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

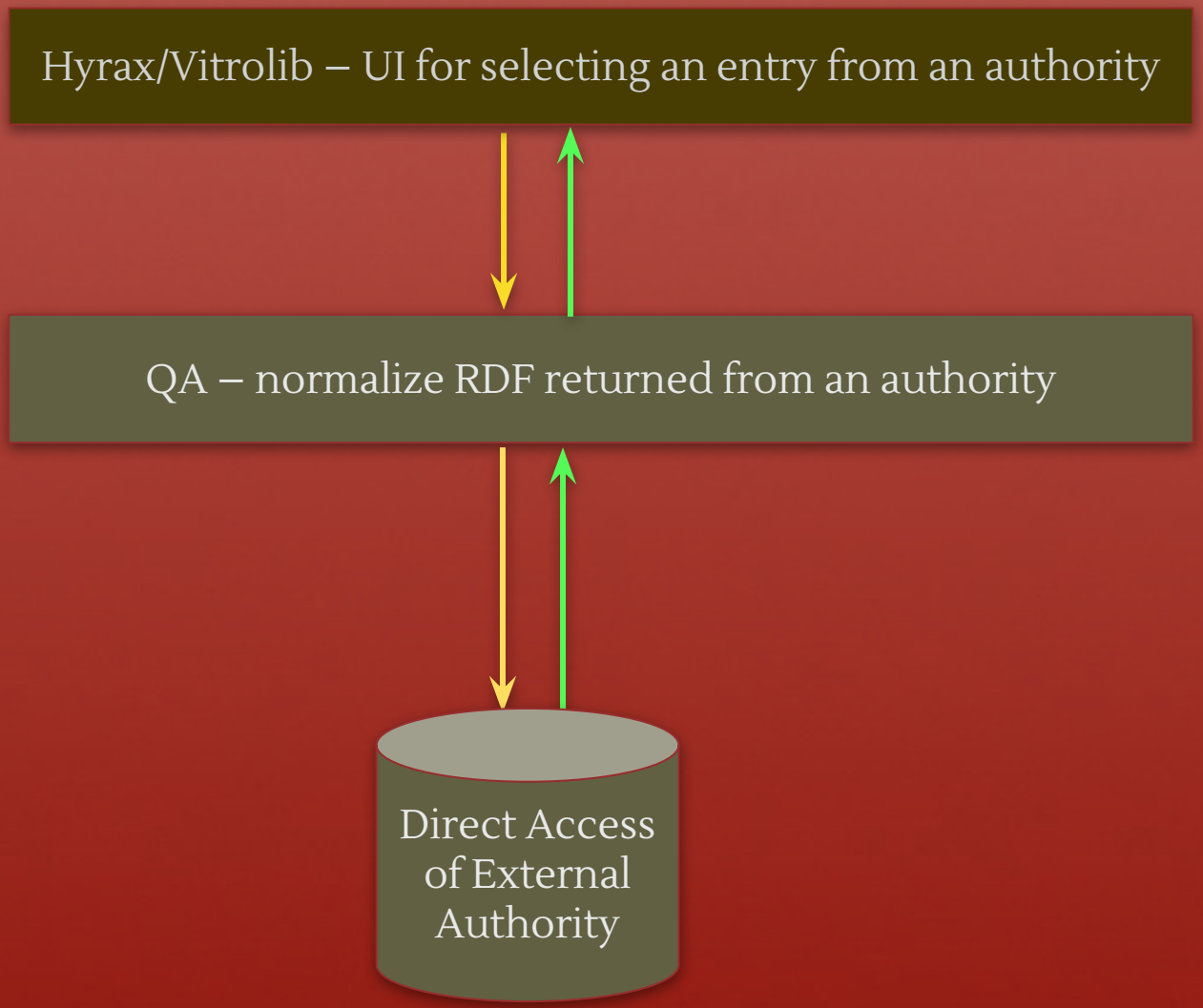
1. Finding Documentation
2. Linked Data Access API
   e.g. no support, partial support, requires login credentials,
   sparql query endpoint only
3. Varying Results Formats
   e.g. rdf-xml, json-ld, turtle, n-triples, etc.
4. Varying Ontologies
   e.g. SKOS, schema.org, madsrdf, dbpedia, geonames

   - complexity and richness of the data varies

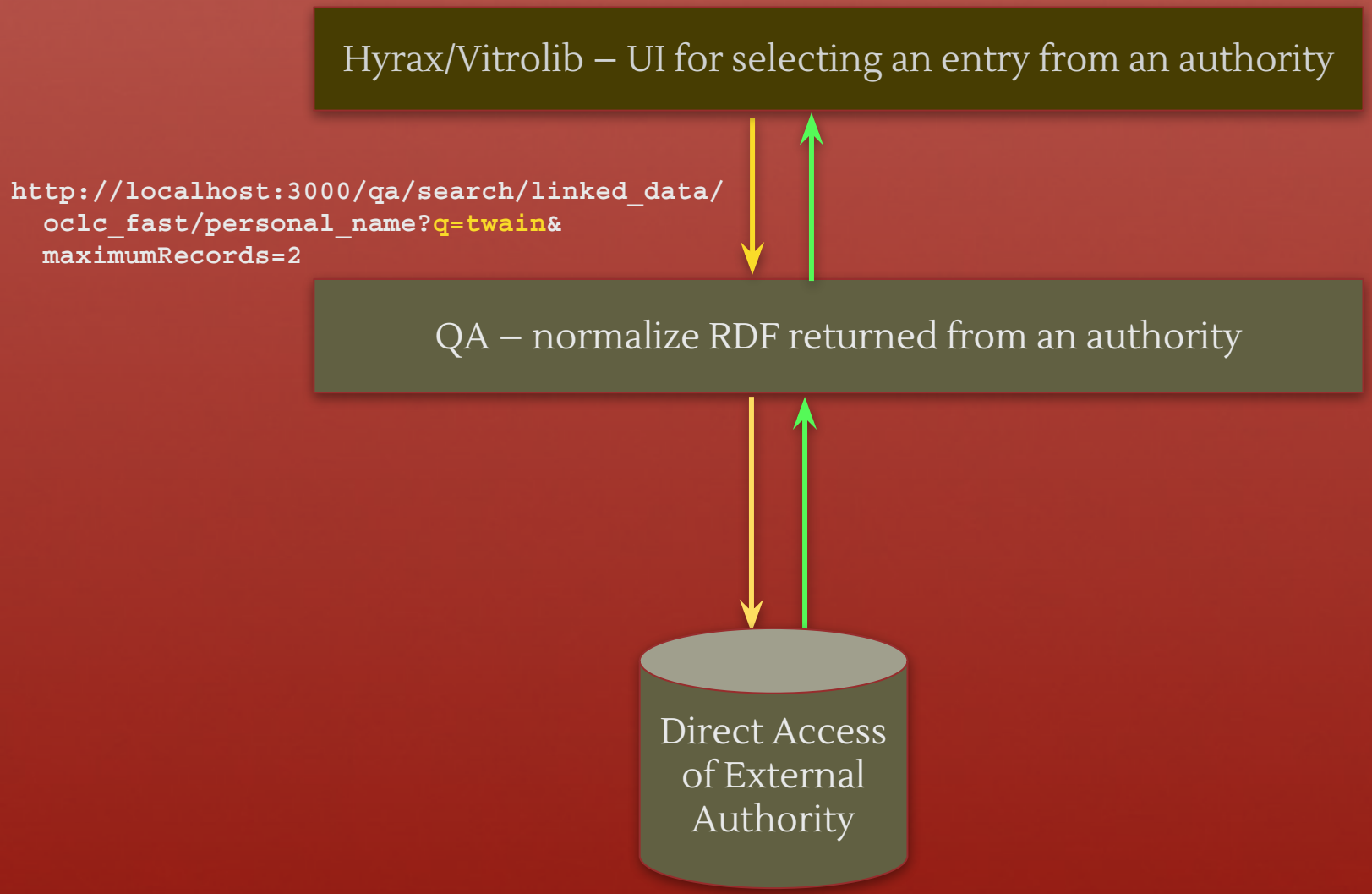
Multi-Server Architecture

QA – normalize RDF returned from an authority

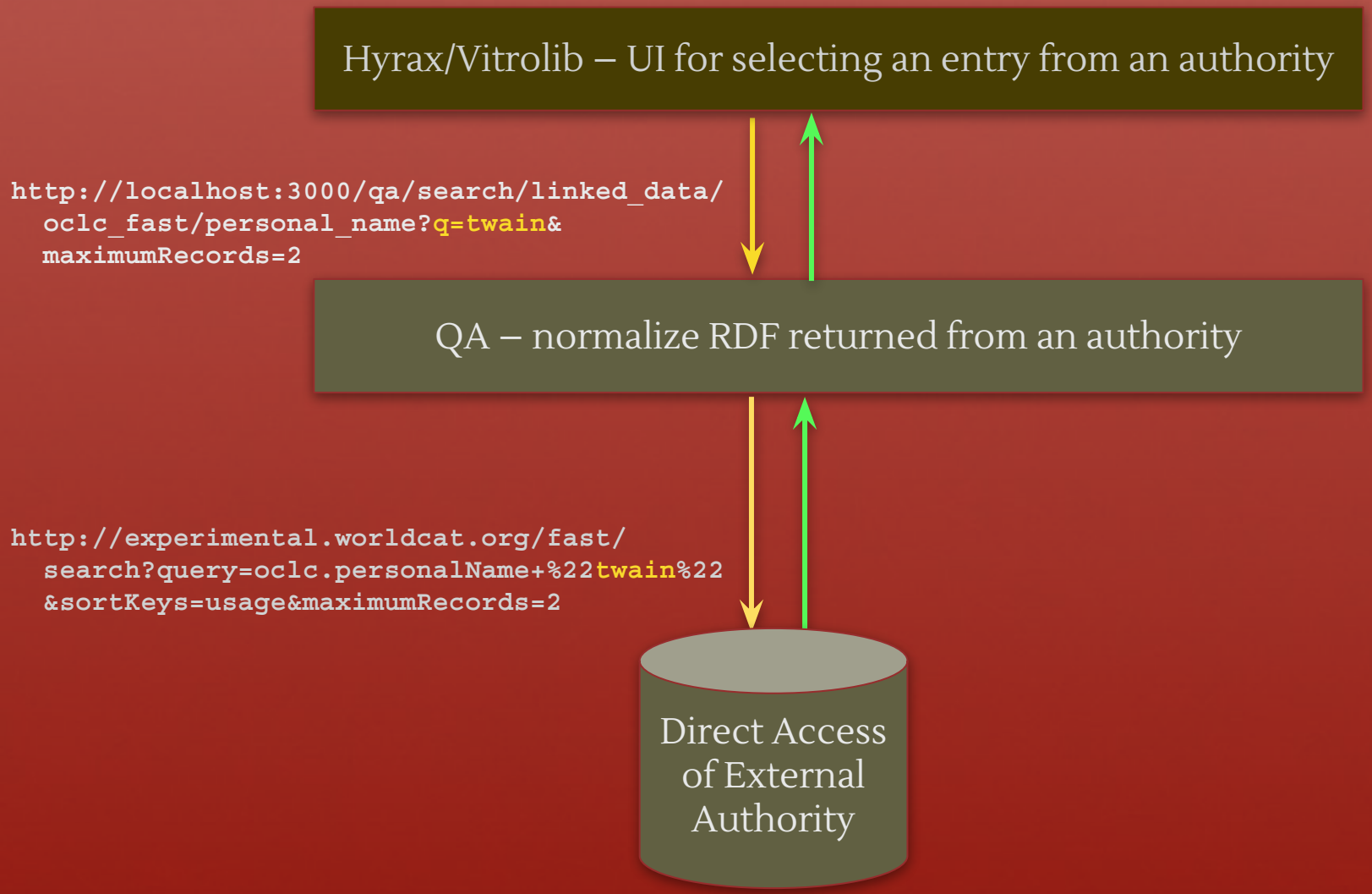
Multi-Server Architecture



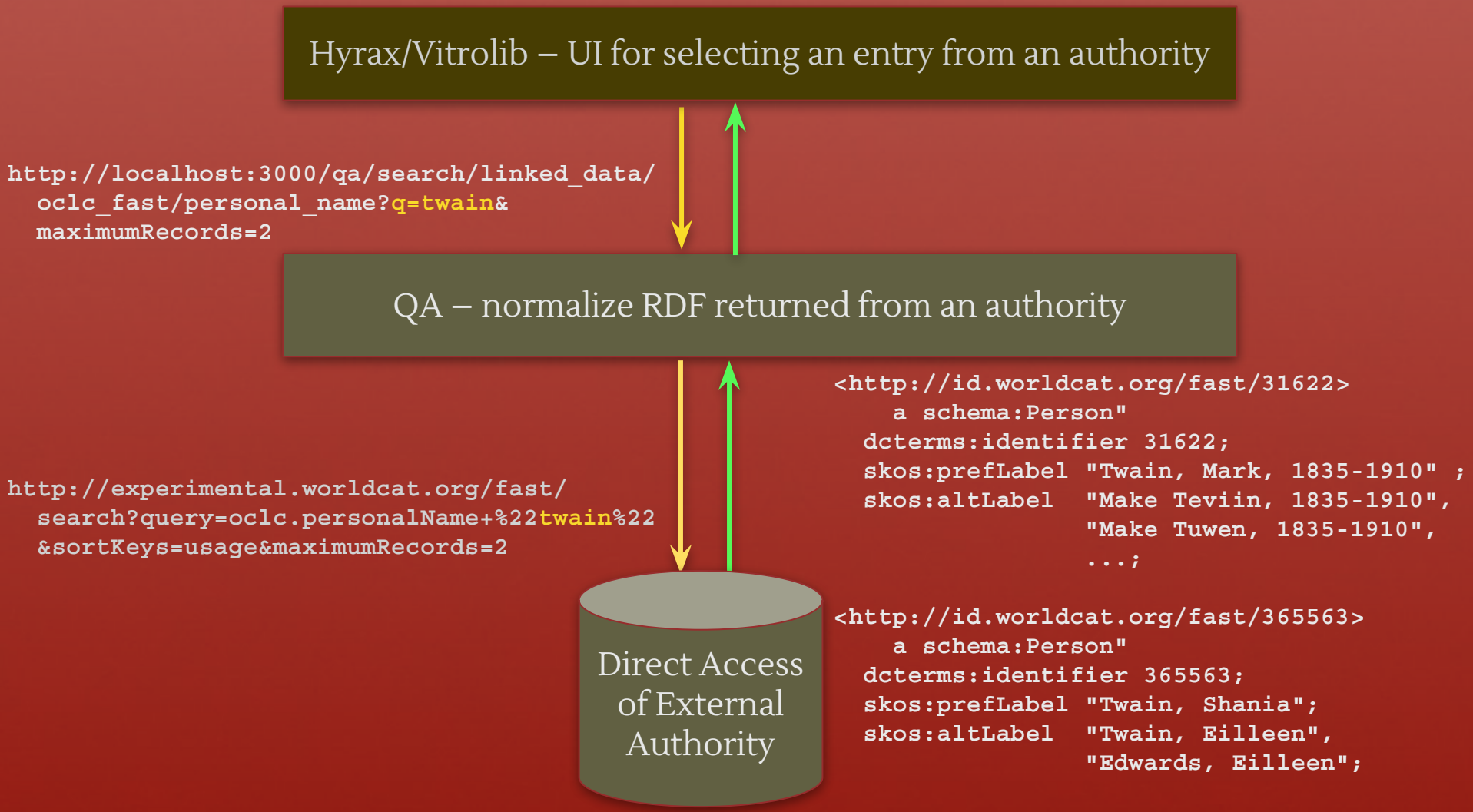
Multi-Server Architecture



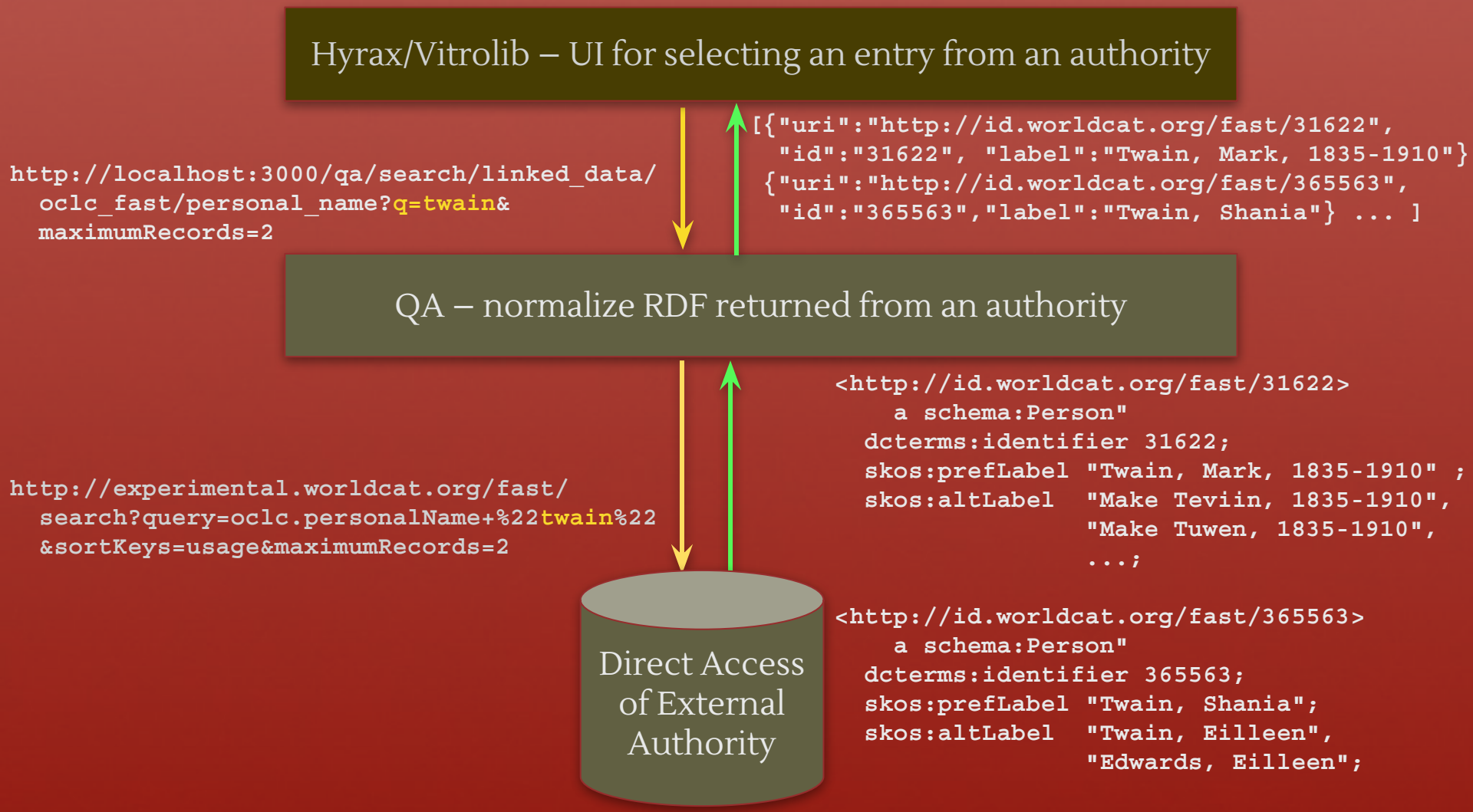
Multi-Server Architecture



Multi-Server Architecture



Multi-Server Architecture



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

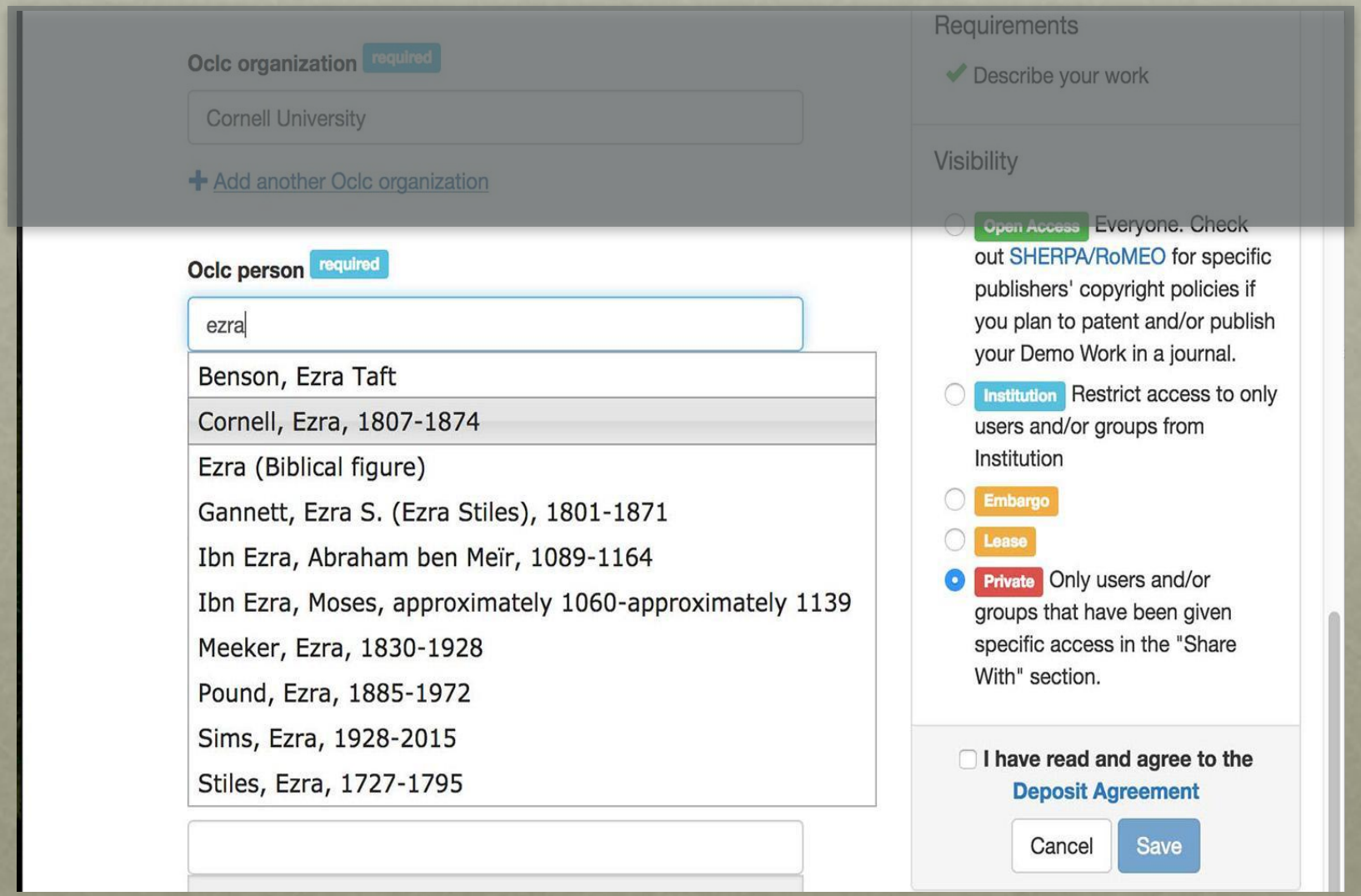

```
[{"uri":"http://id.worldcat.org/fast/31622",
  "id":"31622",
  "label":"Twain, Mark, 1835-1910"},
 {"uri":"http://id.worldcat.org/fast/365563",
  "id":"365563",
  "label":"Twain, Shania"}]

[{"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"}]
```

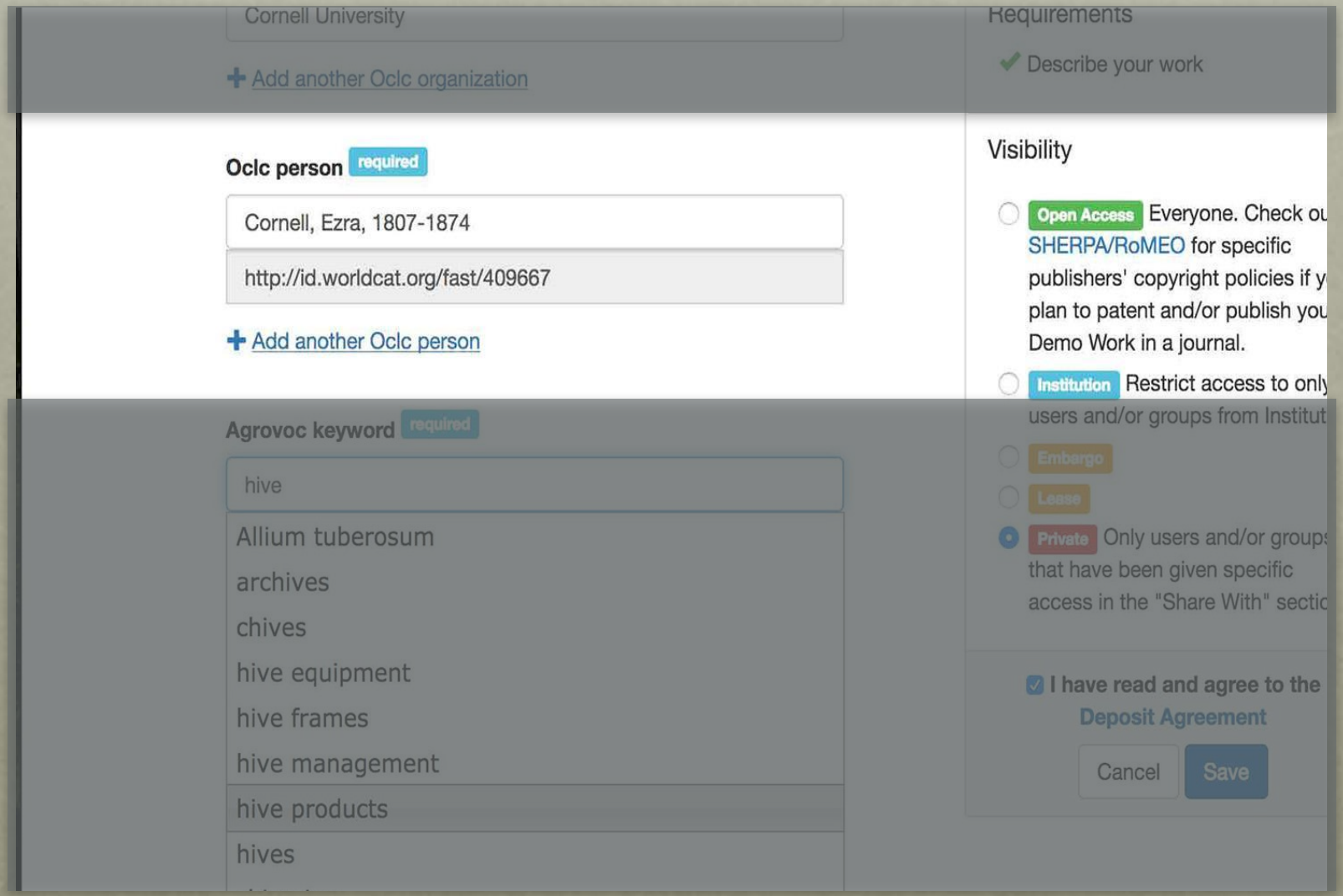
Autocomplete Saving String and URI

Authority: OCLC FAST Subauthority: PersonName



Selected String and URI

Saves both string and URI



Access and Normalization Layer

Questioning Authority

- ruby gem that can be deployed as a standalone server (ld4l-labs/qa_server)

Challenge 2 Access API ✔

- configurable access to authorities providing a single api for end applications (ld4l-labs/linked_data_authorities)

Challenge 3 Results Format ✔

- uses rdf/linkeddata gem for processing any number of formats (e.g. jsonld, rdf-xml, ttl, nt, etc.)

Challenge 4 Ontologies ✔

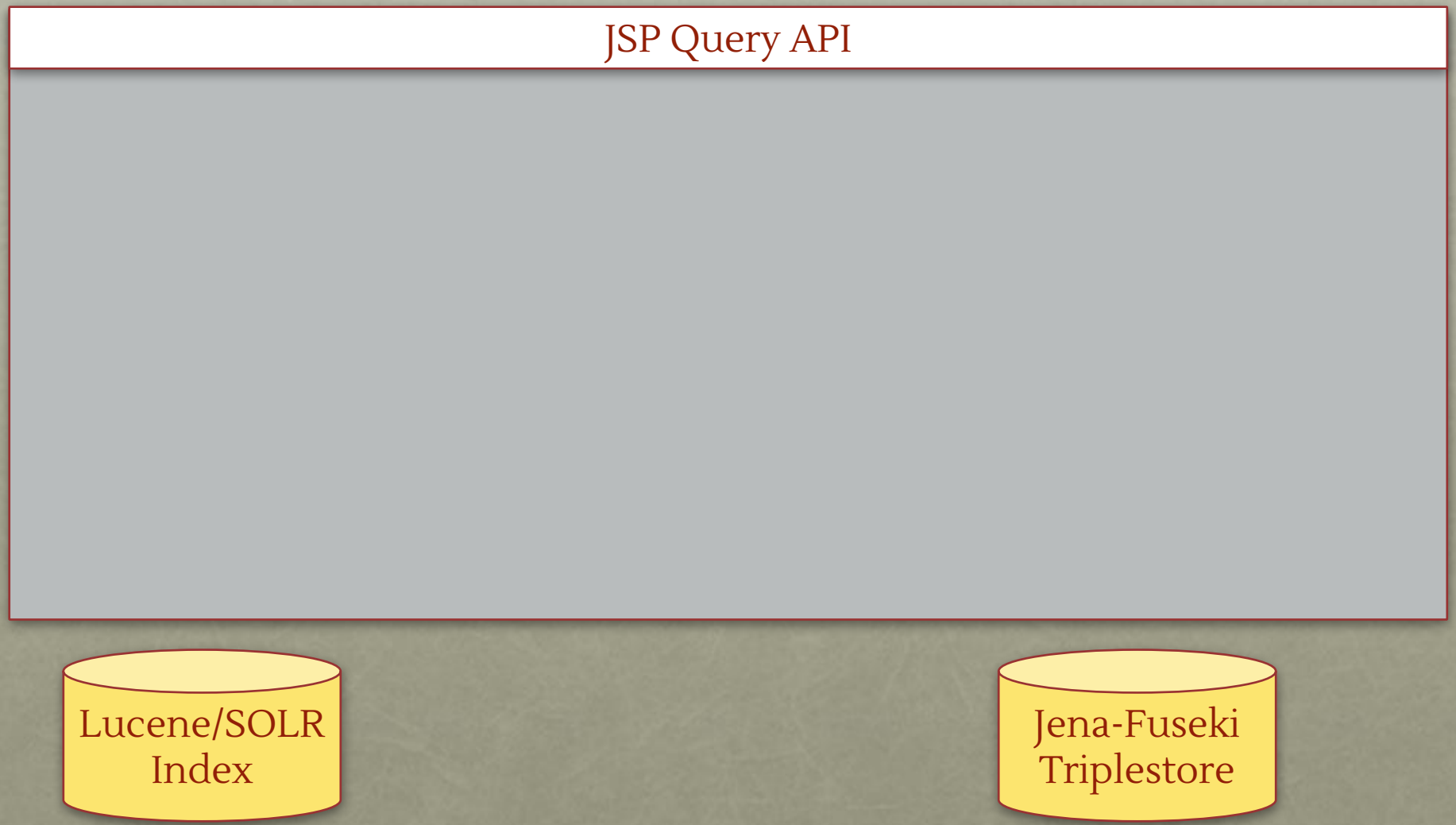
- normalizes results to simplify end applications UI

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

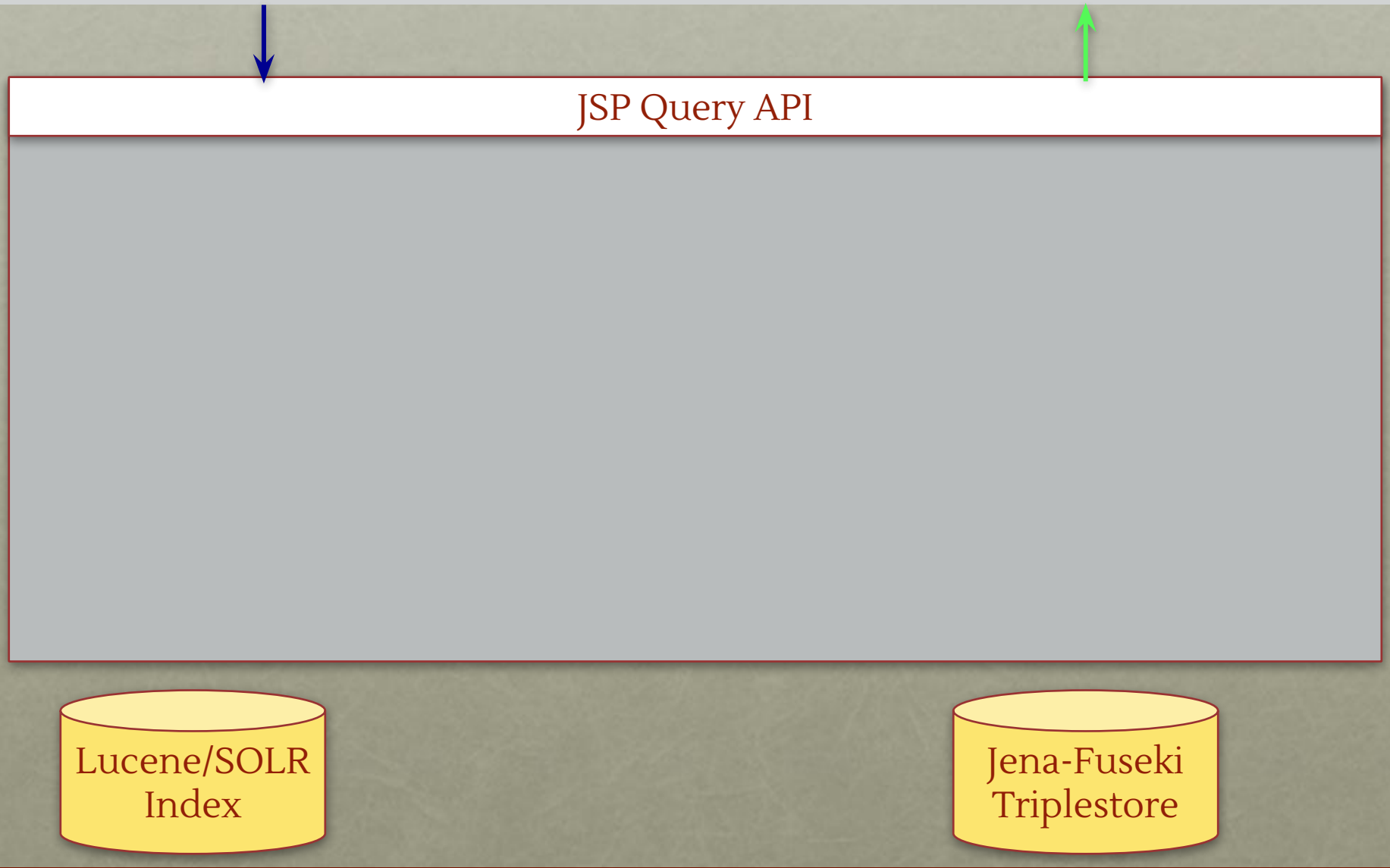
One full setup per authority



Cache Server Query Process

One full setup per authority

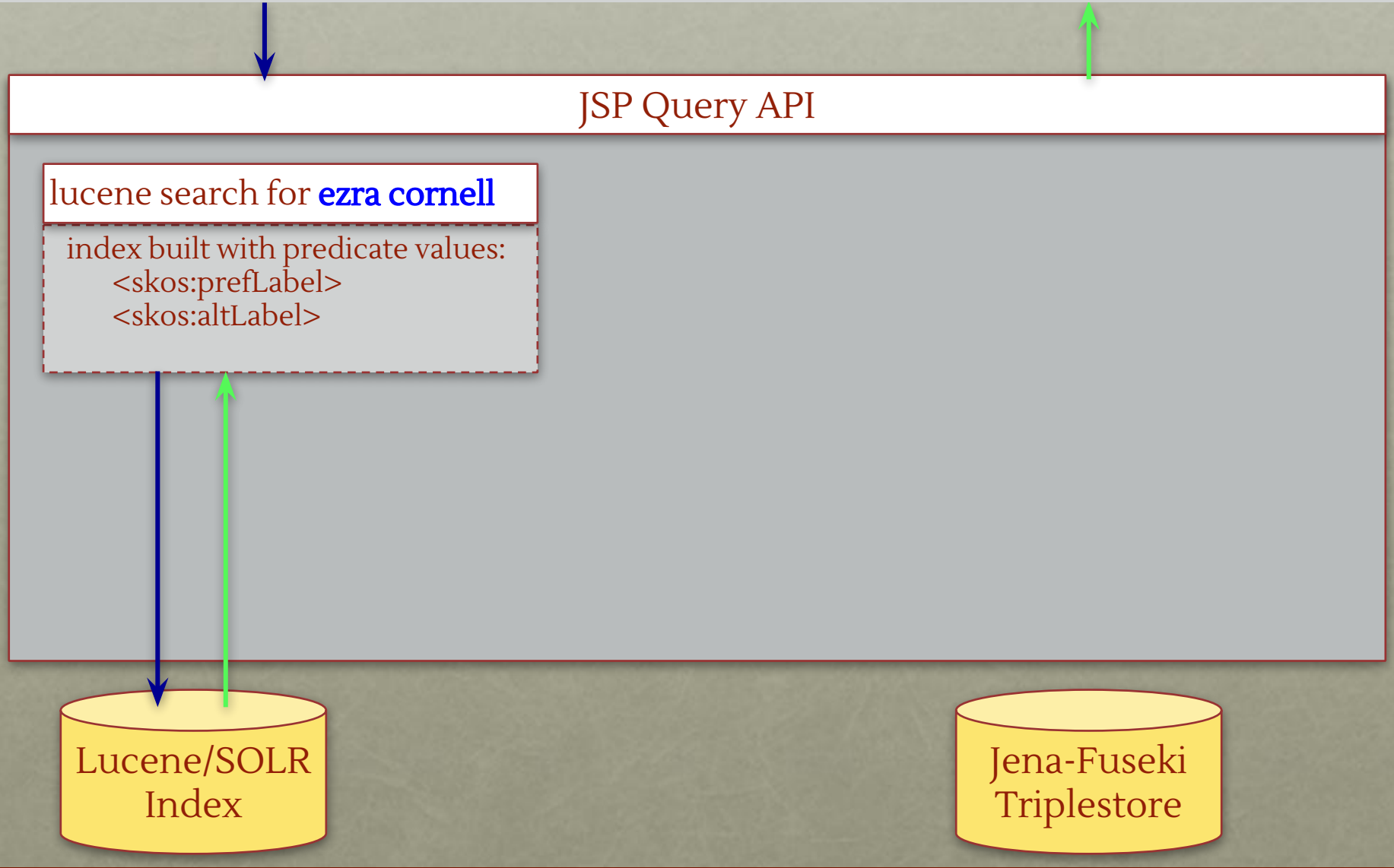
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



Cache Server Query Process

One full setup per authority

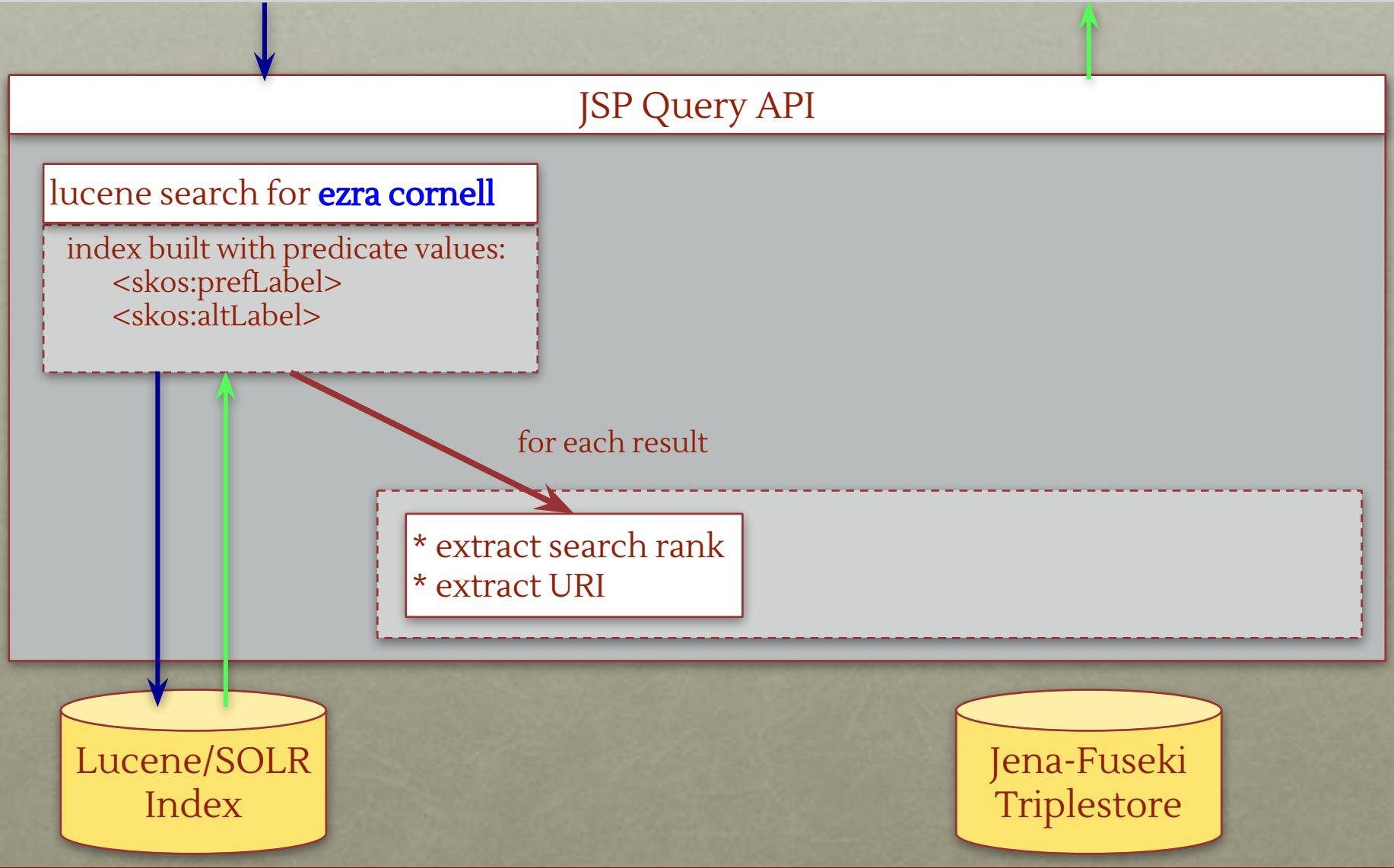
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



Cache Server Query Process

One full setup per authority

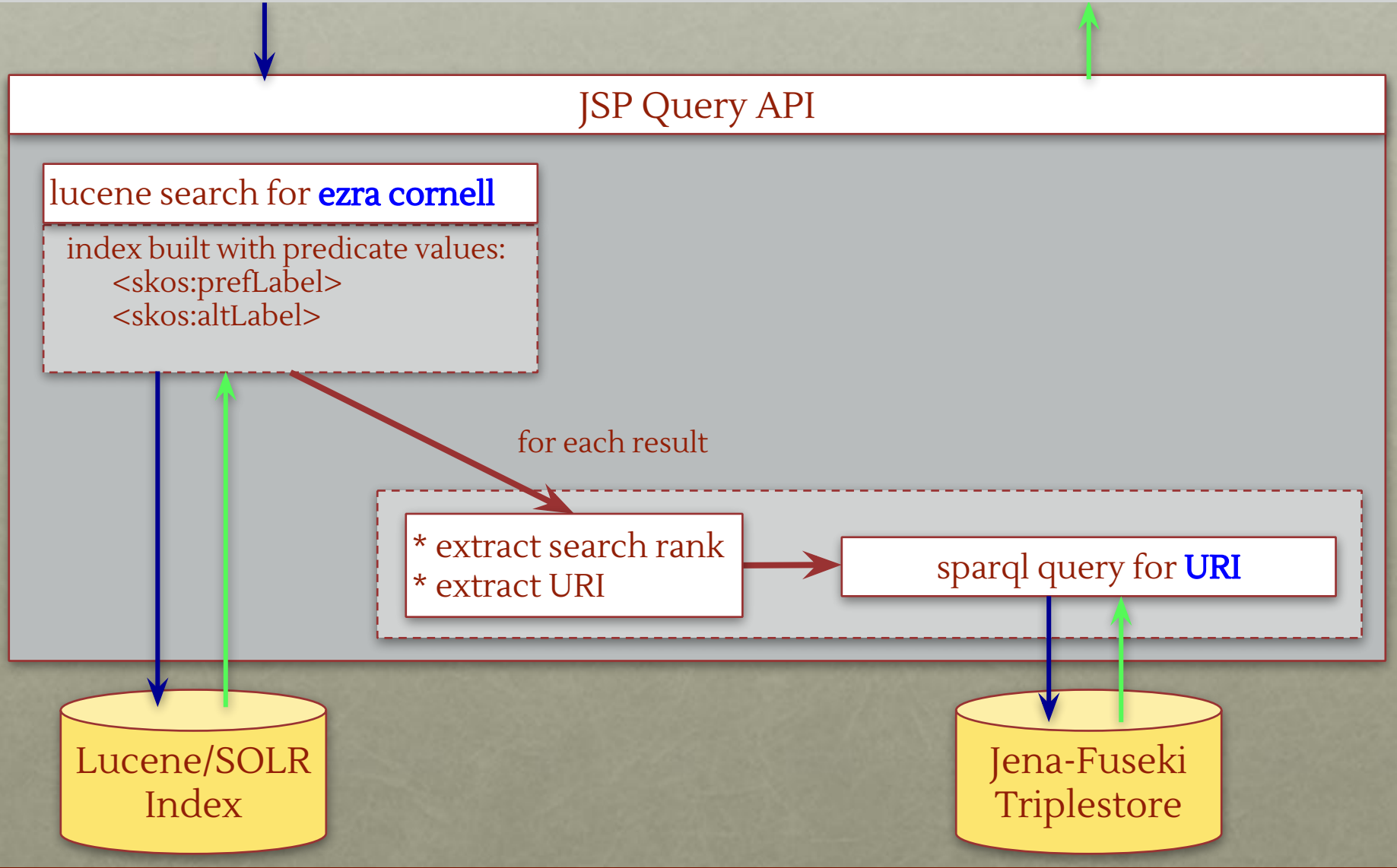
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



Cache Server Query Process

One full setup per authority

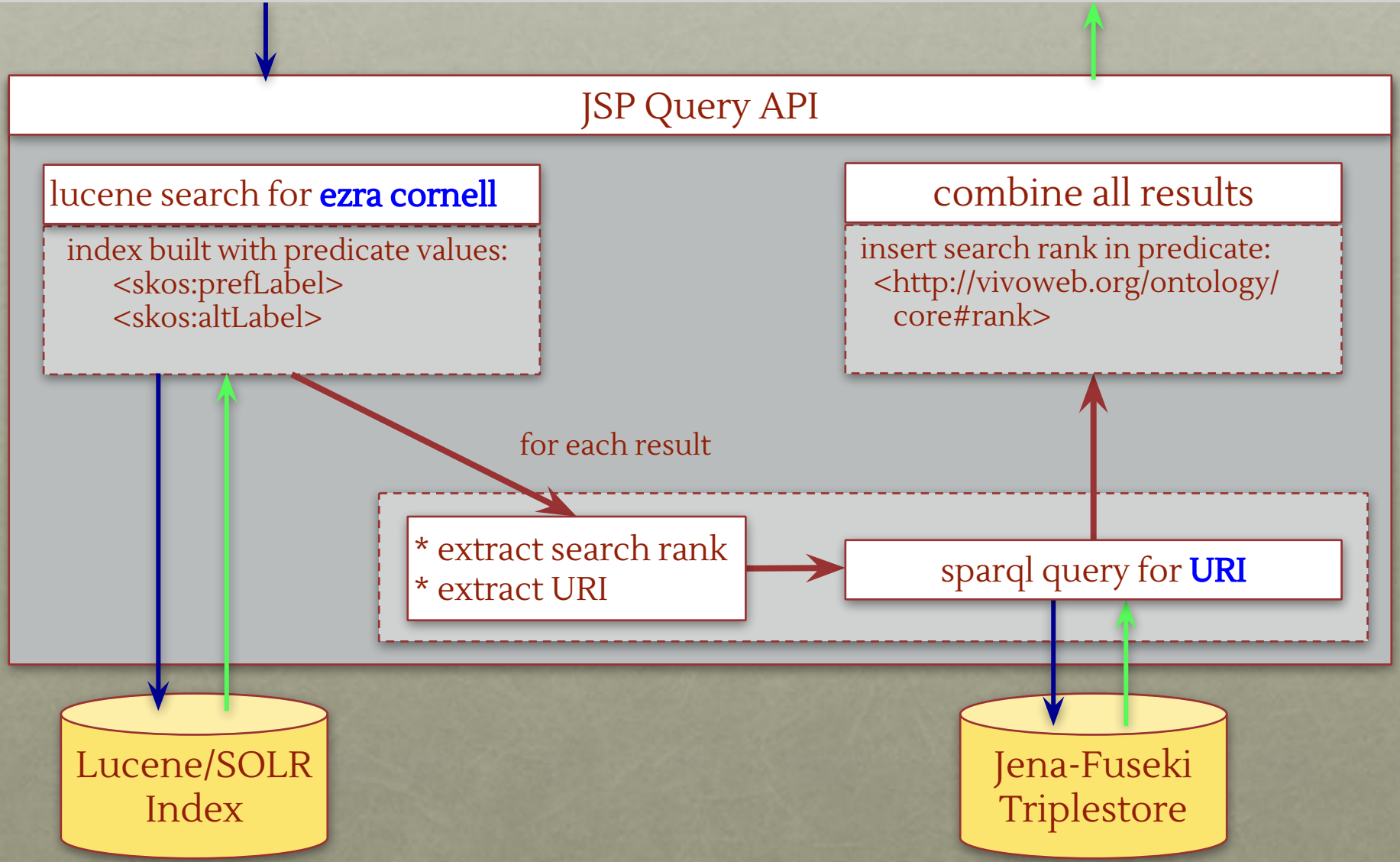
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



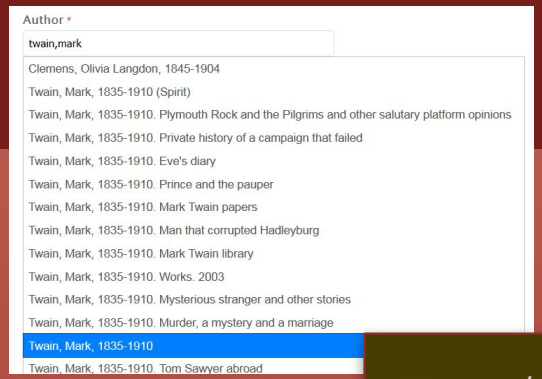
Cache Server Query Process

One full setup per authority

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



UI-QA-Authority



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

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

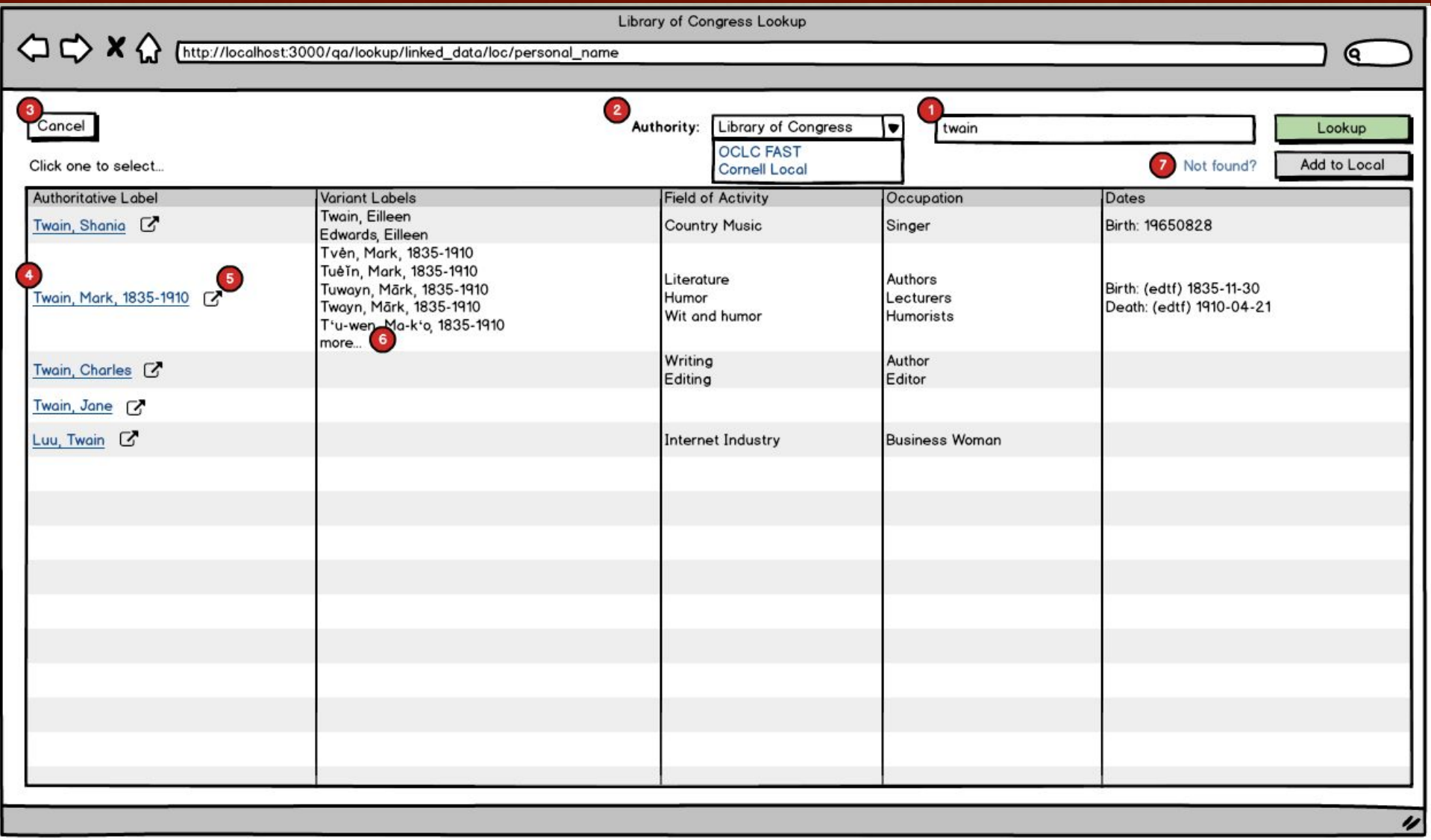
Challenge 7 Order of Results ✔
- addition of search rank predicate provides ability to sort rdf graph search results for consistent presentation to the user

Third Set of Challenges

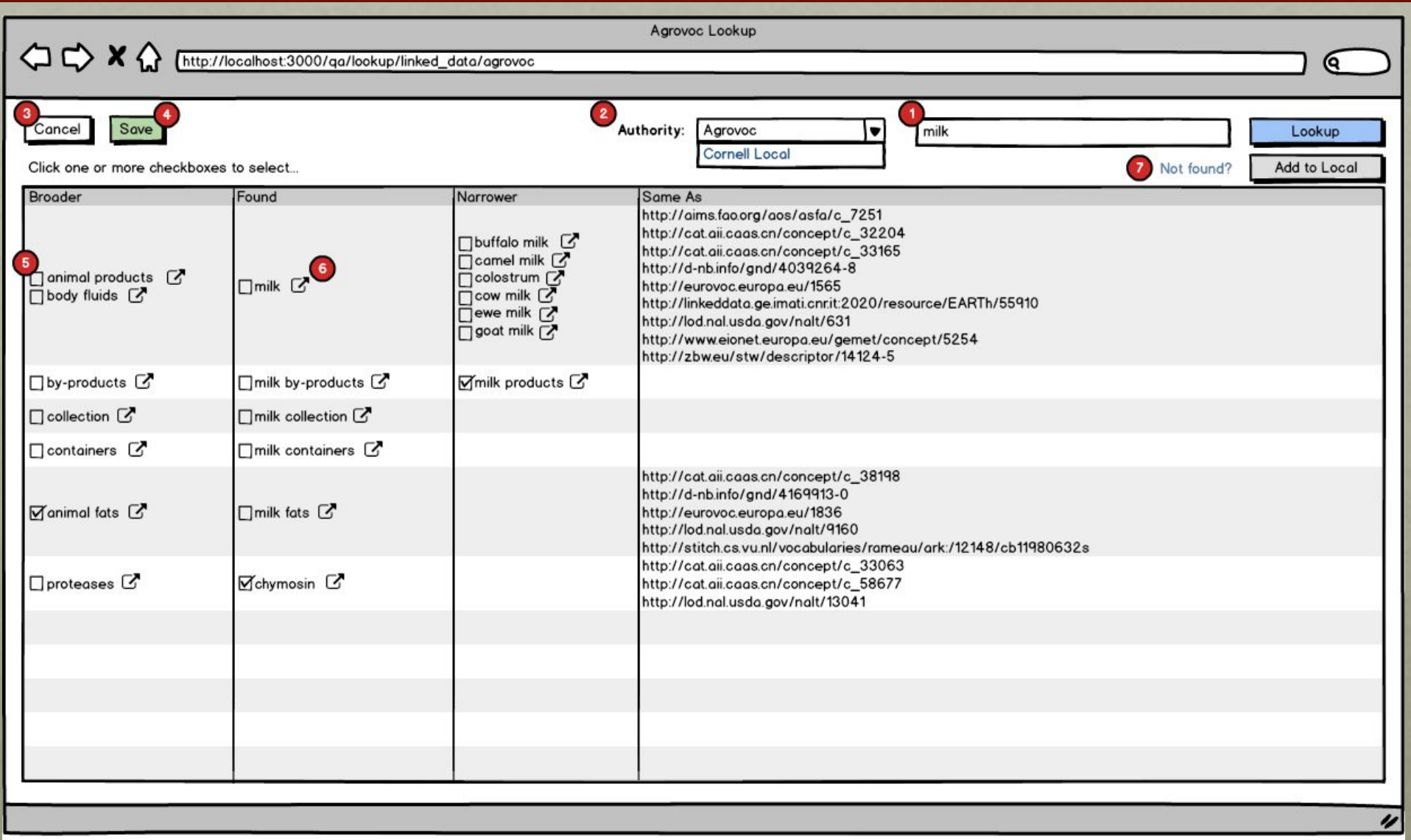
8. Disambiguation through better context
   e.g. expand from just prefLabel to preLabel, altLabel, birth/death dates, occupation, etc.

9. Reconciliation across multiple sources
   e.g. match LoC URI to OCLC FAST URI

Selecting a Term using Lookup with Context



Selecting a Term using Lookup with Context



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 uri	http://id.worldcat.org/fast/409667

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

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

Ezra Cornell | Search

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