

Efficient and Reliable Access to Authorities with Contextual Lookup

E. Lynette Rayle
Cornell University

and

Dave Eichmann
University of Iowa



Benefits of Working with External Authorities Data

- Controlled Vocabulary
- Data integrity
- Shared concepts across many institutions
- URI providing an exact match and disambiguation

List of Authorities

- LoC – Library of Congress
 - Name Authority (personal, corporate, etc.)
 - Subjects
 - Genres
- OCLC FAST
- GeoNames
- AGROVOC – agricultural keywords
- NALT – agricultural keywords thesaurus
- DBpedia

First Set of Challenges

1. Finding Documentation

2. Linked Data Access API

e.g. no support, partial support, requires login credentials, sparql query endpoint only

3. Varying Results Formats

e.g. rdf-xml, json-ld, turtle, n-triples, etc.

4. Varying Ontologies

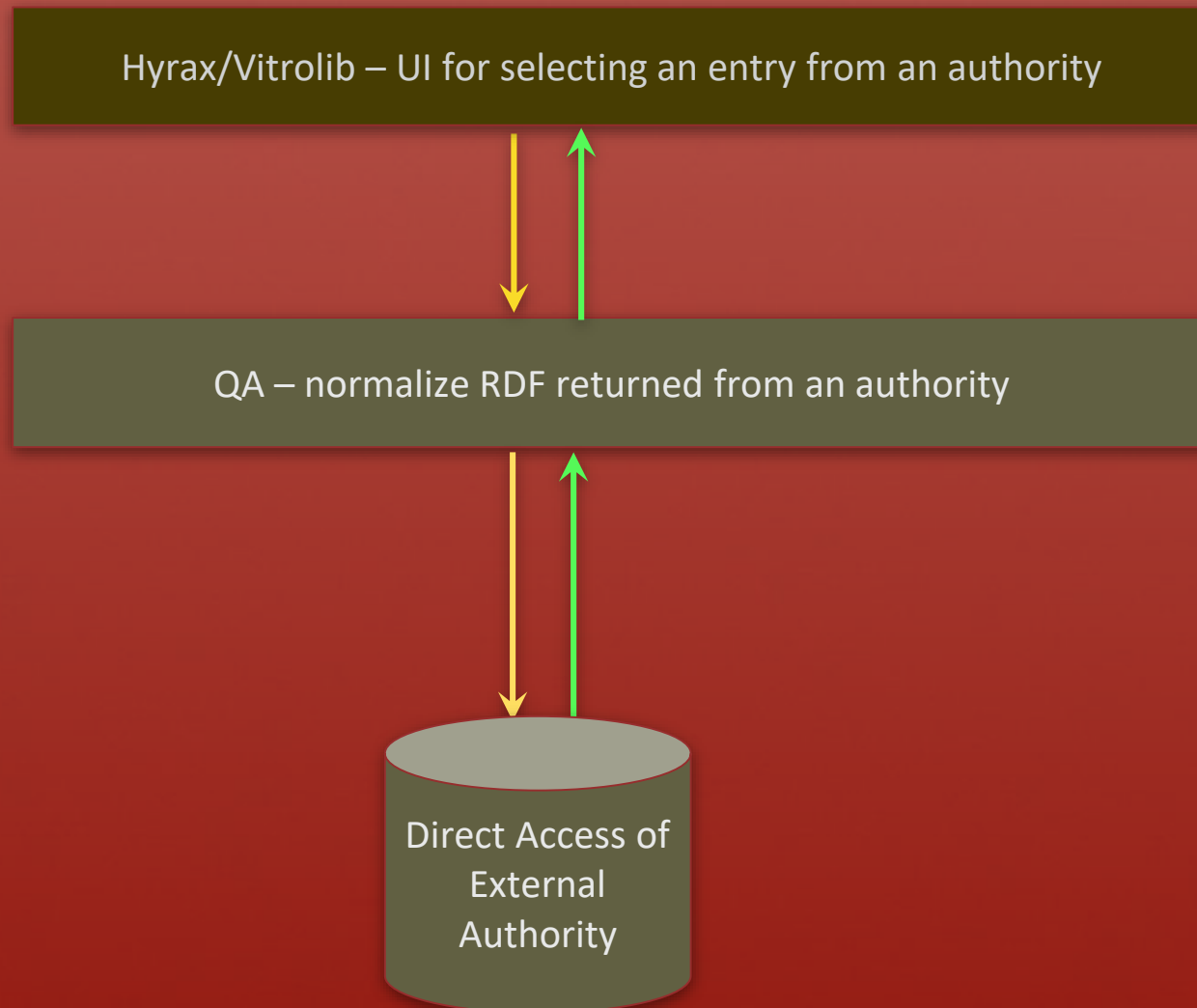
e.g. SKOS, schema.org, madsrdf, dbpedia, geonames

- complexity and richness of the data varies

Multi-Server Architecture

QA – normalize RDF returned from an authority

Multi-Server Architecture



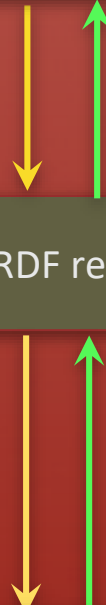
Multi-Server Architecture

Hyrax/Vitrolib – UI for selecting an entry from an authority

`http://localhost:3000/qa/search/linked_data/
oclc_fast/personal_name?q=twain&
maximumRecords=2`

QA – normalize RDF returned from an authority

Direct Access of
External
Authority



Multi-Server Architecture

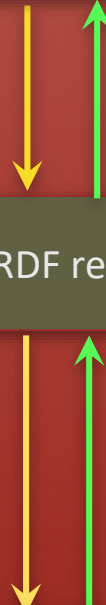
Hyrax/Vitrolib – UI for selecting an entry from an authority

`http://localhost:3000/qa/search/linked_data/
oclc_fast/personal_name?q=twain&
maximumRecords=2`

QA – normalize RDF returned from an authority

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

Direct Access of
External
Authority



Multi-Server Architecture

Hyrax/Vitrolib – UI for selecting an entry from an authority

`http://localhost:3000/qa/search/linked_data/
oclc_fast/personal_name?q=twain&
maximumRecords=2`

QA – normalize RDF returned from an authority

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

`<http://id.worldcat.org/fast/31622>
 a schema:Person"
 dcterms:identifier 31622;
 skos:prefLabel "Twain, Mark, 1835-1910"
 skos:altLabel "Make Teviin, 1835-1910",
 "Make Tuwen, 1835-1910",
 ...;`

`<http://id.worldcat.org/fast/365563>
 a schema:Person"
 dcterms:identifier 365563;
 skos:prefLabel "Twain, Shania";
 skos:altLabel "Twain, Eilleen",
 "Edwards, Eilleen";`

Direct Access of
External
Authority

Multi-Server Architecture

Hyrax/Vitrolib – UI for selecting an entry from an authority

`http://localhost:3000/qa/search/linked_data/
oclc_fast/personal_name?q=twain&
maximumRecords=2`

```
[{"uri": "http://id.worldcat.org/fast/31622",  
  "id": "31622", "label": "Twain, Mark, 1835-1910"},  
 {"uri": "http://id.worldcat.org/fast/365563",  
  "id": "365563", "label": "Twain, Shania"} ... ]
```

QA – normalize RDF returned from an authority

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

```
<http://id.worldcat.org/fast/31622>  
  a schema:Person"  
  dcterms:identifier 31622;  
  skos:prefLabel "Twain, Mark, 1835-1910" ;  
  skos:altLabel "Make Teviin, 1835-1910",  
                "Make Tuwen, 1835-1910",  
                ...;
```

Direct Access of
External
Authority

```
<http://id.worldcat.org/fast/365563>  
  a schema:Person"  
  dcterms:identifier 365563;  
  skos:prefLabel "Twain, Shania";  
  skos:altLabel "Twain, Eilleen",  
                "Edwards, Eilleen";
```

Direct Access Query API

Direct against authority...

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

```
http://api.geonames.org/search?q=ithaca&maxRows=2  
&username=demo&type=rdf
```

```
http://artemide.art.uniroma2.it:8081/agrovoc/rest/v1/search/  
?query=*milk*&lang=en&maxhits=2
```

Normalized Query API

Through QA normalization layer...

```
http://localhost:3000/qa/search/linked_data/oclc_fast?  
q=twain&maxRecords=2
```

```
http://localhost:3000/qa/search/linked_data/geonames?  
q=ithaca&maxRecords=2
```

```
http://localhost:3000/qa/search/linked_data/agrovoc?  
q=milk&maxRecords=2&lang=en
```

Normalized Results

```
[{"uri": "http://id.worldcat.org/fast/31622",  
  "id": "31622",  
  "label": "Twain, Mark, 1835-1910"},  
{"uri": "http://id.worldcat.org/fast/365563",  
  "id": "365563",  
  "label": "Twain, Shania"}]
```

```
[{"uri": "http://sws.geonames.org/2162552/",  
  "id": "http://sws.geonames.org/2162552/",  
  "label": "Ithaca (AU)"},  
{"uri": "http://sws.geonames.org/4515289/",  
  "id": "http://sws.geonames.org/4515289/",  
  "label": "Ithaca (US)"}]
```

```
[{"uri": "http://aims.fao.org/aos/agrovoc/c_8602",  
  "id": "http://aims.fao.org/aos/agrovoc/c_8602",  
  "label": "acidophilus milk"},  
{"uri": "http://aims.fao.org/aos/agrovoc/c_16076",  
  "id": "http://aims.fao.org/aos/agrovoc/c_16076",  
  "label": "buffalo milk"}]
```

Autocomplete Saving String and URI

Authority: OCLC FAST

Subauthority: PersonName

Oclc organization required

Cornell University

+ Add another Oclc organization

Oclc person required

ezra

Benson, Ezra Taft

Cornell, Ezra, 1807-1874

Ezra (Biblical figure)

Gannett, Ezra S. (Ezra Stiles), 1801-1871

Ibn Ezra, Abraham ben Meïr, 1089-1164

Ibn Ezra, Moses, approximately 1060-approximately 1139

Meeker, Ezra, 1830-1928

Pound, Ezra, 1885-1972

Sims, Ezra, 1928-2015

Stiles, Ezra, 1727-1795

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 specific access in the "Share With" section.

☐ I have read and agree to the [Deposit Agreement](#)

Cancel

Save

Selected String and URI

Saves both string and URI

<div>Cornell University</div> <div>+ Add another Oclc organization</div>	<div>Requirements</div> <div><input checked="" type="checkbox"/> Describe your work</div>
<div>Oclc person required</div> <div>Cornell, Ezra, 1807-1874</div> <div>http://id.worldcat.org/fast/409667</div> <div>+ Add another Oclc person</div>	<div>Visibility</div> <div><input type="radio"/> 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.</div> <div><input type="radio"/> Institution Restrict access to only users and/or groups from Institution</div> <div><input type="radio"/> Embargo</div> <div><input type="radio"/> Lease</div> <div><input checked="" type="radio"/> Private Only users and/or groups that have been given specific access in the "Share With" section</div>
<div>Agrovoc keyword required</div> <div>hive</div> <div>Allium tuberosum</div> <div>archives</div> <div>chives</div> <div>hive equipment</div> <div>hive frames</div> <div>hive management</div> <div>hive products</div> <div>hives</div>	<div><input checked="" type="checkbox"/> I have read and agree to the Deposit Agreement</div> <div><div>Cancel</div><div>Save</div></div>

Access and Normalization Layer

Questioning Authority

- ruby gem that can be deployed as a standalone server

([ld4l-labs/ga_server](https://github.com/ld4l-labs/ga_server))

Challenge 2
Access API



- configurable access to authorities providing a single api for end applications ([ld4l-labs/linked_data_authorities](https://github.com/ld4l-labs/linked_data_authorities))

Challenge 3
Results Format



- uses rdf/linkedException gem for processing any number of formats (e.g. jsonld, rdf-xml, ttl, nt, etc.)
- normalizes results to simplify end applications UI

Challenge 4
Ontologies



Second Set of Challenges

5. Reliability & Efficiency

e.g. server uptime, server load

6. Accuracy

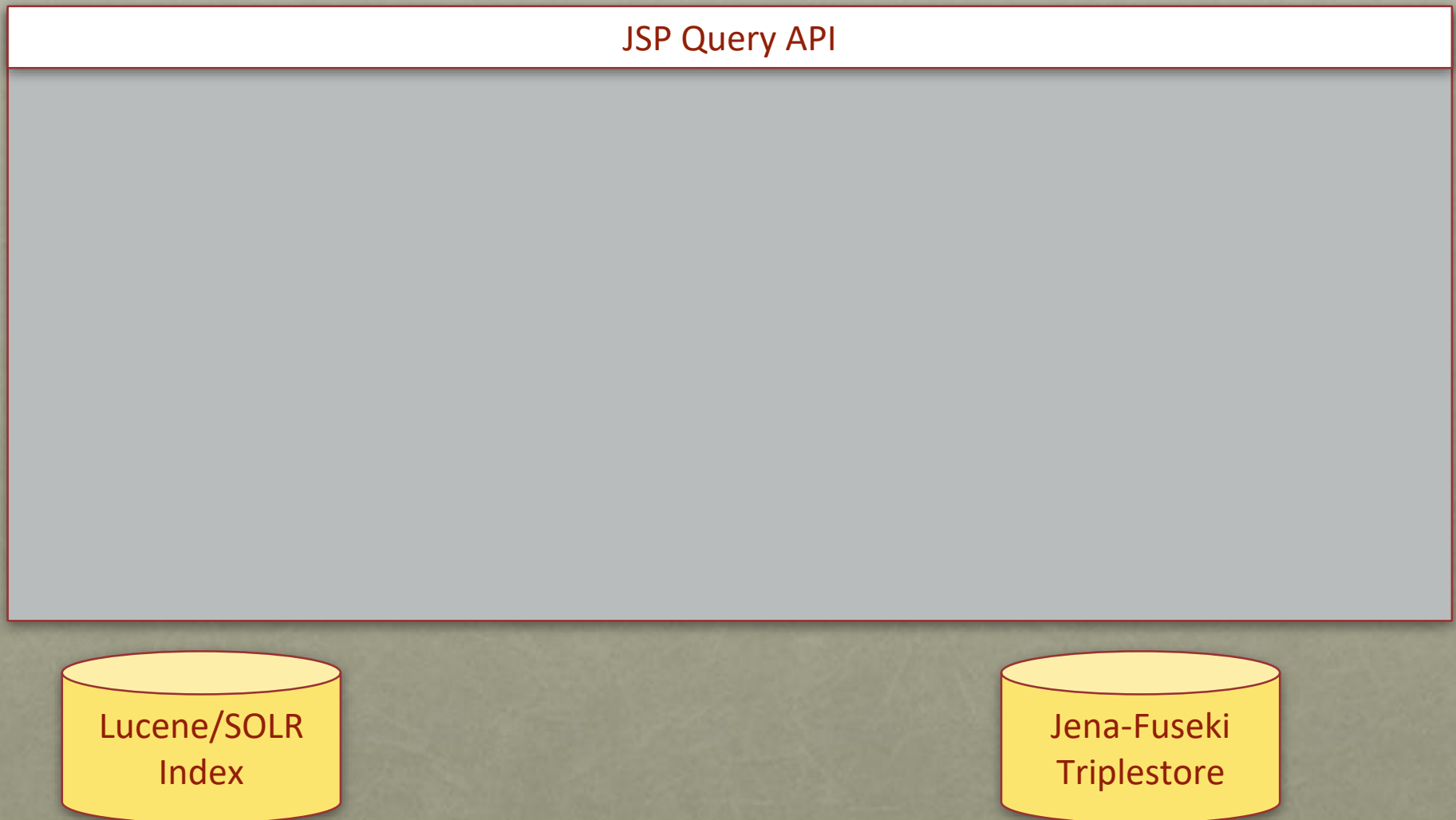
e.g. select results based on usage data, lexical match, custom weighting, other?

7. Order Ranking

e.g. How to order a graph?

Cache Server Query Process

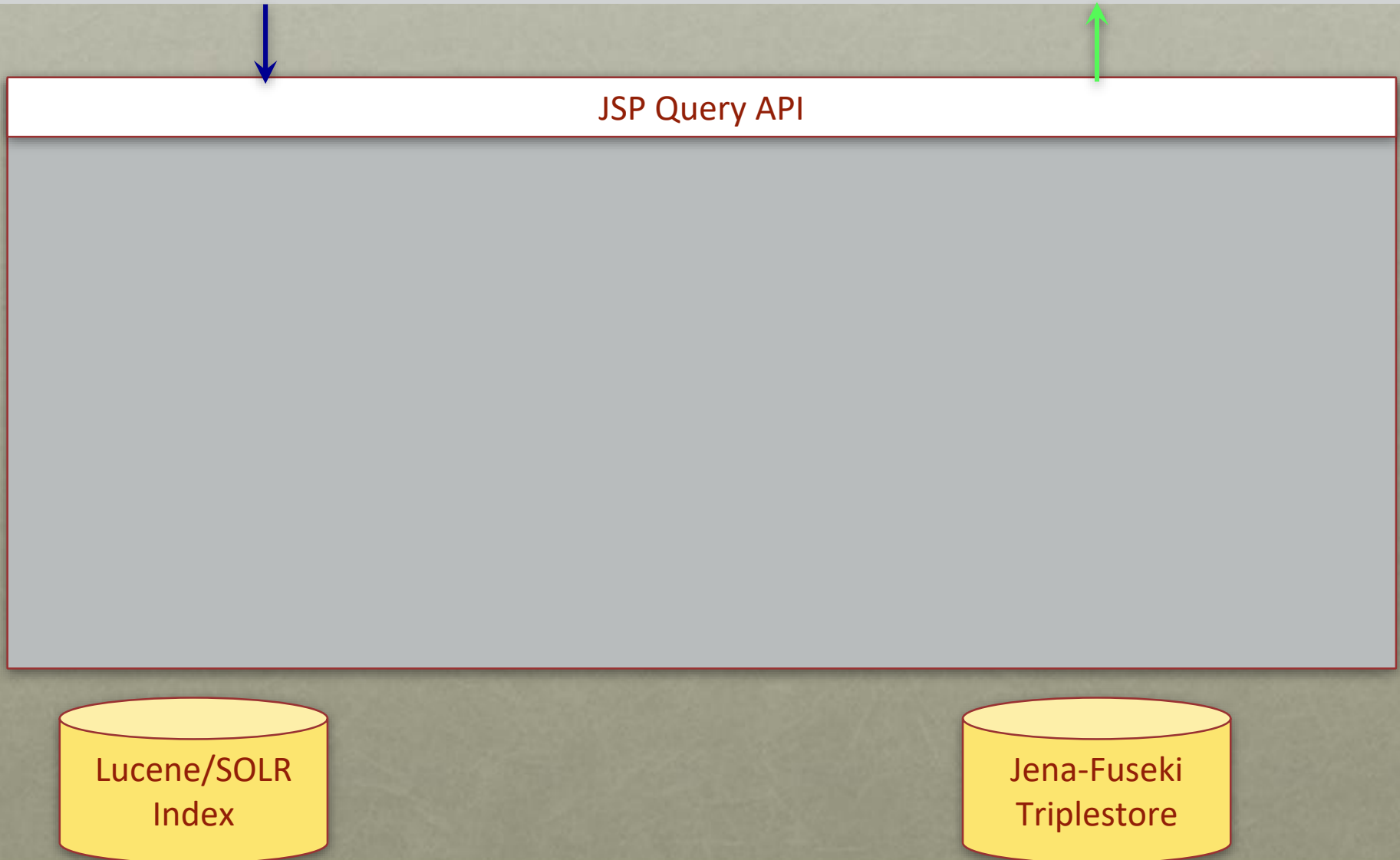
One full setup per authority



Cache Server Query Process

One full setup per authority

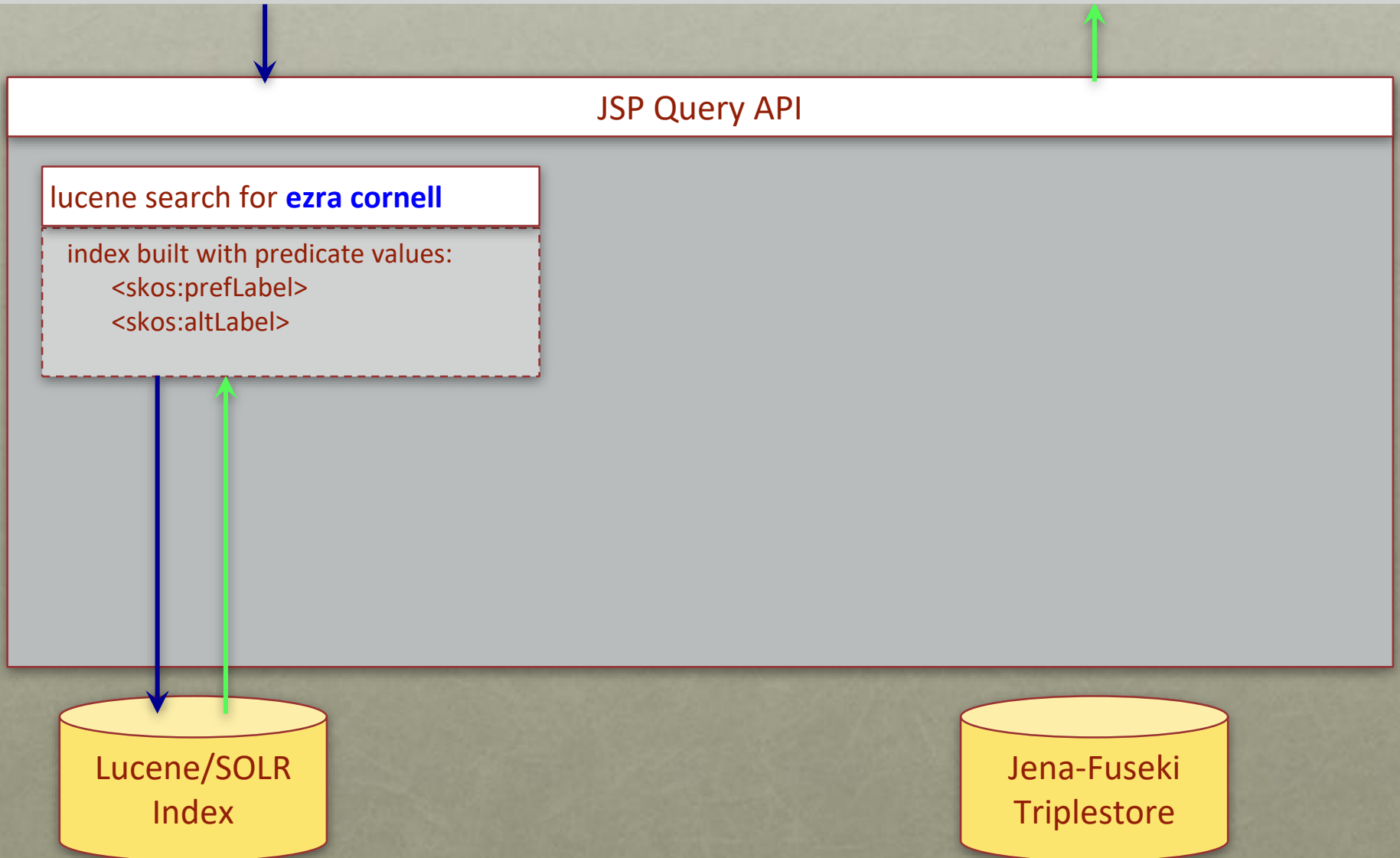
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



Cache Server Query Process

One full setup per authority

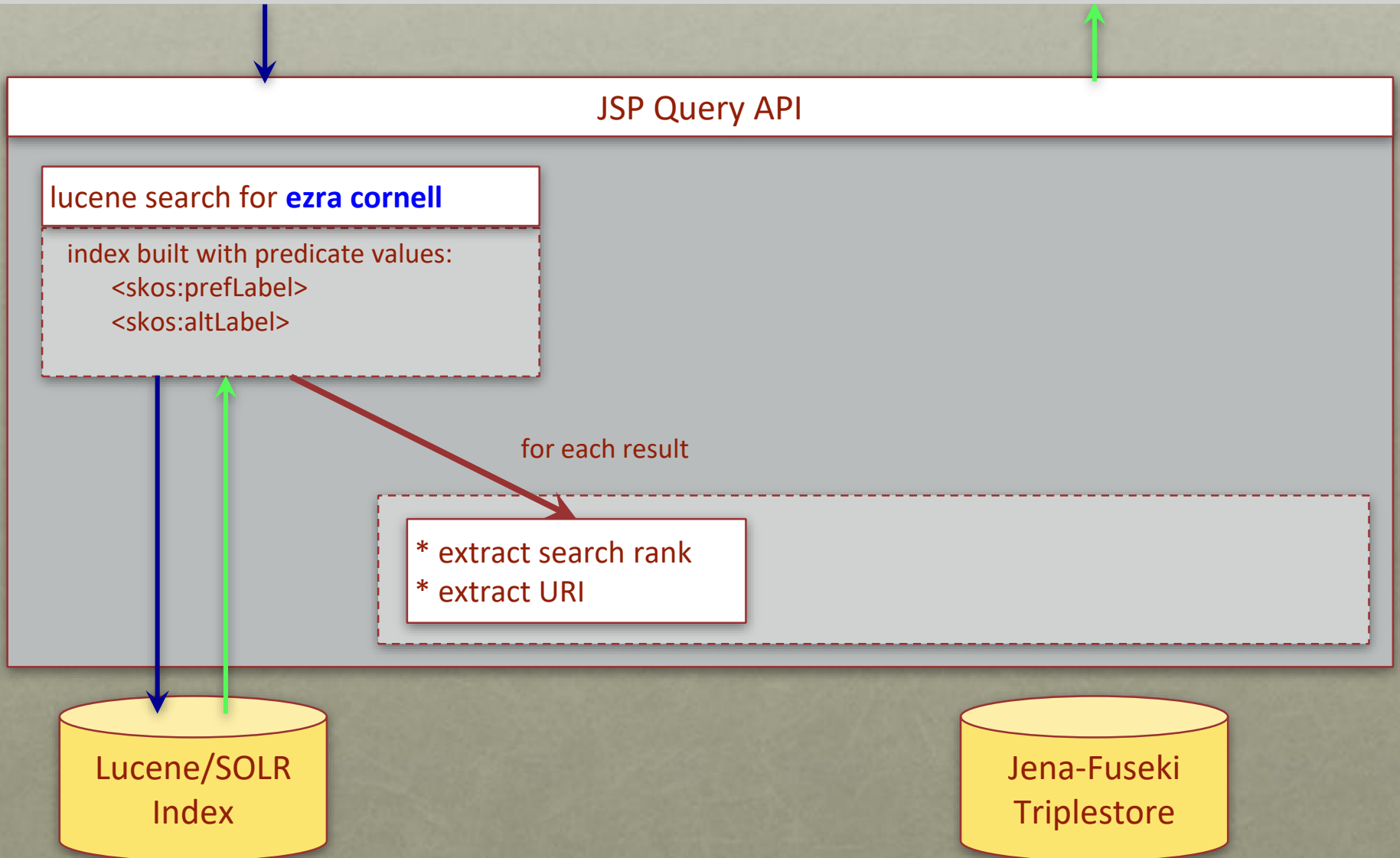
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



Cache Server Query Process

One full setup per authority

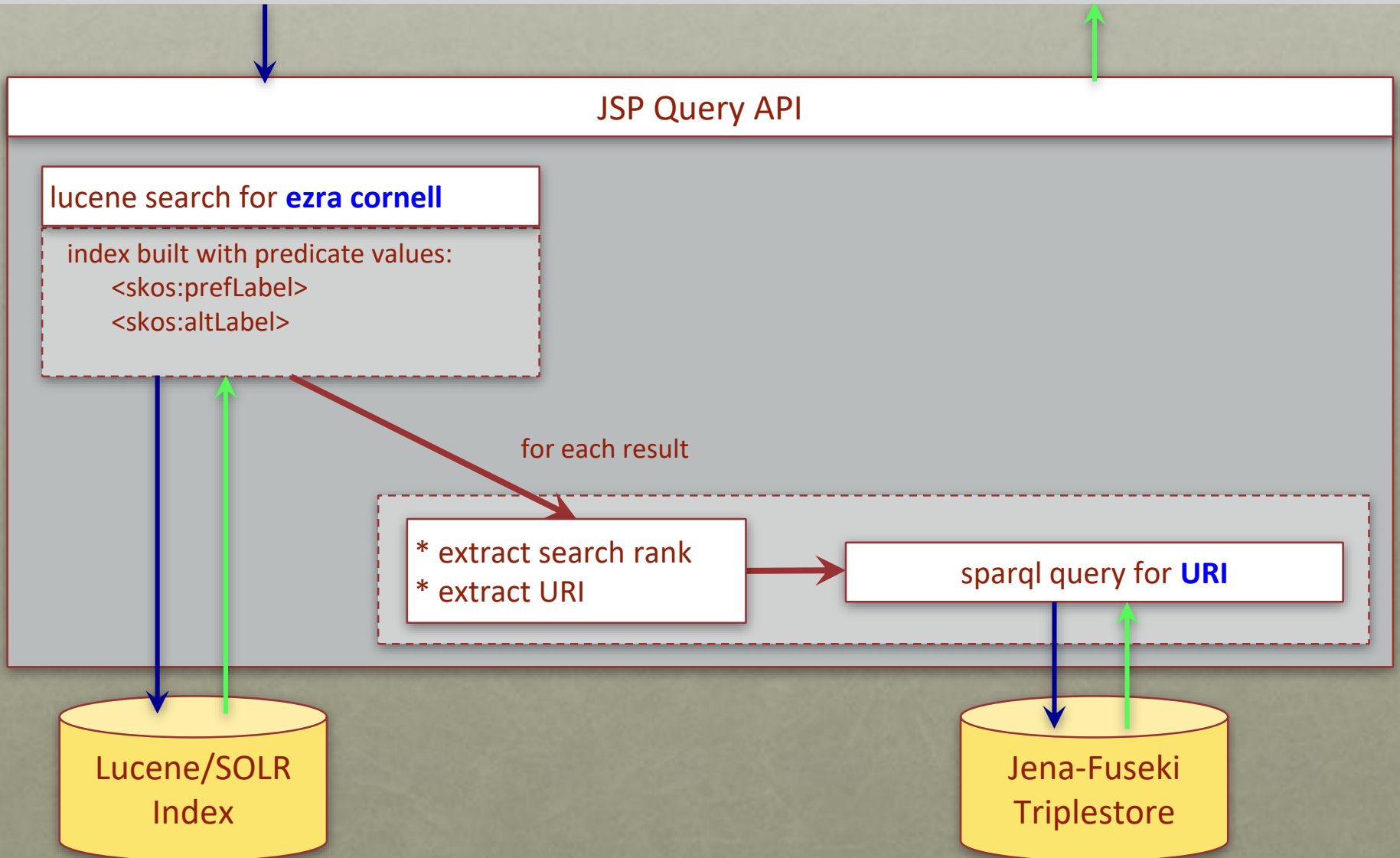
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



Cache Server Query Process

One full setup per authority

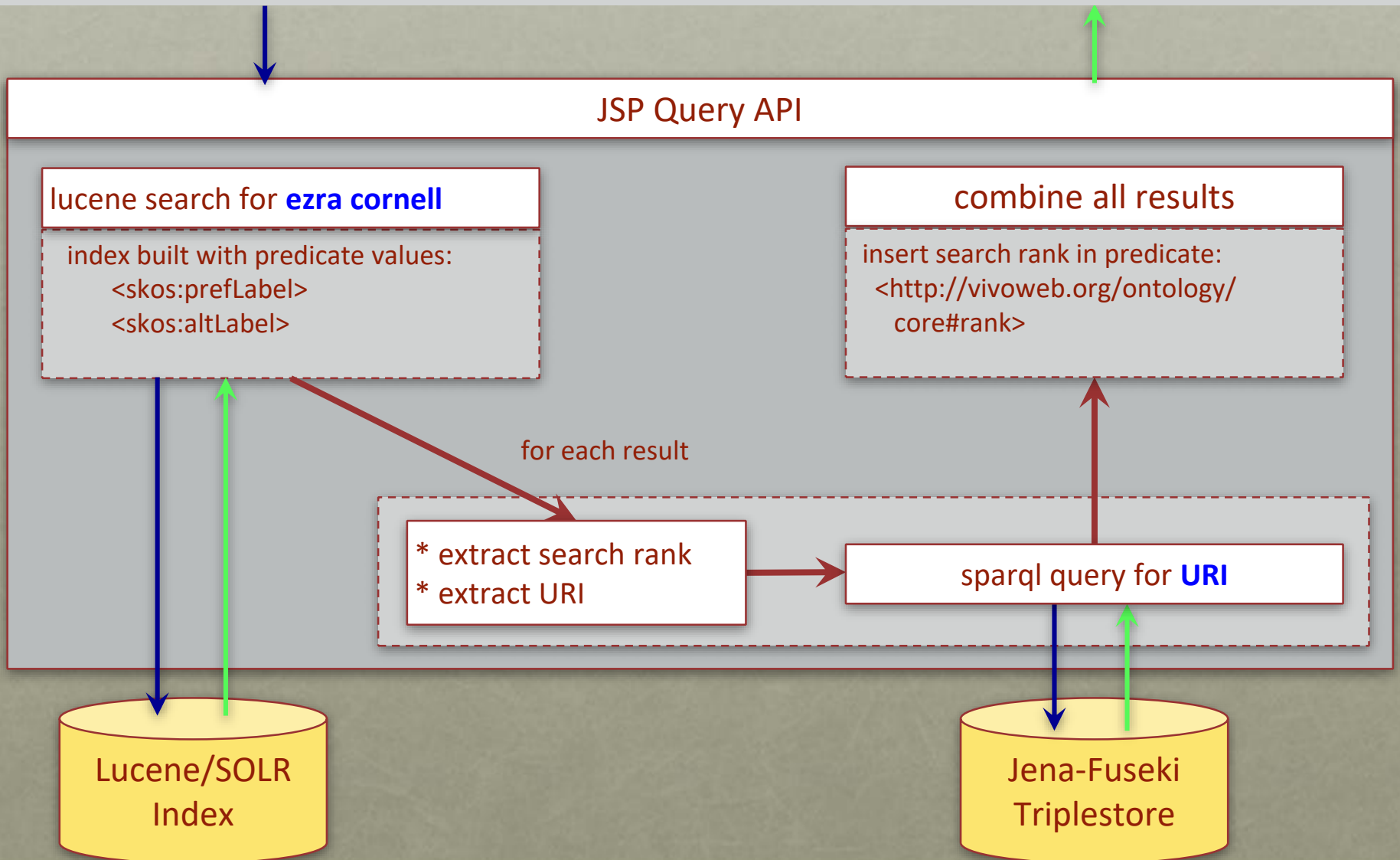
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



Cache Server Query Process

One full setup per authority

http://services.id4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



UI-QA-Authority

Author *
twain,mark
Clemens, Olivia Langdon, 1845-1904
Twain, Mark, 1835-1910 (Spirit)
Twain, Mark, 1835-1910. Plymouth Rock and the Pilgrims and other salutary platform opinions
Twain, Mark, 1835-1910. Private history of a campaign that failed
Twain, Mark, 1835-1910. Eve's diary
Twain, Mark, 1835-1910. Prince and the pauper
Twain, Mark, 1835-1910. Mark Twain papers
Twain, Mark, 1835-1910. Man that corrupted Hadleyburg
Twain, Mark, 1835-1910. Mark Twain library
Twain, Mark, 1835-1910. Works. 2003
Twain, Mark, 1835-1910. Mysterious stranger and other stories
Twain, Mark, 1835-1910. Murder, a mystery and a marriage
Twain, Mark, 1835-1910
Twain, Mark, 1835-1910. Tom Sawyer abroad

Hyrax/Vitrolib – UI for selecting an entry from an authority

http://localhost:3000/qa/search/linked_data/oclc_fast/personal_name?q=twain&maximumRecords=2

```
[{"uri": "http://id.worldcat.org/fast/31622", "id": "31622",  
  "label": "Twain, Mark, 1835-1910"},  
 {"uri": "http://id.worldcat.org/fast/365563", "id": "365563",  
  "label": "Twain, Shania"}]
```

QA – normalize RDF returned from an authority

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

RDF of
search
results

Direct Access of
External
Authority

Jena-Fuseki-
Lucene Cache*

LDF Cache

Active-Triples
LDF Cache
(Marmotta or
Blazegraph)

* search of cache performed via Lucene index

Cached Linked Data

Challenge 5
Reliability &
Efficiency



- controlled server allows for control of uptime, throughput, and speed

Challenge 6
Accuracy



- lucene indexing and pre-determined sparql queries provide accuracy of search results

Challenge 7
Order of Results



- addition of search rank predicate provides ability to sort rdf graph search results for consistent presentation to the user

Third Set of Challenges

8. Disambiguation through better context

e.g. expand from just prefLabel to preLabel, altLabel, birth/death dates, occupation, etc.

9. Reconciliation across multiple sources

e.g. match LoC URI to OCLC FAST URI

Mockups

<https://wiki.duraspace.org/x/mxUZBQ>

Selecting a Term using Lookup with Context

[illegible]

Selecting a Term using Lookup with Context

Agrovoc Lookup

http://localhost:3000/qa/lookup/linked_data/agrovoc

3 Cancel 4 Save

2 Authority: Agrovoc
Cornell Local

1 milk

7 Not found? Add to Local

Click one or more checkboxes to select...

Broader	Found	Narrower	Same As
<p>5</p> <input type="checkbox"/> animal products	<p>6</p> <input type="checkbox"/> milk	<input type="checkbox"/> buffalo milk <input type="checkbox"/> camel milk <input type="checkbox"/> colostrum <input type="checkbox"/> cow milk <input type="checkbox"/> ewe milk <input type="checkbox"/> goat milk	http://aims.fao.org/aos/asfa/c_7251 http://cat.ii.caas.cn/concept/c_32204 http://cat.ii.caas.cn/concept/c_33165 http://d-nb.info/gnd/4039264-8 http://eurovoc.europa.eu/1565 http://linkeddata.ge.imati.cnrit.2020/resource/EARTH/55910 http://lod.nal.usda.gov/nalt/631 http://www.eionet.europa.eu/gemet/concept/5254 http://zbw.eu/stw/descriptor/14124-5
<input type="checkbox"/> by-products <input type="checkbox"/> collection <input type="checkbox"/> containers	<input type="checkbox"/> milk by-products <input type="checkbox"/> milk collection <input type="checkbox"/> milk containers	<input checked="" type="checkbox"/> milk products	
<input checked="" type="checkbox"/> animal fats <input type="checkbox"/> proteases	<input type="checkbox"/> milk fats <input checked="" type="checkbox"/> chymosin		http://cat.ii.caas.cn/concept/c_38198 http://d-nb.info/gnd/4169913-0 http://eurovoc.europa.eu/1836 http://lod.nal.usda.gov/nalt/9160 http://stitch.cs.vu.nl/vocabularies/rameau/ark:/12148/cb11980632s http://cat.ii.caas.cn/concept/c_33063 http://cat.ii.caas.cn/concept/c_58677 http://lod.nal.usda.gov/nalt/13041

Cached Linked Data

Challenge 8 Disambiguation



- providing additional context at selection time supports metadata specialists ability to select the most accurate term

Challenge 9 Reconciliation




- reconciliation of terms from multiple authorities

Leveraging Linked Data

Getting more from the same authority?

Agrovoc Keyword


hive products  http://aims.fao.org/aos/agrovoc/c_3655

narrower: propolis  http://aims.fao.org/aos/agrovoc/c_15919

narrower: royal jelly  http://aims.fao.org/aos/agrovoc/c_26817

narrower: honey  http://aims.fao.org/aos/agrovoc/c_3652

narrower: beeswax  http://aims.fao.org/aos/agrovoc/c_866

narrower: honeycomb extracts  http://aims.fao.org/aos/agrovoc/c_29026

broader: animal products  http://aims.fao.org/aos/agrovoc/c_438

sameas:  <http://d-nb.info/gnd/4006529-7>

sameas:  http://cat.aii.caas.cn/concept/c_12939

Getting more from other authorities?

Oclc person [Cornell, Ezra, 1807-1874](#)

(source:  http://dbpedia.org/resource/Ezra_Cornell)

Birth: 1807-01-11

Death: 1874-12-09

Ezra Cornell (January 11, 1807 – December 9, 1874) was an American businessman, politician, philanthropist and educational administrator. He was the founder of Western Union and a co-founder of Cornell University. He also served as President of the New York Agriculture Society and as a New York state Senator.

Oclc person  <http://id.worldcat.org/fast/409667>
uri

Questions?

Appendix for Challenges 1-4

Challenge 1: Documentation

LoC	http://id.loc.gov/techcenter/ C. Harlow notes on reconciling LoC - https://github.com/cmh2166/lc-reconcile
OCLC FAST	https://www.oclc.org/developer/develop/web-services/fast-api/linked-data.en.html
GeoNames	http://www.geonames.org/export/geonames-search.html
AGROVOC	http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus swagger config: https://github.com/NatLibFi/Skosmos/blob/master/swagger.json
NALT	https://agclass.nal.usda.gov/
DBpedia	http://wiki.dbpedia.org/OnlineAccess#1.2%20Public%20Faceted%20Web%20Service%20Interface

Challenge 2: Linked Data Access API

	for Search Query	for Term Fetch
LoC	not supported	URI
OCLC FAST	http://experimental.worldcat.org/fast/search?query={?subauth}+all+%22{?query}%22&sortKeys=usage&maximumRecords={?maximumRecords}	URI
GeoNames	http://api.geonames.org/search?q={?query}&maxRows={?maxRows}&username={?username}&type=rdf	URI
AGROVOC	http://artemide.art.uniroma2.it:8081/agrovoc/rest/v1/search/?query={?query}&lang={?lang}	http://artemide.art.uniroma2.it:8081/agrovoc/rest/v1/data?uri=http://aims.fao.org/aos/agrovoc/{?term_id}
NALT	http://skosmos.library.cornell.edu/rest/v1/nalt/search/?query={?query}&lang={?lang}	http://skosmos.library.cornell.edu/rest/v1/nalt/data?uri={?term_uri}
DBpedia		

Challenge 3: Varying Results Formats

	for Search Query	for Term Fetch
LoC	not supported	rdf+xml
OCLC FAST	rdf+xml	rdf+xml
GeoNames	rdf+xml	rdf+xml
AGROVOC	json-ld	rdf+xml, json-ld, turtle
NALT	json-ld	rdf+xml, json-ld, turtle
DBpedia		

Challenge 4: Varying Ontologies

	Primary Ontology	Flat vs. Navigation required
LoC	madsrdf SKOS	navigation required
OCLC FAST	schema.org SKOS	flat
GeoNames	geonames	flat hierarchical
AGROVOC	SKOS	flat hierarchical
NALT	SKOS	flat hierarchical
DBpedia	dbpedia	flat

Configurations for Questioning Authority

See [LD4L Labs linked data authorities README](#) for documentation on how to use these configurations.

LoC	https://github.com/ld4l-labs/linked_data_authorities/tree/master/qa_loc/config/authorities/linked_data
OCLC FAST	https://github.com/ld4l-labs/linked_data_authorities/tree/master/qa_oclcfast/config/authorities/linked_data
GeoNames	https://github.com/ld4l-labs/linked_data_authorities/tree/master/qa_geonames/config/authorities/linked_data
AGROVOC	https://github.com/ld4l-labs/linked_data_authorities/tree/master/qa_agrovoc/config/authorities/linked_data
NALT	https://github.com/ld4l-labs/linked_data_authorities/tree/master/qa_nalt/config/authorities/linked_data
DBpedia	https://github.com/ld4l-labs/linked_data_authorities/tree/master/qa_dbpedia/config/authorities/linked_data

Appendix for Challenges 5-7

Creating a Cache Server

Hardware

- 8-core, 64gb 3Ghz Mac Pro (late 2013), macOS Sierra (10.12.6)
- 32tb Pegasus-2 Thunderbolt RAID configured as RAID-5

Triplestore

- Apache Jena Fuseki 2.4.0 provides SPARQL endpoint
- Apache Tomcat 9.0 runs custom web application(s)
- Apache Lucene 3.6 provides search interface

Customizations

- custom per-data-source JSP web application provides search/browse/download functionality
- custom (generic) SPARQL Tag Library provides API for web apps (available at <https://github.com/eichmann/lod-utilities>)
- custom (generic) Lucene Tag Library provides API for web apps

Loading a New Vocabulary

- download RDF
- if necessary, convert to n-triples (required for GeoNames data, for instance)
- use tdbloader2 to populated triplestore
- configure Fuseki server(s) with triplestore details
- create new JSP project in Eclipse
- write one or more indexer programs that populate Lucene indices and run indexer(s)
- write search/browse/download application logic using the SPARQL and Lucene tags
- package project as war
- deploy to Apache Tomcat server(s)
- add new service to Apache HTTPD virtual host specification

UI Access to Cache Server

http://services.ld4l.org/ld4l_services/loc_name.jsp

LD4L Reconciliation Services

Home

DBpedia

By Name

By Entity

By Person

Library of Congress

By Genre

By Name

By Subject

VIAF

By Entity

By Person

FAST

By Entity

Catalog

By Work

By Person

More BibLeo

Cornell Catalog

Harvard Catalog

LoC Search by Name

Ontology class?

PersonalName



Result Format?

☐ Return a list of triples

☒ Display as HTML table

Search Results: Ezra Cornell

Result Count: 770

- ⌘ [Cornell, Ezra, 1807-1874@EN](#)
- ⌘ [Brown, Ezra \(Ezra A.\)@EN](#)
- ⌘ [Ezra, Derek@EN](#)
- ⌘ [Cornell, Margaret@EN](#)
- ⌘ [Cornell, Bryan@EN](#)
- ⌘ [Cornell, Drucilla@EN](#)
- ⌘ [Cornell, Judith@EN](#)
- ⌘ [Cornell, Mimi@EN](#)
- ⌘ [Cornell, Tyson@EN](#)
- ⌘ [Cornell, Heather@EN](#)
- ⌘ [Cornell, Ross@EN](#)

Downloads

LoC	http://id.loc.gov/download/ (n-triples OR rdf-xml)
OCLC FAST	http://www.oclc.org/research/themes/data-science/fast/download.html (n-triples)
GeoNames	http://www.geonames.org/ontology/documentation.html (custom format – see notes for processing)
AGROVOC	https://aims-fao.atlassian.net/wiki/spaces/AGV/pages/2949126/Releases (n-triples OR rdf-xml)
NALT	https://agclass.nal.usda.gov/download.shtml (rdf-xml)
DBpedia	http://wiki.dbpedia.org/downloads-2016-04

Potential Options for Reconciliation

- [VIAF](#) for name reconciliation – we are doing some work with this
- [Europeana Pro](#) – I've heard that they are working on Reconciliation issues but haven't yet explored in depth
 - From the KB in the Netherlands: https://media.ccc.de/v/wikidatacon2017-10019-wikidata_as_un...
 - Jason Evans from the National Library of Wales: https://media.ccc.de/v/wikidatacon2017-10018-wikidata_loves...
 - Europeana of Wikidata for their application in their collections: https://media.ccc.de/v/wikidatacon2017-10016-structured_dat...

Doing a google search for 'linked data reconciliation' returns a large number of articles and presentations on this concept.