Performed Music Ontology - a BIBFRAME-based ontology for performed music in all formats, with a particular emphasis on clarifying and expanding on the modelling of works, events, and their contributors
- Rare Materials Ontology - define an extension ontology for the description of rare materials, with a particular focus on instance and item-level data

LD4L Ontology



Using the LD4L Ontology in VIVO

- LD4L represents the work, not just the instance (preprint, journal article, book chapter)
- LD4L models relationships among works (derivative, performance, sequel)
- LD4L provides broader range of work types, particularly for non-text works
- LD4L provides richer predicates for subject, genre/form, and geographic & temporal settings of works
- LD4L can model many types of work contributions, e.g., Illustrator, Composer, Conductor, Narrator, etc.
- LD4L provides rich set of identifiers to link local entities to external data sources

National Strategy for Shareable Local Authorities

- Funded at Cornell by IMLS – May 2016-April 2017
- Expert Forum to be held in October 2016
- Need to shift from existing model of local/national authorities to more flexible LD shared authorities
- Need to leverage LD work in VIVO, ORCID, MusicBrainz, SNAC, ISNI, GRID, etc.
- Challenges: governance, workflow, persistence, licensing, technical infrastructure
- Outputs: minimum requirements for creating and publishing local authorities; best practices for linked data authorities; opportunities for peer-to-peer sharing; strategy for entity/vocabulary reconciliation

Thoughts on VIVO and LD4L

- LD4L is working on a broad range of tools and data sources that VIVO can leverage
- VIVO can represent scholarship at or across institutions – not just people's profiles
- VIVO instances can leverage LD4L concepts to be part of a comprehensive LD ecosystem representing the full range of scholarly activity
- **Don't silo your VIVO!**

More Info: http://ld4l.org

Questions?