A Step Forward: Adding Linked Data Vocabularies to Digital Repositories

5-24-2018

Sai Deng
University of Central Florida, sai.deng@ucf.edu

Find similar works at: https://stars.library.ucf.edu/ucfscholar

University of Central Florida Libraries http://library.ucf.edu

Part of the Cataloging and Metadata Commons

Original Citation


This Conference Presentation is brought to you for free and open access by STARS. It has been accepted for inclusion in Faculty Scholarship and Creative Works by an authorized administrator of STARS. For more information, please contact lee.dotson@ucf.edu.
A Step Forward: Adding Linked Data Vocabularies to Digital Repositories

Sai Deng, University of Central Florida Libraries
Linked Data Basics

• Linked data:
  • “A term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF.”
    -Wikipedia

• Resource Description Framework (RDF): Represent information in triples
  • **Subject**: a resource, identified with a URI
  • **Predicate**: specification of the relationship
  • **Object**: a resource the subject is related to
Linked Data Basics

• The Four Design Principles of Linked Data
  • Use Uniform Resource Identifiers (URIs) as names for things.
  • Use HTTP URIs so that people can look up those names.
  • When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
  • Include links to other URIs so that they can discover more things.

• URIs for naming things, RDF for describing things, SPARQL for querying data; link to other URIs to find more things.
Linked Data: An Example

Comedy of The tempest

Is an example of schema:exampleOfWork

The tempest

Is about schema:about

Plays

Spirits

Fathers and daughters

Shipwreck victims

Is written by schema:creator

Henry V

A midsummer night's dream

Is written by schema:creator

The Norton Shakespeare

Has format schema:bookFormat

John C. Crowther

Is contributed by schema:contributor

PrintBook

*Based on information from worldcat.org
Linked Open Data Building Blocks

- **The LOD Data Model**: RDF triples
- **Content Rules**: relies on external vocabularies and ontologies
- **Metadata schema**
  - RDF Schema
  - Vocabulary Building Blocks: SKOS, OWL
  - Open Archives Initiative Object Reuse and Exchange (OAI-ORE)
- **Serializations of LOD**
  - RDF Notation-3/Ns, Turtle, N-Triples; RDFa; JSON-LD
- **Exchanging LOD**
  - SPARQL
- **Mitchell, Erik T.** Building Blocks of Linked Open Data in Libraries. *Library Technology Reports* (vol. 49, no. 5)
Five Star-schema of Linked Open Data (LOD)

• Five star-schema of Linked Open Data (Tim Berners-Lee)

★ Available on the web (whatever format) but with an open licence, to be Open Data

★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)

★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)

★★★★ All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

★★★★★ All the above, plus: Link your data to other people’s data to provide context

Source: http://5stardata.info/en/
Linked Data Examples and Datasets

- LC Linked Data Service: Authorities and Vocabularies [https://id.loc.gov/](https://id.loc.gov/)
- FAST: Faceted Application of Subject Terminology [http://id.worldcat.org/fast](http://id.worldcat.org/fast)
- Virtual International Authority File (VIAF) [https://viaf.org/](https://viaf.org/)
- Worldcat Identities [www.worldcat.org/identities](http://www.worldcat.org/identities)
- Open Researcher and Contributor ID (ORCID) [https://orcid.org/](https://orcid.org/)
- International Standard Name Identifier (ISNI) [http://www.isni.org/](http://www.isni.org/)
- Schema.org Vocabulary [http://schema.org/docs/gs.html#schemaorg](http://schema.org/docs/gs.html#schemaorg)
- The Linked Open Data Cloud [https://lod-cloud.net/](https://lod-cloud.net/)
Linked Data Implementations

- Google Knowledge Graph
Linked Data Implementations

- **Libraries, Archives and Museums Community**
  - Europeana [http://europeana.eu](http://europeana.eu)
  - DPLA [http://dp.la](http://dp.la)
  - LIBRIS [https://libris.kb.se/](https://libris.kb.se/)
  - British National Bibliography [http://bnb.bl.uk/](http://bnb.bl.uk/)
  - BIBFRAME [http://BIBFRAME.org/](http://BIBFRAME.org/)

- **The Web and the Library Community**
  - Schema.org
    - Bibliographic Extension
    - Extend Schema.org to represent archives
      [https://archival.github.io/schema-org/](https://archival.github.io/schema-org/)

- **More Library Experiments/Cases**
  - WorldCat Works (schema:exampleOfWork) [http://experiment.worldcat.org/entity/work/data/1080130963](http://experiment.worldcat.org/entity/work/data/1080130963)
  - Check out ALA conference programs on LOD in the past few years, e.g., Linked Library Data IG
Current Digital Repositories and Linked Data

• Islandora
  • Effort in mapping MODS to RDF;
  • Fedora 4, LDP implementation; Islandora CLAW

• Experiments and Cases:
  • Islandora CLAW
  • TU Delft repository example: Colonial Architecture (Delftdora module, Fedora 3)
  • A Lightweight Structured Data Implementation Using JSON-LD and Schema.org for Digital Repository (Mak et al.)
Current Digital Repositories and Linked Data

- **CONTENTdm**
- Not inside the system, but work on exported data;

- **UIUC.** Challenges of Mapping Digital Collections Metadata to Schema.org (Lampron et al. 2016)


A UIUC Experiment: [http://imagesearch-test1.library.illinois.edu/cdm/ref/collection/motley-new/id/599](http://imagesearch-test1.library.illinois.edu/cdm/ref/collection/motley-new/id/599)
Project Site: [http://publish.illinois.edu/linkedspcollections/](http://publish.illinois.edu/linkedspcollections/)
Current Digital Repositories and Linked Data

• Omeka
  • Open source web publishing platform. Supports Dublin Core. Can export METS, dc-rdf, JSON etc.
  • [Ozmeka (plug-ins)](https://wiki.duraspace.org/display/DSDOC6x/Linked%20(Open)%20Data)/ Omeka linked data

• DSpace
  • DSpace Linked Data support spans all three Layers: “the storage layer with a triple store, the business logic with classes to convert stored contents into RDF, and the application layer with a module to publish RDF serializations.”

  - Donohue. Linked (Open) Data.
    ([https://wiki.duraspace.org/display/DSDOC6x/Linked+%28Open%29+Data](https://wiki.duraspace.org/display/DSDOC6x/Linked+%28Open%29+Data))
Current Digital Repositories and Linked Data

• Digital Commons
  • Allow to add extra fields for links (not RDF)

• Preparing for Linked Data
  • Include links (e.g., id.loc.gov, viaf.org) in added additional fields for author, advisor, department names etc.;

  • Data (in spreadsheet) cleanup and reconciliation using OpenRefine
    • Name review can be tricky and time-consuming: Google, Wikipedia, obituary, university and departmental websites, publications, CVs...
Adding Linked Data Vocabularies to UCF Digital Repositories

• Adding LOD Links to Fields in Digital Repositories
  • Authors/Creators
  • Contributors, e.g., thesis advisors
  • Corporate names, e.g., university, college and departmental names
  • Subjects
  • Person identifiers

• OpenRefine
  • Download and installation
    http://openrefine.org/download.html
  • Documentation
    https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users
  • Features
    • Import
    • Facet
    • Edit cells: Clustering...
    • Edit columns: Add Column based on this column; Fetching URLs from Web Services (*Enrich existing data columns)...
    • Expressions
    • Export
    • History...
Adding LC and VIAF Links to Advisor Names

Open OpenRefine — Choose Files — Select advisors-RTDs.xlsx — Click “Next”
Adding LC and VIAF Links to Advisor Names

Click “Create Project”
Adding LC and VIAF Links to Advisor Names

After clicking “Facet – Text facet”
Adding LC and VIAF Links to Advisor Names

Click “Cluster” on the left, or click “Edit Cells – Cluster and edit” under “advisor” column
Adding LC and VIAF Links to Advisor Names

Cluster & Edit column “advisor”
(Method: key collision, Keying Function: fingerprint)

E.g., Hagedoorn, A. Henry, Hagedoorn, Henry A. → Hagedoorn, A. Henry
Adding LC and VIAF Links to Advisor Names

### Clustering

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th>Row Count</th>
<th>Values in Cluster</th>
<th>Merge?</th>
<th>New Cell Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8</td>
<td>• Frank, Fredric D. (7 rows)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Frank, Frederic D. (1 row)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>• Schrader, George F. (5 rows)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Schrader, George A. (1 row)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>• Snelson, Franklin F. (4 rows)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Snelson, Franklin (1)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cluster & Edit column “advisor”: Try other algorithms
Adding LC and VIAF Links to Advisor Names

Reconciliation

Sort by advisor;
Click “Reconcile—Start reconciliation”
Adding LC and VIAF Links to Advisor Names

Reconciliation

How to set up LoC Reconciliation Service

Click “Add Standard Service”;

Enter the service's URL:

https://lc-reconcile.herokuapp.com/

Add Service
Adding LC and VIAF Links to Advisor Names

Reconciliation

Leave “Names” as default; Click “Start Reconciling”

Matched 313 (about 32.33%).
Adding LC and VIAF Links to Advisor Names

Review

- For names linked, click the name to check;
- If it’s not the right person, click “Choose new match”;
- It opens several options, re-check and apply.
### Adding LC and VIAF Links to Advisor Names

**Review**

- For names unlinked, click the other options to check;
- If it’s the right person, click ✓ (apply to this cell only) or ✓ ✓ (apply to all identical cells) to apply.

<table>
<thead>
<tr>
<th>802.</th>
<th>DP0015862</th>
<th>Acierno, Louis J.</th>
<th></th>
<th></th>
<th>![edit.png]</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>Acierno, Louis J., 1920- (86)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Click the link under “Acierno, Louis J.”, it opens: [http://id.loc.gov/authorities/names/n85099714.html](http://id.loc.gov/authorities/names/n85099714.html)

**Acierno, Louis J., 1920-**

**Sources**

- found: His Comprehensive cardiac rehabilitation and prevention, c1985: CIP t.p. (Louis J. Acierno, M.D., F.A.C.C., F.A.C.P.; associate prof. of pub. health, Univ. of Central Fla.) data sheet (b. 1920)
- found: His A history of cardiology, 1993: CIP t.p. (L.J. Acierno; Univ. of Florida, USA)
Adding LC and VIAF Links to Advisor Names

Clausen, Chris A., 1940- http://id.loc.gov/authorities/names/n78032138.html

Sources
found: His Principles of industrial ... 1978: t.p. (Chris A. Clausen, Dept. of Chem., Fla. Tech. U., Orlando) pub. info. (b. 12/7/40)

Doering, Robert D., 1925- http://id.loc.gov/authorities/names/no2018055990.html

Has Affiliation
Organization: Florida Technological University
Organization: University of Central Florida. Department of Industrial Engineering and Management Systems
Associated Locale Orlando (Fla.)

Review

• For names unlinked, click the other options to check (e.g., the LC links below);
• If it’s the right person, click ✓ or ✓✓ to apply;
• If no right name found, click “Create new topic”;
• Input the right name and apply.
Adding LC and VIAF Links to Advisor Names

Narsingh Deo, 1936-
http://id.loc.gov/authorities/names/n2006056273.html

Sources
found: Sysło, Maciej M. *Discrete optimization algorithms*, 2007: CIP t.p. (Narsingh Deo)
found: LC database, July 20, 2006: (Deo, Narsingh, 1936-)


...Narsingh Deo is a professor and Charles N. Millican Endowed Chair of the Department of Computer Science, University of Central Florida...

Books

Review

- If LCNAF doesn’t provide enough information, further check the web (e.g., Wikipedia, Amazon books, obituaries, university website, newsletter...)
- When confirmed that it’s the right person, click ✓ or ✓✓ to apply.
Adding LC and VIAF Links to Advisor Names

Add Column for URLs

- Edit column—Add column based on this column
- In the opened new window, fill in:
  - New column name: `advisor_lc`
  - Expression: `cell.recon.match.id`
Adding LC and VIAF Links to Advisor Names

Add Column for URLs: LC Links added
Adding LC and VIAF Links to Advisor Names

Add a duplicate column for advisor

New column name: advisor_c
Expression: value

[advisor_c] Reconciliation
Select Virtual International Authority File service
(*Give more options to choose from if not matched directly)

Or, select VIAF-LC service

How to set up VIAF-LC Reconciliation Service
Click “Add Standard Service”;
Enter the service’s URL:
http://refine.codefork.com/reconcile/viaf/LC

How to set up Virtual International Authority File service
Enter the service’s URL:
http://iphylo.org/~rpage/phyloinformatics/services/reconciliation_viaf.php
Adding LC and VIAF Links to Advisor Names

Add VIAF Links

New column name: advisor_viaf
Expression: "https://viaf.org/viaf/" + cell.recon.match.id

VIAF links added
Adding LC and VIAF Links to Advisor Names

• Display in the IR

Adding LOD links for the advisor names is a work in progress.

*Add both links or only one type of link for advisors (e.g., VIAF, or LC)
Adding LC and VIAF Links to Author Names

- Using Expression to Format Multiple Links
  - Expression Language: General Refine Expression Language (GREL)
  - First, run reconciliations and get VIAF and LC links;
  - Then add columns, e.g.,
    - Edit column – Add column based on this column
    - New Column name: author1_viafurl
      - Expression: "http://viaf.org/viaf/" + cell.recon.match.id
    - Edit column – Add column based on this column
      - New column name: author1_lcurl
        - Expression: cell.recon.match.id
Adding LC and VIAF Links to Author Names

- Using Expression to Format Multiple Links
  - Add column based on this column
  - New column name: author1_viafurl_lcurl
  - Expression
    
    ```
    cells["author1_viafurl"].value + " " + cells["author1"].value + "[VIAF] " + cells["author1_lcurl"].value + " " + cells["author1_copy"].value + "[LC]"
    ```

```

<table>
<thead>
<tr>
<th></th>
<th>author1_lcurl</th>
<th>author1_viafurl</th>
<th>author1_viafurl_lcurl</th>
</tr>
</thead>
</table>
```
Creating a project for one “UCF Forum” dataset in OpenRefine;
- Its keywords field has a mixture of controlled and uncontrolled terms taken from the articles and supplied by the Metadata Librarian;
- The goal is to add LOD links for the controlled terms.

Splitting Keywords into Several Columns

Under “Keywords”, click “Edit column – Split into several columns”; “by separator “;”
Adding LC Subject Links to IR Records

- Create a project for one “UCF Forum” dataset in OpenRefine;
- Keywords splitted into Several Columns

Top graph:
Keywords splitted (by default, e.g., keywords 1, 2, 3...)

Rename the fields, delete the space in field names
keywords 1, 2, 3... → keyword1, 2, 3...

Left graph:
Reconcile column “keyword1”
Reconcile each cell against LoC Reconciliation Service
Adding LC Subject Links to IR Records

• Review, select and confirm matched values (when needed)

Edit column – Add column based on this column
Adding LC Subject Links to IR Records

• Adding column for links

New column name
keyword1_url

Expression
cell.recon.match.id + " | " + cell.recon.match.name

Links added for keyword1;
Repeating the same process for keyword2, 3, 4...
Adding LC Subject Links to IR Records

• Putting Links together

Putting all subjects and their links together in a column

**New column name**  subjects_linked

**Expression**  

```grel
if(isBlank(cells["keyword1_url"].value), " ",
cells["keyword1_url"].value) + " ",
if(isBlank(cells["keyword2_url"].value), " ",
cells["keyword2_url"].value) + " ",
if(isBlank(cells["keyword3_url"].value), " ",
cells["keyword3_url"].value) + " ",
if(isBlank(cells["keyword4_url"].value), " ",
cells["keyword4_url"].value) + " ",
if(isBlank(cells["keyword5_url"].value), " ",
cells["keyword5_url"].value) + " ",
if(isBlank(cells["keyword6_url"].value), " ",
cells["keyword6_url"].value)
```

(*can add more values)

Subject links display together

* Use “|” or other dividers to separate subjects
Adding LOD Links to IR Records

An Example in the IR
http://stars.library.ucf.edu/ucf-forum/186/
Adding FAST by Fetching URLs

Add column by fetching URLs based on Topical Subject column
(Click “Edit column—add column by fetching URLs” under “Topical_Subject”)

New column name
fast_url

Expression
"http://experimental.worldcat.org/fast/search?query=oclc.heading+exact+%22"+value.replace(/(\s)/,'%20')+%22&httpAccept=application/rdf%2bxml"

*Query FAST for RDFXML return
Adding FAST by Fetching URLs

Need to extract FAST ID from the returned content:
<dct:identifier>799558</dct:identifier>
...
<dc:format>application/rdf+xml</dc:format>
Adding FAST by Fetching URLs

Text Transformation

Edit cells – Transform...
Adding FAST by Fetching URLs

Text Transformation
(Extract FAST identifier from the returned content)

Expression
"http://experimental.worldcat.org/fast/"+value.parseHtml().select('dct|identifier')[0].htmlText()
Adding FAST by Fetching URLs

*If returned text doesn’t include a FAST identifier, it’ll display the original text; If more than one match, will pick up the first one (index[0])
Other Possibilities

• Augment Data with Reconciliation Services
  • Reconcile genre values to AAT terms
  • Reconcile places to GeoNames
  • Reconcile ORCID...

• Reconcile Against Local Vocabularies
  • reconciliation-csv
    • Possibilities: Reconcile various fields in OpenRefine against local controlled term lists, e.g., medals, ranks for Veterans Oral History

• Next Steps
  • Keep adding more LOD links to personal names, subjects and college/department names to digital collections;
  • Continue to explore OpenRefine and related services to enrich metadata and improve workflow;
  • Promote, explore and link authors/researchers to personal identifiers.
Resources

- Berners-Lee, Tim. Linked Data https://www.w3.org/DesignIssues/LinkedData.html
- BIBFRAME http://BIBFRAME.org/
- DBpedia http://wiki.dbpedia.org/
- DPLA http://dp.la
- Europeana http://europeana.eu
- FAST: Faceted Application of Subject Terminology http://id.worldcat.org/fast
- GeoNames http://www.geonames.org/ontology/documentation.html
- International Standard Name Identifier (ISNI) http://www.isni.org/
- LC Linked Data Service: Authorities and Vocabularies https://id.loc.gov/
- Mitchell, Erik T. Building Blocks of Linked Open Data in Libraries. Library Technology Reports (vol. 49, no. 5)
- Mitchell, Erik T. Three Case Studies in Linked Open Data. Library Technology Reports (vol. 49, no. 5)
- OCLC Linked Data https://www.oclc.org/developer/develop/linked-data.en.html
- Open Researcher and Contributor ID (ORCID) https://orcid.org/
Resources

- OpenRefine. Reconcilable data sources. [https://github.com/OpenRefine/OpenRefine/wiki/Reconcilable-Data-Sources](https://github.com/OpenRefine/OpenRefine/wiki/Reconcilable-Data-Sources)
- Orbis Cascade Alliance. OpenRefine for Metadata Cleanup. [https://www.orbiscascade.org/openrefine-for-metadata-cleanup/](https://www.orbiscascade.org/openrefine-for-metadata-cleanup/)
- Ozmeka. [https://github.com/ozmeka/ozmeka](https://github.com/ozmeka/ozmeka)
  Reconcile-csv in GitHub: [https://github.com/okfn/reconcile-csv](https://github.com/okfn/reconcile-csv)
- Schema.org Vocabulary [http://schema.org/docs/gs.html#schemaorg](http://schema.org/docs/gs.html#schemaorg)
- Tay, Aaron. OpenRefine, & reconciliation services - learning the ropes. [http://musingsaboutlibrarianship.blogspot.com/2017/03/openrefine-reconciliation-services.html](http://musingsaboutlibrarianship.blogspot.com/2017/03/openrefine-reconciliation-services.html)
- The Linked Open Data Cloud [https://lod-cloud.net/](https://lod-cloud.net/)
- Virtual International Authority File (VIAF) [https://viaf.org/](https://viaf.org/)
- Worldcat Identities [www.worldcat.org/identities](http://www.worldcat.org/identities)
Acknowledgements

• Lucas Mak, Metadata and Catalog Librarian, Michigan State University Libraries

• Systems, University of Central Florida Libraries

• Image credits
  • 5 Star Open data (http://5stardata.info/en/)
  • Linked Data (https://www.w3.org/DesignIssues/LinkedData.html)
  • Linked Open Data Cloud (https://lod-cloud.net/)
Thank you!

Contact

Sai Deng
Metadata Librarian & Associate Librarian
University of Central Florida Libraries
sai.deng@ucf.edu