Authority data

the good, the dirty and the semantic

E. Lynette Rayle
Cornell University

Dave Eichmann
The University of Iowa

Linked Data for Production Grant Funded by Andrew W. Mellon Foundation
Who we are...

This work is part of the LD4P project and done in support of accessing controlled vocabularies from various authorities.

Lynette works on the normalization layer that provides a consistent layout of the data to allow the end applications to work with data in the same way regardless of where the data is coming from and what ontology is used to represent the data.

Dave works on the backend triple store caching system and indexing.
Architecture

Application - Sinopia

Normalization

Authority
Architecture

Application - Sinopia

Normalization - Questioning Authority (QA gem), Lookup Service (QaServer engine)

Authority
Architecture

Application - Sinopia

Normalization - Lookup Service

Authority
Architecture

Application - Sinopia

Normalization - Lookup Service

Jena-Fuseki-Lucene Cache

Direct Access of External Authority
Lookup Search Request

Application - Sinopia

Normalization - Lookup Service

Direct Access to OCLC FAST
Lookup Search Request

Application - Sinopia

https://lookup.ld4l.org/authorities/search/linked_data/oclca_fast/personal_name?q=twain&maximumRecords=2

Normalization - Lookup Service

Direct Access to OCLC FAST
Lookup Search Request

Application - Sinopia

http://lookup.ld4l.org/authorities/search/linked_data/oclcpersonal_name?q=twain&maximumRecords=2

Normalization - Lookup Service

http://experimental.worldcat.org/fast/search?query=oclpersonalName+"twain"&sortKeys=usage&maximumRecords=2

Direct Access to OCLC FAST

* search of cache performed via Lucene index
Lookup Search Request

Application - Sinopia

http://lookup.ld4l.org/authorities/search/linked_data/oclc_fast/personal_name?q=twain&maximumRecords=2

Normalization - Lookup Service

http://experimental.worldcat.org/fast/search?query=oclc.personalName+%22twain%22&sortKeys=usage&maximumRecords=2

Direct Access to OCLC FAST

  <rdf:Description rdf:about="31622">
    <dct:identifier>31622</dct:identifier>
    <dct:replaces rdf:resource="1986616"/>
    <skos:inScheme rdf:resource="ontology/1.0/#fast"/>
    <rdf:type rdf:resource="http://schema.org/Person"/>
    <skos:prefLabel>Twain, Mark, 1835-1910</skos:prefLabel>
    <schema:name>Twain, Mark, 1835-1910</schema:name>
    <skos:altLabel>Clemens, Samuel Langhorne, 1835-1910</skos:altLabel>
    <schema:name>Clemens, Samuel Langhorne, 1835-1910</schema:name>
    <skos:altLabel>Make Tevisin, 1835-1910</skos:altLabel>
    <schema:name>Make Tevisin, 1835-1910</schema:name>
  </rdf:Description>
</rdf:RDF>
Architecture

Application - Sinopia

Normalization - Lookup Service

- Jena-Fuseki-Lucene Cache
- Direct Access of External Authority
Initial Basic Use Case
Simple auto-complete for a field -- type search query

**Creator** required

The person or group responsible for the work. Usually this is the author of the content. Personal names should be entered with the last name first, e.g. "Smith, John."

John Smith
Normalized JSON from the Lookup Service

[
  {
    "uri": "http://id.loc.gov/authorities/names/n85389596",
    "id": "n 85389596",
    "label": "Smith, John W. (John Williamson)"
  },
  {
    "uri": "http://id.loc.gov/authorities/names/n86851637",
    "id": "n 86851637",
    "label": "Smith, John W. (John Williamson), 1897-1986"
  },
  {
    "uri": "http://id.loc.gov/authorities/names/n88005259",
    "id": "n 88005259",
    "label": "Simon, John S. (John Smith), 1843-1933"
  },
  ...
]
Simple auto-complete for a field -- make a selection

Creator <required>

The person or group responsible for the work. Usually this is the author of the content. Personal names should be entered with the last name first, e.g. "Smith, John."

John Smith

- Smith, John W. (John Williamson)
- Smith, John W. (John Williamson), 1897-1986
- Simon, John S. (John Smith), 1843-1933
- St. John-Smith, C. (Christopher), 1956-
- Smith, John Austin, 1954-
- Walker-Smith, John A.
- Smith, John, 1921-

Requirements

- Describe your work

Visibility

- Open Access Everyone. Check out SHERPA/RoMEO for specific publishers' copyright policies if you plan to patent and/or publish your Demo Work in a journal.
- Institution Restrict access to only users and/or groups from Institution
- Embargo
- Lease
- Private Only users and/or groups that have been given
Simple auto-complete for a field -- set value in field

Creator **required**

The person or group responsible for the work. Usually this is the author of the content. Personal names should be entered with the last name first, e.g. "Smith, John."

Smith, John W. (John Williamson), 1897-1986

Requirements

- [x] Describe your work

Visibility

- Open Access: Everyone. Check out SHERPA/RoMEO for specific publishers' copyright policies if you plan to patent and/or publish your Demo Work in a journal.
- Institution: Restrict access to only users and/or groups from Institution
- Embargo
- Lease
- Private: Only users and/or groups that have been given access
Simple auto-complete for a field -- also store URI

Creator

The person or group responsible for the work. Usually this is the author of the content. Personal names should be entered with the last name first, e.g. "Smith, John."

Smith, John W. (John Williamson), 1897-1986

Uri: http://id.loc.gov/authorities/names/n86851637
Extended Context
Lookup with Context

LOC Names: locnames_id4i_cache/person

Search

Barcode*
Enumeration and chronology

Frequency (RDA 2.14)

Search

Note on Frequency (RDA 2.17.12)

Note

Mint URI
Add

LITERAL WITH DEFAULT

LITERAL WITH DEFAULT

DLC X Language English
Lookup with Context -- type search query

LOC Names_lochames_lo41_cache/person

Search

Barcode
+ Enumeration and chronology

Frequency (RDA 2.14)

Frequency (RDA 2.14)

Note on Frequency (RDA 2.17.12)

Note
++ Note

LITERAL WITH DEFAULT

LITERAL WITH DEFAULT

DLC X Language English
Lookup with Context -- selected entity

LOC Names: lochones ld41_cache/person

Taylor, Sam

Search

Frequency (RDA 2.14)

Note on Frequency (RDA 2.17.12)

Note

LITERAL WITH DEFAULT

LITERAL WITH DEFAULT

DLC X Language English
Lookup with Context -- application stores URI

Uri: http://id.loc.gov/authorities/names/n87118067
Bringing in data to multiple fields

(On the near horizon for Sinopia)
Sinopia is a project created with public domain collections.

Mellon Foundation. All metadata

0.1.0 Universal Public Domain
[Clone|Edit] Name of Resource

**Title Information**

- **Instance Title**
  - Title Proper (RDA 2.3.2) (BIBFRAME: Main title)
    - Title Proper (RDA 2.3.2) (BIBFRAME: Main title) [Sorry / Why Remind Me]
  - Other Title Information (RDA 2.3.4) (BIBFRAME: Subtitle)
    - Part number
    - Part name
    - Note on title

- **Genres**
  - Genre Form (BIBFRAME: Genre Form)
    - Genre Form (BIBFRAME: Genre Form) [Jazz]
  - Base Material (BIBFRAME: Base Material)
    - Base Material (BIBFRAME: Base Material) [Vinyl]
  - Carrier (BIBFRAME: Carrier)
    - Carrier (BIBFRAME: Carrier) [audio disc]
<table>
<thead>
<tr>
<th>Instance of</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIBFRAME Work</td>
</tr>
<tr>
<td>Title Information</td>
</tr>
<tr>
<td>(Time) Coverage of the Content (RDA 7.3)</td>
</tr>
<tr>
<td>Notes about the Work</td>
</tr>
<tr>
<td>Content Type (RDA 6.9)</td>
</tr>
<tr>
<td>Content Type (RDA 6.9)</td>
</tr>
<tr>
<td>Lookup</td>
</tr>
<tr>
<td>Illustrious Content (RDA 7.15)</td>
</tr>
<tr>
<td>Color Content (RDA 7.17)</td>
</tr>
<tr>
<td>Has BIBFRAME Instance</td>
</tr>
<tr>
<td>Authorized Access Point Representing the Work (RDA 6.27.1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Title Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instance Title</td>
</tr>
<tr>
<td>Title Proper (RDA 2.3.2) (BIBFRAME: Main title)</td>
</tr>
<tr>
<td>Title Proper (RDA 2.3.2) (BIBFRAME: Main title)</td>
</tr>
<tr>
<td>Sorry / Why Remind Me</td>
</tr>
<tr>
<td>Other Title Information (RDA 2.3.4) (BIBFRAME: Subtitle)</td>
</tr>
<tr>
<td>Part number</td>
</tr>
<tr>
<td>Part name</td>
</tr>
<tr>
<td>Note on title</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genre Form (BIBFRAME: Genre Form)</td>
</tr>
<tr>
<td>Genre Form (BIBFRAME: Genre Form)</td>
</tr>
<tr>
<td>Jazz</td>
</tr>
<tr>
<td>Base Material (BIBFRAME: Base Material)</td>
</tr>
<tr>
<td>Base Material (BIBFRAME: Base Material)</td>
</tr>
<tr>
<td>Vinyl</td>
</tr>
<tr>
<td>Carrier (BIBFRAME: Carrier)</td>
</tr>
<tr>
<td>Carrier (BIBFRAME: Carrier)</td>
</tr>
<tr>
<td>audio disc</td>
</tr>
</tbody>
</table>
Lots of Linked Data Authorities
Direct Access

- AGROVOC
- DBPEDIA_DIRECT *(term fetch only)*
- GEONAMES_DIRECT
- LOC_DIRECT *(term fetch only)*
- NALT_DIRECT *(term fetch only)*
- OCLCFAST_DIRECT

Exploring

- Wikidata *(APIs are not linked data)*
- ISNI *(not available as linked data... yet)*

LD4P Cache

- AGROVOC_LD4L_CACHE
- DBPEDIA_LD4L_CACHE
- GEONAMES_LD4L_CACHE
- Getty
  - GETTY_AAT_LD4L_CACHE
  - GETTY_TGN_LD4L_CACHE
  - GETTY_ULAN_LD4L_CACHE
- Library of Congress
  - LOCDEMOGRAPHICS_LD4L_CACHE
  - LOCGENRES_LD4L_CACHE
  - LOCNAMES_LD4L_CACHE
  - LOCPERFORMANCE_LD4L_CACHE
  - LOCSUBJECTS_LD4L_CACHE
- MESH_LD4L_CACHE
- NALT_LD4L_CACHE
- OCLCFAST_LD4L_CACHE
- RDA Registry
- Share-VDE data for 20 institutions

A few more authorities available through QA that access authority APIs that are not supporting linked data.
Architecture

Application - Sinopia

Normalization - Lookup Service

Jena-Fuseki-Lucene Cache

Direct Access of External Authority
Why use a cache instead of direct access?

- Limited access by original data provider
  - Authority only provides a download of RDF
  - Authority supports reconcilable URI for term fetch, but no search API

- Search results do not include rank predicate

- Provide access to extended content in search results

- Control over indexing approach to improve accuracy of searches

- Control over when data changes
Challenges of caching data

- Equipment required to store the data locally
- Expertise in triplestores and scalability issues
- Synchronizing data updates
- Difficulty in providing real time updates
  - Original source doesn’t provide access to data in real time
Cache Loading Process

Original Source Data
e.g. LCGFT

download

file
e.g. ntriples, rdfxml, jsonld

ingest via loader scripts

Jena-Fuseki

ingest via indexer script via SPARQL queries

Lucene Index
Current Services Stack

- Apache
- Tomcat

- `_AUTHNAME_batch.jsp`
  (search for query string)

- `_AUTHNAME_lookup.jsp`
  (fetch single term)

- Lucene Index

- Jena-Fuseki
Cache API Access – **Fetch One**

**Normalization - Lookup Service**

- CURL request to `cache lookup jsp`
- RDF serialization

---

**Apache**

- `_AUTHNAME_batch.jsp`
  - (search for query string)

---

**Tomcat**

- `_AUTHNAME_lookup.jsp`
  - (fetch single term)

---

- request a single term by URI
- RDF serialization

---

**Lucene Index**

---

**Jena-Fuseki**
Cache API Access – **Search for multiple** (part 1)

Normalization - Lookup Service

- CURL request to `cache_batch.jsp`

- **_AUTHNAME_batch.jsp** (search for query string)
  - perform lucene search
  - matching results

- **_AUTHNAME_lookup.jsp** (fetch single term)
  - matching results

- **Lucene Index**

- **Jena-Fuseki**

- **Apache**

- **Tomcat**
Cache API Access – Search for multiple (part 2)

Normalization - Lookup Service

Apache
Tomcat

_AUTHNAME_batch.jsp
(search for query string)
- submit SPARQL query to get exact match (mark as rank #1)
- submit SPARQL to get basic & extended context using URIs from lucene results
- inject rank triples based on order of results from lucene search
- consolidate RDF to return

_AUTHNAME_lookup.jsp
(fetch single term)

Jena-Fuseki

RDF serialization

SPARQL queries
A Simple Lookup Request

```xml
<sparql:query var="result" endpoint="${ld4l}" resultType="triple">
  SELECT DISTINCT ?p ?o WHERE {
    ?s ?p ?o
  }
  <c:if test="${not empty param.lang}">
    FILTER(!isLiteral(?o) || lang(?o) = "" || langMatches(lang(?o), "${param.lang}"))
  </c:if>
</sparql:query>

<c:forEach items="${result.rows}" var="row" varStatus="rowCounter">
  <$param.uri> ${row.p} ${row.o} .
</c:forEach>
```
A Typical Complex Lookup Request

```sparql
<sparql:construct var="graph" endpoint="${ld4l}">
  OPTIONAL {
    ?o ?q ?r
    FILTER (isBlank(?o))
    <c:if test="${not empty param.lang}">
      FILTER(!isLiteral(?r) || lang(?r) = "" || langMatches(lang(?r), "${param.lang}"))
    </c:if>
  }
  OPTIONAL {
    ?t gn:name ?u
  }
  OPTIONAL {
    ?s gn:parentADM1 ?v .
    ?v gn:name ?w
  }
  OPTIONAL {
    ?s gn:parentADM2 ?x .
    ?x gn:name ?y
  }
</sparql:construct>
```
A Typical Complex Lookup Request, Part 2

<sparql:query var="result" graph="${graph}" resultType="triple">
  SELECT ?s ?p ?o WHERE {
  } ORDER BY ?s ?p
</sparql:query>

<c:forEach items="${result.rows}" var="row" varStatus="rowCounter">
  ${row.s} ${row.p} ${row.o} .
</c:forEach>
Tuning “Knobs”

- At the Lucene index level (i.e. matching entities and the order in which they are returned)
  - Tokenization of input query
  - Relative weighting of terms
  - Semantics of query (disjunction vs. conjunction, etc.)

- At the Fuseki endpoint level (i.e., the shape of the response graph for each matching URI)
  - SPARQL query (or queries) can be extended as needed for complex contextual requirements
A User Interface for the Lookup Service
## Authority List

See [Check Status](#) to test whether an authority is online.

<table>
<thead>
<tr>
<th>Authority/Subauthority</th>
<th>Service</th>
<th>Action</th>
<th>Sample URL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ld4l_cache</td>
<td>search</td>
<td>/authorities/search/linked_data/agrovoc_ld4l_cache?q=milk&amp;maxRecords=4</td>
</tr>
<tr>
<td></td>
<td>ld4l_cache</td>
<td>search</td>
<td>/authorities/search/linked_data/dbpedia_ld4l_cache?q=Barack+Obama&amp;maxRecords=4</td>
</tr>
</tbody>
</table>
# Authority Lookup Service

## Check Status

- Select authority...
- Check Connection Status only
- Check Accuracy only
- Run all checks

## Connection Checks

<table>
<thead>
<tr>
<th>Status</th>
<th>Subauthority</th>
<th>Service</th>
<th>Action</th>
<th>Sample URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>✅</td>
<td>direct</td>
<td>search</td>
<td></td>
<td><code>/authorities/search/linked_data/geonames_direct?q=lthaca&amp;maxRecords=4</code></td>
</tr>
</tbody>
</table>
Performance Graph

monthly data

all requests
Performance Graph

day data

all requests
Performance Graph

year data

all requests
Performance Graph

monthly data

search requests
Performance Graph

monthly data

fetch requests
Performance Graph

Identifying poorly performing authorities
Real Code! (Just add data…)

To set up a normalization lookup service...

- qa_server - [https://github.com/ld4p/qa_server](https://github.com/ld4p/qa_server) (start here)
- QA - [https://github.com/samvera/questioning_authority](https://github.com/samvera/questioning_authority) (documentation)
- linked data authorities - [https://github.com/LD4P/linked_dataAuthorities](https://github.com/LD4P/linked_dataAuthorities)

To set up a cache system...

- ld4l_services web app - [https://github.com/eichmann/ld4l_services](https://github.com/eichmann/ld4l_services)
- Lucene Tag Library - [https://github.com/eichmann/LuceneTagLib](https://github.com/eichmann/LuceneTagLib)
- SPARQL Tag Library - [https://github.com/eichmann/SPARQLTagLib](https://github.com/eichmann/SPARQLTagLib)
Try out our server:  https://lookup.ld4l.org

Exploration only. Not for production use beyond LD4P project.