Getting Started with Library Linked Open Data: Lessons from UNLV and NCSU

Eric Hanson, North Carolina State University
Cory Lampert, University of Nevada, Las Vegas
Silvia B. Southwick, University of Nevada, Las Vegas

June 28, 2015

Co-Sponsors:
ALCTS Creative Ideas in Technical Services Interest Group
LITA/ALCTS Library Linked Data Interest Group
Organization Name Linked Data

Alphabetical Index

# A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Overview

The NCSU Organization Name Linked Data (ONLD) is based on the NCSU Organization Name Authority, a tool maintained by the Acquisitions & Discovery department since 2009 to manage the variant forms of name for journal and e-resource publishers, providers, and vendors in E-Matrix, our locally-developed electronic resource management system (ERMS). The Organization Name Authority was first described by Kristen Blake (Wilson) & Jacque Samples in "Creating organization name authority within an electronic resources management system," Library Resources and Technical Resources, 53(2), 2009, p. 94-97, which is available online here.

The names chosen as the authorized form reflect an acquisitions, rather than bibliographic, orientation. For example, in the Library of Congress Name Authority File, the Institute of Electrical and Electronics Engineers is represented by the full name, whereas as in the NCSU ONLD, it appears as “IEEE,” which is how it is generally known among acquisitions staff. Also, there are many subsidiary units with valid headings in the LC Name Authority File but for the purpose of managing journals and electronic resources they are simply considered to be variant forms of names for the parent organization that manages acquisitions and licensing-related functions for the subsidiaries.

The information in the NCSU Organization Name Linked Data are represented as RDF triples using properties from the SKOS, RDF Schema, FOAF, and OWL vocabularies. Clicking on the name of each property will take users to the property’s definition. The authorized
NCSU Organization Name Linked Data

- [http://www.lib.ncsu.edu/ld/onld/](http://www.lib.ncsu.edu/ld/onld/)

- Based on the NCSU Organization Name Authority, a tool used to manage the variant forms of name for serial and e-resource publishers, providers, and vendors in E-Matrix, our locally-developed electronic resource management system.

- Published in July 2014
Linked Data Creation Process

- Model
- Clean Up
- Augment
- Convert
- Publish
Model

- Find terms from linked data vocabularies to describe the source data you have selected for a linked data project

- General vocabularies can be used to describe a wide variety of data

  RDF Schema [http://www.w3.org/TR/rdf-schema/](http://www.w3.org/TR/rdf-schema/)
  OWL [http://www.w3.org/TR/owl-ref/](http://www.w3.org/TR/owl-ref/)
Model

• Search for more specific vocabulary terms when necessary

Open Metadata Registry [http://metadataregistry.org/](http://metadataregistry.org/)
Linked Open Vocabularies [http://lov.okfn.org/dataset/lov/](http://lov.okfn.org/dataset/lov/)
Clean Up

- Make sure the source data is in a structured format such as XML and as clean as possible before converting to RDF

- Types of errors vary with each data set: duplicate entries, inconsistent formatting, diacritics

- Do your best before conversion, but data clean up will likely happen throughout the conversion process as errors become visible in new formats
Clean Up

- The NCSU Libraries project team used a MS Access database for cleaning up the source data.
- After clean up, the data was exported as an XML file using the MS Access Reports function.
Augment

- Add links to related information in other linked data sets

- Links to other data sets allow users to easily combine information from multiple data sets

**Oxford University Press**

<table>
<thead>
<tr>
<th>VIAF</th>
<th>DBpedia</th>
<th>ISNI</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>LCNAF</th>
<th>Freebase</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://id.loc.gov/authorities/names/n80126136">http://id.loc.gov/authorities/names/n80126136</a></td>
<td><a href="http://rdf.freebase.com/ns/m/0czzl">http://rdf.freebase.com/ns/m/0czzl</a></td>
</tr>
</tbody>
</table>
Augment

- Use the owl:sameAs or skos:closeMatch property with the Uniform Resource Identifier (URI) for related information in another linked data set

  http://www.lib.ncsu.edu/ld/onld/00000008
  Oxford University Press in the NCSU ONLD

  skos:closeMatch

  http://viaf.org/viaf/140705396
  Oxford University Press in VIAF
Augment

- Manually search linked data sets for equivalent URIs
Batch searching for equivalent URIs with OpenRefine

The project team tested an OpenRefine GREL script developed by Matt Carruthers at the University of Michigan, available here: https://github.com/mcarruthers/LCNAF-Named-Entity-Reconciliation
Augment

- This script was modified to create the Organization Name URI Look Up Script which queries the VIAF API for VIAF, LC, and ISNI identifiers and searches DBpedia for matches in the URL

- The results require manual review to clean up incorrect matches

- NCSU’s modified script and a sample data set are available for download at http://www.lib.ncsu.edu/ld/onld/
Convert

• Transform the source data into RDF serializations using a tool of your choice

• The NCSU project team used XSLT to transform our data into various forms of RDF

• XSLT is easy to learn for those with HTML and XML experience
Convert

- Don’t know XSLT or other scripting languages? Learn for free!
  
  W3Schools [http://www.w3schools.com/](http://www.w3schools.com/)

- Troubleshooting specific problems?
  
  Users helping each fix coding problems
Convert

- Choose what RDF serializations to use for publishing
  - N-Triples  http://www.w3.org/TR/n-triples/
  - RDF-XML    http://www.w3.org/TR/rdf-syntax-grammar/
  - N3/Turtle   http://www.w3.org/TeamSubmission/n3/,
                http://www.w3.org/TR/turtle/
  - RDFa        http://www.w3.org/TR/xhtml-rdfa-primer/
  - JSON-LD     http://json-ld.org/
Oxford University Press

Authorized Name

skos:prefLabel  Oxford University Press

Website

foaf:homepage  http://global.oup.com/
Convert

HTML with RDFa - HTML View

</ul><br><h2 style="font-weight:bold;">Oxford University Press</h2><br><h3>Authorized Name</h3><hr><table><tr><td><a href="http://www.w3.org/2004/02/skos/core#prefLabel">skos:prefLabel</a></td><td><span property="skos:prefLabel">Oxford University Press</span></td></tr></table><h3>Website</h3><hr><table><tr><td><a href="http://xmlns.com/foaf/0.1/homepage">foaf:homepage</a></td></tr></table>
Publish

• Make your new linked data set available on the web with an open data license such as CC0 and ODC-BY

• Announce to listservs and colleagues

• Register your data sets with linked data registries such as http://www.datahub.io

• Provide contact information so users can report errors

• Create a plan for updating and maintaining the data set
Publish

• Show your work!

• Linked data is about sharing information so make your process and tools available for others to use as a model

• We made our XSLT files and some sample XML data available on the project website http://www.lib.ncsu.edu/ld/onld/
• For more detail on the NCSU Libraries’ first linked data project:

Cory Lampert (Presenter)
Head, Digital Collections
UNLV University Libraries
cory.lampert@unlv.edu

Silvia Southwick
Digital Collections Metadata Librarian
UNLV University Libraries
silvia.southwick@unlv.edu

Thank you to Melissa Stoner (UNLV) for her work testing the NCSU data.
UNLV Linked Open Data Project Goals

Study the feasibility of developing a common process that would allow the conversion of our collection records into linked data preserving their original expressivity and richness.

Publish data from our collections in the Linked Open Data Cloud to improve discoverability and connections across our collections and with data from other related data sets on the Web.
More information on UNLV Linked Data Project

Leading to Linking: Introducing Linked Data to Academic Library Digital Collections:

A Guide for Transforming Digital Collections Metadata into Linked Data Using Open Source Technologies:
http://www.tandfonline.com/doi/pdf/10.1080/19386389.2015.1007009

UNLV Linked Data Blog (videos posted here):
https://www.library.unlv.edu/linked-data

Data published at:

Or contact us via email!
Actions

Phase 1
- Clean data
- Export data

Phase 2
- Import data
- Prepare data
- Reconcile
- Generate triples
- Export RDF

Phase 3
- Import data
- Publish

Technologies

CONTENTdm

OpenRefine

Mulgara / Virtuoso
Principles Guiding the Project

Consistency
Interoperability
Transformation Integrity
ID all things with URIs
Sustainability
Sharing

-> Result: New metadata focus
Our focus on metadata

- Why?
  - Metadata is essential for establishing relationships

- What metadata?
  - Ability of discovering relationships is directly affected by metadata quality

- It is critical to:
  - Use well-established Controlled Vocabularies (particularly if they are linked data ready)
  - Rigorously control local terms
  - Re-use URIs
  - Assign URIs for local terms
Metadata Creation Issues - current

- Focus is on the collection being created
  - Usually metadata consistency is managed within collections

- Not much rigor is used to enter controlled vocabulary terms
  - Exs.: Misspellings, use of terms that do not match the preferred terms, etc.

- Limited control of local terms

- Implications:
  - Ability to identify relationships within and across collections is decreased
Evolution of Metadata

UNLV Application Profile

Shared Controlled Vocabularies

Increased rigor in managing CV; workflow modifications

Cleaning data “lake” for downstream quality in linked data
Re-thinking Metadata

• Well-established CVs ➔ allow re-use of URIs
• Rigorous rules of data entry ➔ facilitate reconciliation
• Local Controlled vocabularies ➔ allow interlinking among local terms / names within collections
• Shared vocabularies ➔ allow interlinkage among local terms / names across collections
All these actions:

Allow creation of a single process to transform digital collections into linked data
Metadata Transformation Process

Learning: [The Story of the Super Skeleton]

New Process:

• Script 1: Cleaning the Data
• Script 2: Reconciliation with Authority Files
  *STOP : Manual Cleaning*
• Script 3: Fetching URIs
• Script 4: Creating the Skeleton/Generate Triples
Trim leading and trailing white spaces in columns that will be reconciled for work creators, creators, title, original format, geographic location, work headings, title.
Cleaning (Script 1), con’t.

Split multi-value cells in:
- work creators,
- work geographic locations,
- subjects,
- creators,
- original format,
Reconciliation (Script 2)

Before reconciliation happens you must have the RDF extension.

Extension available at: refine.deri.ie

Once downloaded move the extracted file into the extensions folder of the webapp folder w/in your google-refine folder.
Reconciliation (Script 2), con’t.
Reconciliation (Script 2), con’t.
Pause – Manual Review

OpenRefine shows a green bar for reconciliation progress

At this point in the transformation, we verify that all fields have a full green bar – if less than 100% are matched we search for matches.
Fetching URIs (Script 3)

Create new column called `work creators uri`, this will be based on the column `work creators`. URIs from LoC will be placed in this column.
Fetching URIs (Script 3), con’t.

Select +add prefix; a New Prefix box will appear, in the prefix: field enter edm. The URI: field will fill in automatically, click ok. Continue to add namespaces (prefix) for: dc; dcterms; and skos.
Skeleton (Script 4)
Linked Data Triples (product)
Data Exchange Project

Photo: Photograph of Bluebells posing outside of Pan Am jet, 1958, Donn Arden Collection
Data Exchange

- Testing how our tools can be adapted for use with other data sets
- UNLV sent NCSU a list of organizations for use with the Organization Name URI Look Up Script
- NCSU sent UNLV a sample of digital collections metadata to test their transformation process on a non-UNLV created data set.
Data Exchange (NCSU)

- Ran the Organization Name URI Look Up Script with a list of organizations from UNLV

- Modifying the script was simple because both lists contained organization names that could be searched against VIAF’s “Corporate Names” index
The Organization Name URI Look Up Script had a lower match rate because many of the organizations on the UNLV list were not represented in VIAF but the script also missed some that were.

Adapting a linked data look up tool to a new set of source data requires consideration of what linked data sets that tool should be searching against.

Testing against new data sets can identify ways to improve the tool.
Data Exchange (UNLV):
How Common is the “Common Process”? 

- Loaded the NCSU data into OpenRefine
- Ran 4 scripts on the data
  - Cleaning
  - Reconciliation
  - Fetching URIs
  - Skeleton

- Consulted with NCSU on the match rates and discussed advantages of each method
Data Exchange (UNLV):
Challenges of Automated Reconciliation

System Terms
• Full Match
• Partial Match
• No Match

Real Life Scenarios
• No match (true) - local
• No match (false) – ambiguous term
• Partial match – content context
• Full match (false)-content context
• Full match (true) – YAY!
Lessons Learned: Metadata Practice

Repositories using a common application profile will be more successful.

Using multiple CVs in one field can be problematic with the transformation process (1 field : 1 property).

Consistency in use of standards is important for mapping fields to linked data model (DC Original Format).
Lessons Learned: Linked Data Projects

Different approaches have different goals / advantages

NCSU process very good for determining “same as”

Unexpected matching results led to investigations/future research
Getting started with linked data at your library

- Try a simple project as a learning opportunity
- The skills learned and tools built for your first project will make the second project much easier
- Librarians have a chance in these early stages of linked data to help establish best practices
- Librarians have been working with authority control for decades and we should share our expertise as linked data moves forward
From experimentation to implementation

Manage cost/value of work:
• Cleaning and sharing controlled vocabularies from legacy collections (time consuming)

Change in staff roles:
• Re-training metadata creators
• Re-designing workflows
• Delegating additional data management responsibilities
Questions?

Eric Hanson  emhanso2@ncsu.edu
Cory Lampert  cory.lampert@unlv.edu
Silvia Southwick  silvia.southwick@unlv.edu
Thank you for attending this ALCTS program

Your feedback is important to us! Using the QR code or URL below please take a moment to complete a short online evaluation form and enter to win the ALCTS prize drawing

http://tinyurl.com/alctsac15