

Linking the Data: Building Effective Authority and Identity Lookup

Huda Khan and E. Lynette Rayle
Cornell University

Collaborators:

Dave Eichmann (University of Iowa)
Simeon Warner and Dean Krafft (Cornell)



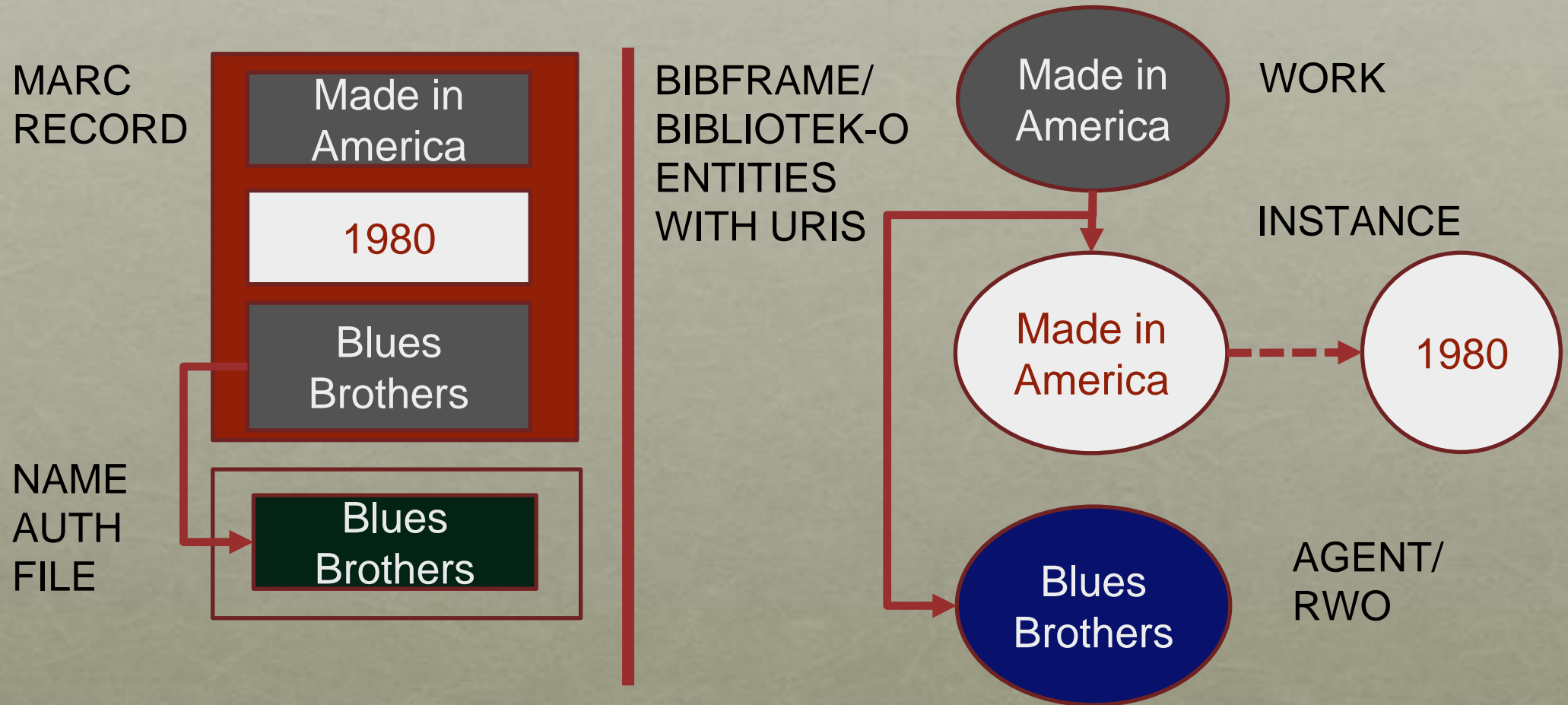
Overview

- Background and Motivation
- Examples:
 - VitroLib
 - Hyrax
- Architecture overview
- Future work
- Questions

Background

- Mellon Foundation-funded LD4 Projects
- Transition library systems to linked data
- Link better, explore better
 - Flat record -> Discrete entities with well-defined relationships
 - String identifiers -> URIs
 - Relationships with other linked data

Background



Background

“A cataloger is an individual responsible for the processes of ***description, subject analysis, classification, and authority control*** of library materials. Catalogers serve as the ‘foundation of all library service, as they are the ones who organize information in such a way as to make it easily accessible’.” (*Emphasis* mine)

From <https://en.wikipedia.org/wiki/Cataloging>

Background

- Traditional practices: Authority File
 - E.g. Name Authority Files, Subject Headings, Genre Forms from LOC
 - String as unique identifier, e.g. “Mark Twain”
- Tasks and workflows
 - Identification, “Aboutness”
 - Disambiguation
 - Context and original authority record

Background

- Goals: Design and architecture around accessing authorities
- VitroLib
 - Prototype cataloging editor
 - Creates/uses linked data
 - Enables lookup and use of authorities
- Hyrax
 - Samvera technology stack
 - Incorporate authorities into institutional repository records

VitroLib Demo

[Home](#)[Works](#)[Instances](#)[Items](#)[People](#)[Organizations](#)

Catalog a new resource [?]

[Work +](#)[Statistics](#)

80

Works

65

Instances

17

Items

96

People

50

Organizations

191

Agents

New Work

Title *

Made in America

Type *

Audio

Language *

English

Author or Other Role *

Performer

blues brothers

Blues Brothers Band

Blues Brothers (Motion picture)

Wood Brothers (Blues group)

Soul Brothers (Blues group)

Blues Brothers Horns (Musical group)

Blues Eldorados

Blues, Thomas

Blues Rockers

Blues Brothers

Blues revue

Audio

Language *

English

Author or Other Role *

Performer

Selected: **Blues Brothers** [\(Verify this match or change selection\)](#)

LC Subject Heading

soul music

Soul music--History and criticism

Guitar music (Soul)

Philadelphia soul (Music)

Soul--Buddhism

Soul worship

Soul mates

Soul music

Soul--Judaism

Soul musicians

Soul--Hinduism

Soul in literature

Soul--Christianity

Soul--Jainism

Selected Subject:

Soul music

[\(Verify this match\)](#) or [change selection](#)

Has Instance (RDA Manifestation)

Title

Made in America

Statement of Responsibility

Blues Brothers

Type

Audio disc

- Aperture card
- Archival controlled
- Audio belt
- Audio cartridge
- Audio cassette
- Audio cylinder
- Audio disc
- Audio roll
- Audio wire reel
- Audiotape reel
- Belt
- Card
- Cartridge
- Cassette
- Computer card
- Computer chip
- Computer chip cartridge
- Computer disc
- Computer disc cartridge
- Computer tape cartridge

Select an existing Entity or create a new one.

Type
Audio disc

Issue Number

SD 16025

Matrix Number

UPC

EAN

Publisher

atlantic recording co

Atlantic Recording Corporation

Atlantic Consultants

Atlantic jazzlore

Atlantic mystery

Atlantic Workshop

Sa Atlantic series

* requ Atlantic Institute

Atlantic jazz

Atlantic Manor

Atlantic Underwriters

Type

Audio disc



Issue Number

SD 16025

Matrix Number

UPC

EAN

Selected: Atlantic Recording Corporation [\(Verify this match or change selection\)](#)

Date of Publication

1980

or [Cancel](#)

** required fields*

Resource URI: <http://vitrolib.cornell.edu/individual/n2726>

[Made in America](#)



Audio

(work)

authors and other roles

[Blues Brothers](#) (Performer)

Instance of Work

[Made in America \(Audio disc\)](#) 1980

subject

[Soul music](#)

genre form

Manage Associated Genre Forms

Genre Form	
Video recordings	Remove

Lookup and Add Genre Form

 [Cancel](#)


Search[Home](#) | [Works](#) | [Instances](#) | [Items](#) | [People](#) | [Organizations](#)

Manage Associated Genre Forms

Genre Form

[Video recordings](#)

[Remove](#)

Lookup and Add Genre Form

[Cancel](#)

Manage Associated Genre Forms

Genre Form

[Video recordings](#)

[Remove](#)

Lookup and Add Genre Form

SEARCH

Genre Form	Notes	Additional Info
<input type="checkbox"/> Clay animation television programs Alternate Labels: <ul style="list-style-type: none">Claymation television programsSculptmation television programs	No definition provided.	Broader: <ul style="list-style-type: none">Animated television programs
<input type="checkbox"/> Silhouette animation films Alternate Labels: <ul style="list-style-type: none">Semi-object animation filmsSand animation filmsFlat-figure animation films	No definition provided.	Broader: <ul style="list-style-type: none">Animated films

<input type="checkbox"/> Stop-motion animation films Alternate Labels: <ul style="list-style-type: none"> Object animation films Frame-by-frame animation films Stop-action animation films Stop-motion animated films 	No definition provided.	Narrower: <ul style="list-style-type: none"> Clay animation films
<input type="checkbox"/> Clay animation films Alternate Labels: <ul style="list-style-type: none"> Claymation films Sculptmation films 	No definition provided.	Broader: <ul style="list-style-type: none"> Stop-motion animation films
<input type="checkbox"/> Computer animation television programs Alternate Labels: <ul style="list-style-type: none"> CG television programs (Computer animation) Computer animated television programs CGI television programs (Computer animation) 	No definition provided.	Broader: <ul style="list-style-type: none"> Animated television programs
<input type="checkbox"/> Live-action/animation films Alternate Labels: <ul style="list-style-type: none"> Animated/live-action films, Hybrid 	This heading is used as a genre/form heading for films that feature interaction between live-action and animated elements.	Broader: <ul style="list-style-type: none"> Animated films

Manage Associated Genre Forms

Genre Form

[Video recordings](#)

[Remove](#)

Lookup and Add Genre Form

SEARCH

Genre Form	Notes	Additional Info
<input type="checkbox"/> Clay animation television programs Alternate Labels: <ul style="list-style-type: none">Claymation television programsSculptmation television programs	No definition provided.	Broader: <ul style="list-style-type: none">Animated television programs
<input type="checkbox"/> Silhouette animation films Alternate Labels: <ul style="list-style-type: none">Semi-object animation filmsSand animation filmsFlat-figure animation films	No definition provided.	Broader: <ul style="list-style-type: none">Animated films

Clay animation television programs

From [Library of Congress Genre/Form Terms](#)

[Details](#)[Visualization](#)

Clay animation television programs

URI(s)

> <http://id.loc.gov/authorities/genreForms/qf2011026141>

Instance Of

- > [MADS/RDF GenreForm](#)
- > [MADS/RDF Authority](#)
- > [SKOS Concept](#) 

Scheme Membership(s)

> [Library of Congress Genre/Form Terms](#)

Collection Membership(s)

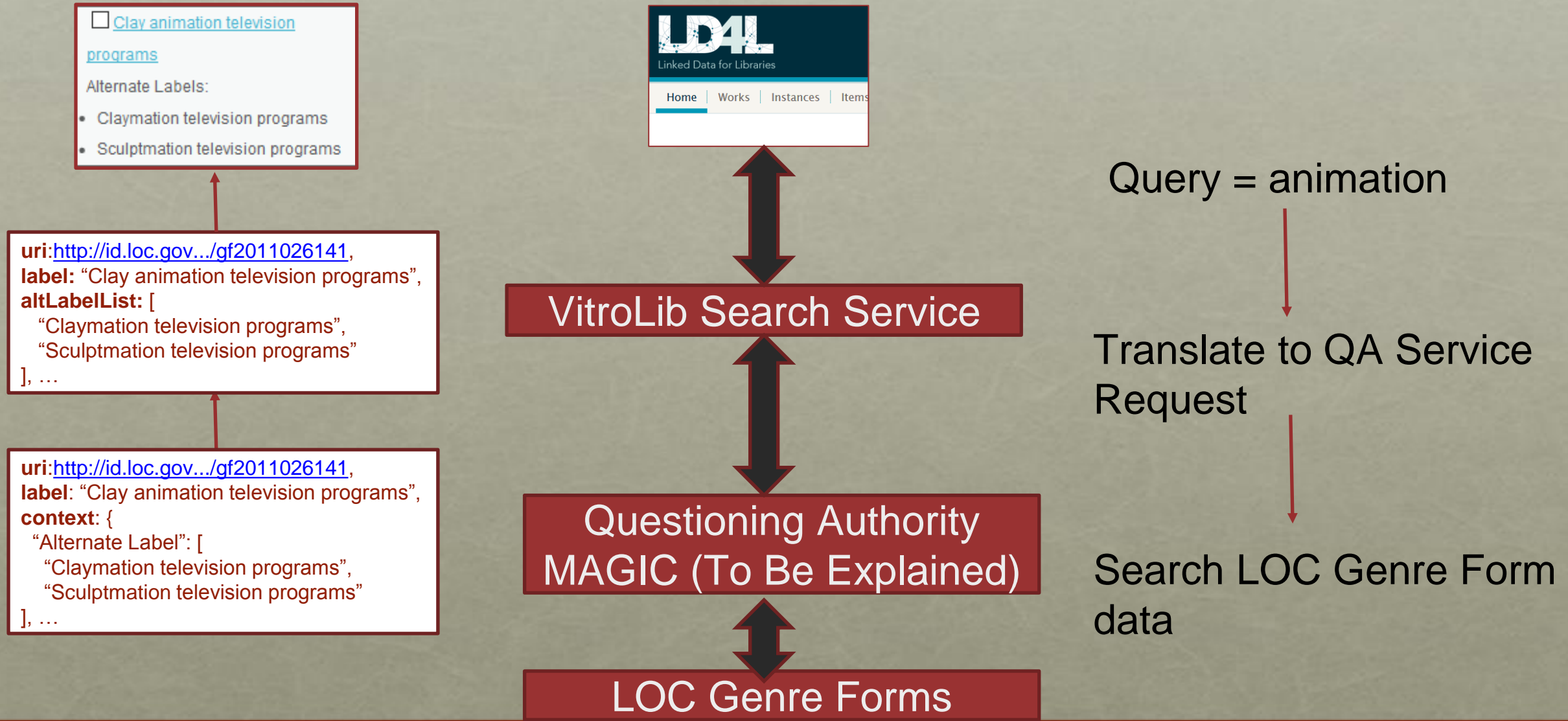
> http://id.loc.gov/authorities/genreForms/collection_LCGFT_General

Form

> Clay animation television programs

Variants

What just happened?



Hyrax Demo

Autocomplete Saving String and URI

Authority: OCLC FAST

Subauthority: PersonName

Oclc person **required**

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

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.

Selected String and URI

Saves both string and URI

Oclc organization required

Cornell University

[+ Add another Oclc organization](#)

Oclc person required

Cornell, Ezra, 1807-1874

<http://id.worldcat.org/fast/409667>

[+ Add another Oclc person](#)

Agrovoc keyword required

hive products

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

Selecting a Term using Lookup with Context

Library of Congress Lookup

http://localhost:3000/qa/lookup/linked_data/loc/personal_name

3 Cancel

2 Authority: Library of Congress
 OCLC FAST
 Cornell Local

1 twain

7 Not found?

Lookup

Add to Local

Click one to select...

Authoritative Label	Variant Labels	Field of Activity	Occupation	Dates
Twain, Shania ↗	Twain, Eilleen Edwards, Eilleen	Country Music	Singer	Birth: 19650828
4 Twain, Mark, 1835-1910 ↗ 5	Tvên, Mark, 1835-1910 Tuêĭn, Mark, 1835-1910 Tuwayn, Mārĭk, 1835-1910 Twayn, Mārĭk, 1835-1910 T'u-wen, Ma-k'ò, 1835-1910 more... 6	Literature Humor Wit and humor	Authors Lecturers Humorists	Birth: (edtf) 1835-11-30 Death: (edtf) 1910-04-21
Twain, Charles ↗		Writing Editing	Author Editor	
Twain, Jane ↗				
Luu, Twain ↗		Internet Industry	Business Woman	

Selecting a Term using Lookup with Context

Agrovoc Lookup

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

Authority:


Click one or more checkboxes to select...

Not found?

Broader	Found	Narrower	Same As
<input type="checkbox"/> animal products	<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.aii.caas.cn/concept/c_32204 http://cat.aii.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.aii.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.aii.caas.cn/concept/c_33063 http://cat.aii.caas.cn/concept/c_58677 http://lod.nal.usda.gov/nalt/13041

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)  <http://id.worldcat.org/fast/409667>

uri

Architecture

Technical Motivation

- Linked data provides...
 - URIs that identify specific terms (as opposed to ambiguity of using strings)
 - Reconciliation to relate terms that are defined in separate authorities
- Goals of implementation...
 - Provide a single process to access many authorities
 - Provide efficient and reliable access to authorities
 - Provide a means for disambiguation that empowers library staff to make the most accurate selections

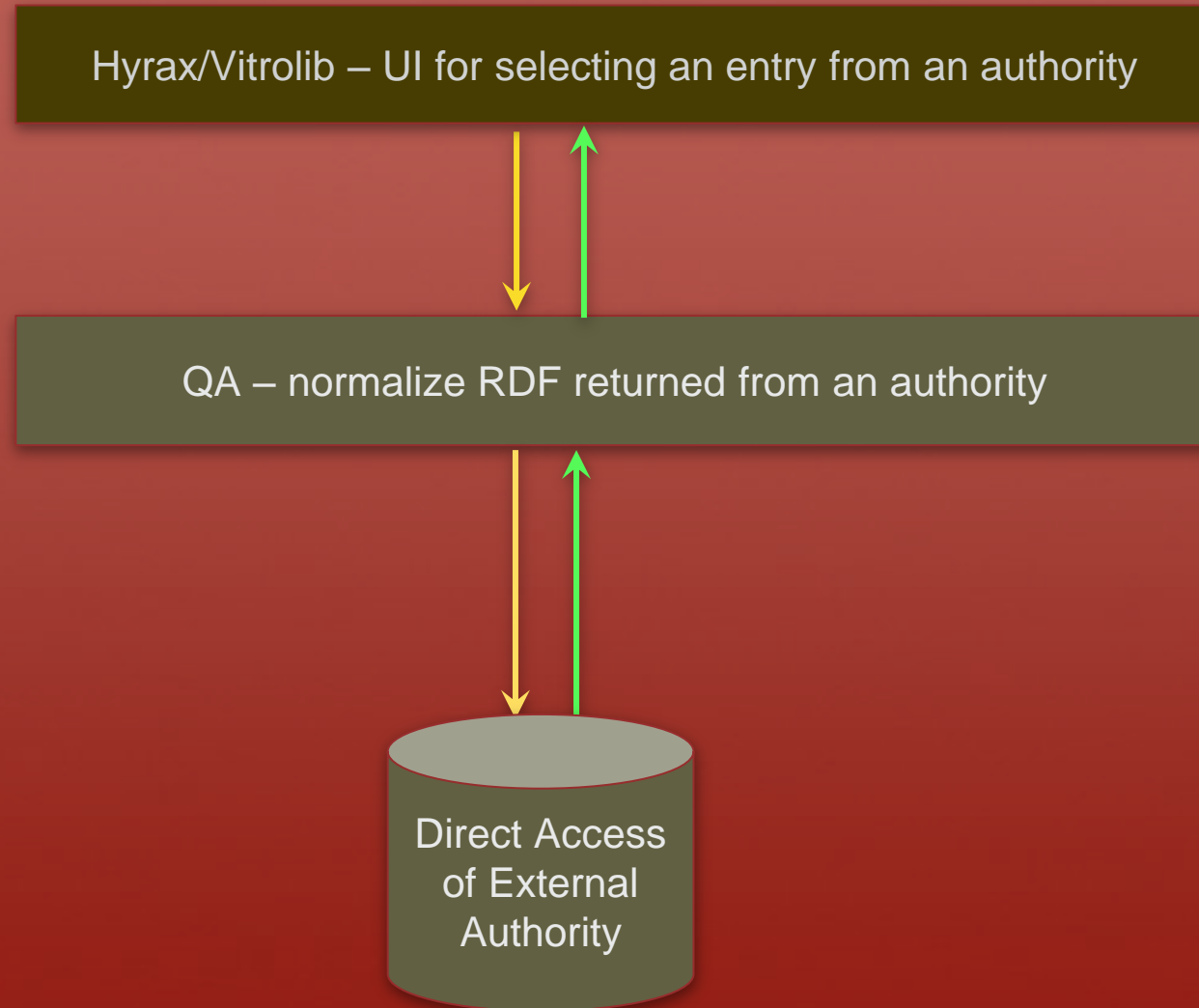
First Set of Challenges

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

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



QA – normalize RDF returned from an authority



Direct Access
of External
Authority

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

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](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://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



QA – normalize RDF returned from an authority



Direct Access
of External
Authority

```
<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";
```

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

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

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://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://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](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

OCLC FAST

```
[{"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"}]
```

GeoNames

```
[{"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)"}]
```

AgroVoc

```
[{"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"}]
```

Second Set of Challenges

5. Reliability & Efficiency

e.g. server uptime, server load

6. Accuracy

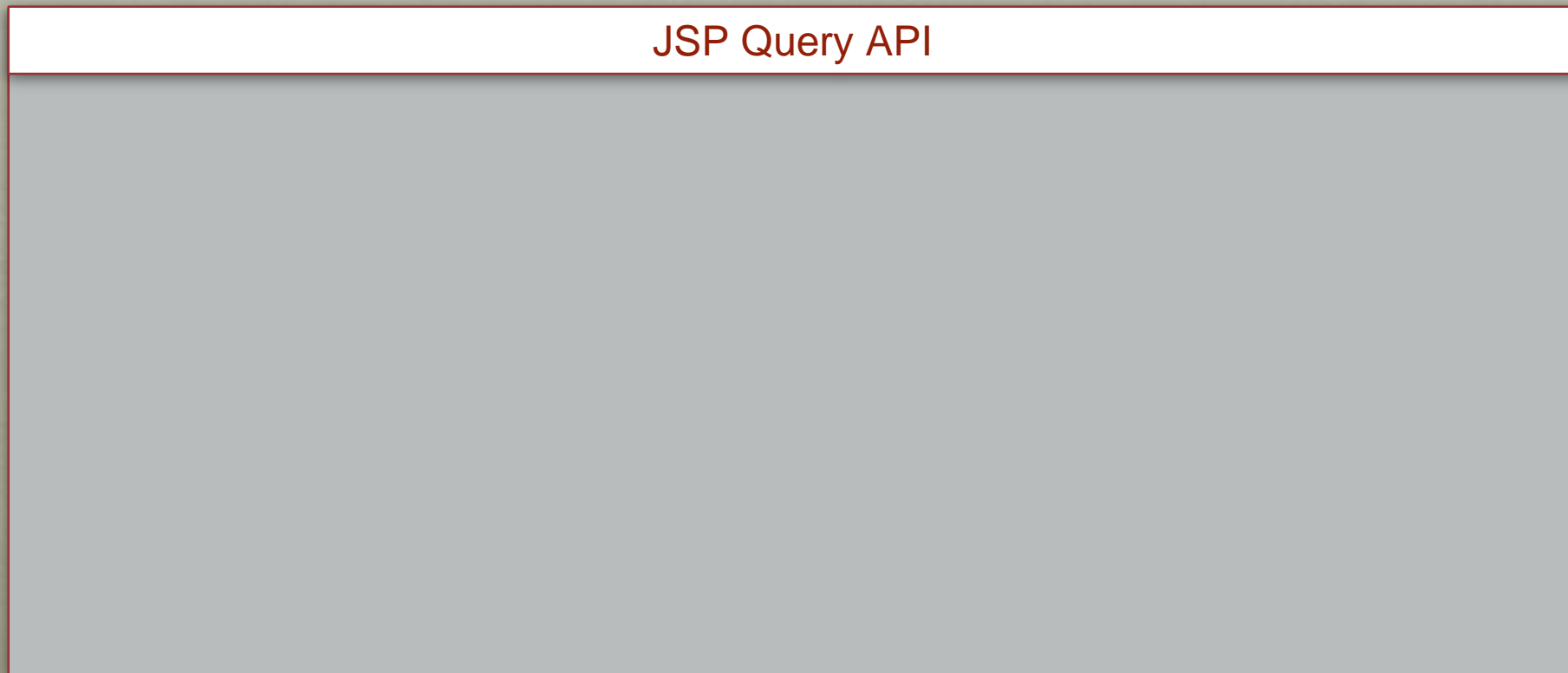
e.g. select results based on usage data, lexical match, custom weighting, other?

7. Order Ranking

e.g. How to order a graph?

Cache Server Query Process

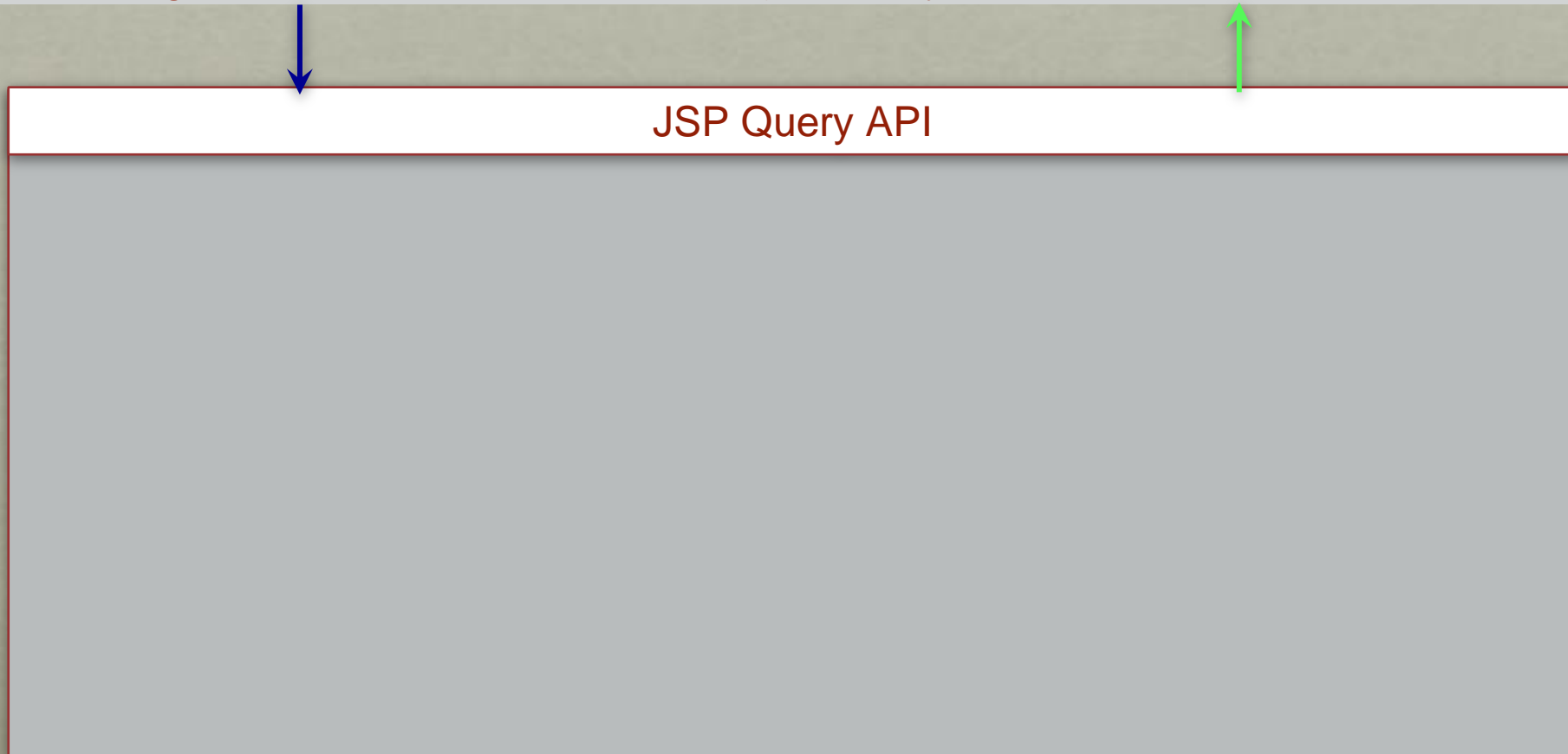
One full setup per authority



Cache Server Query Process

One full setup per authority

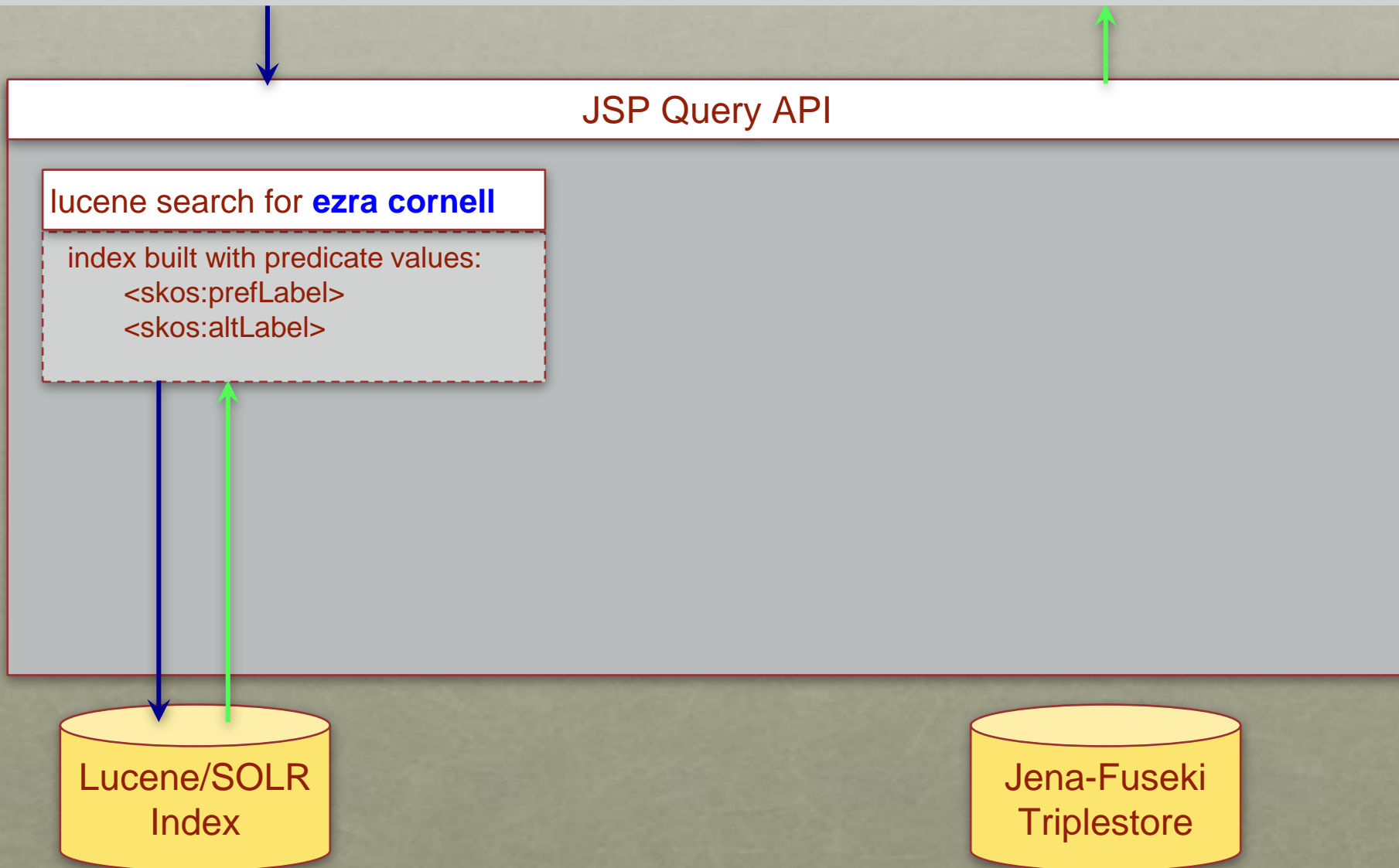
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



Cache Server Query Process

One full setup per authority

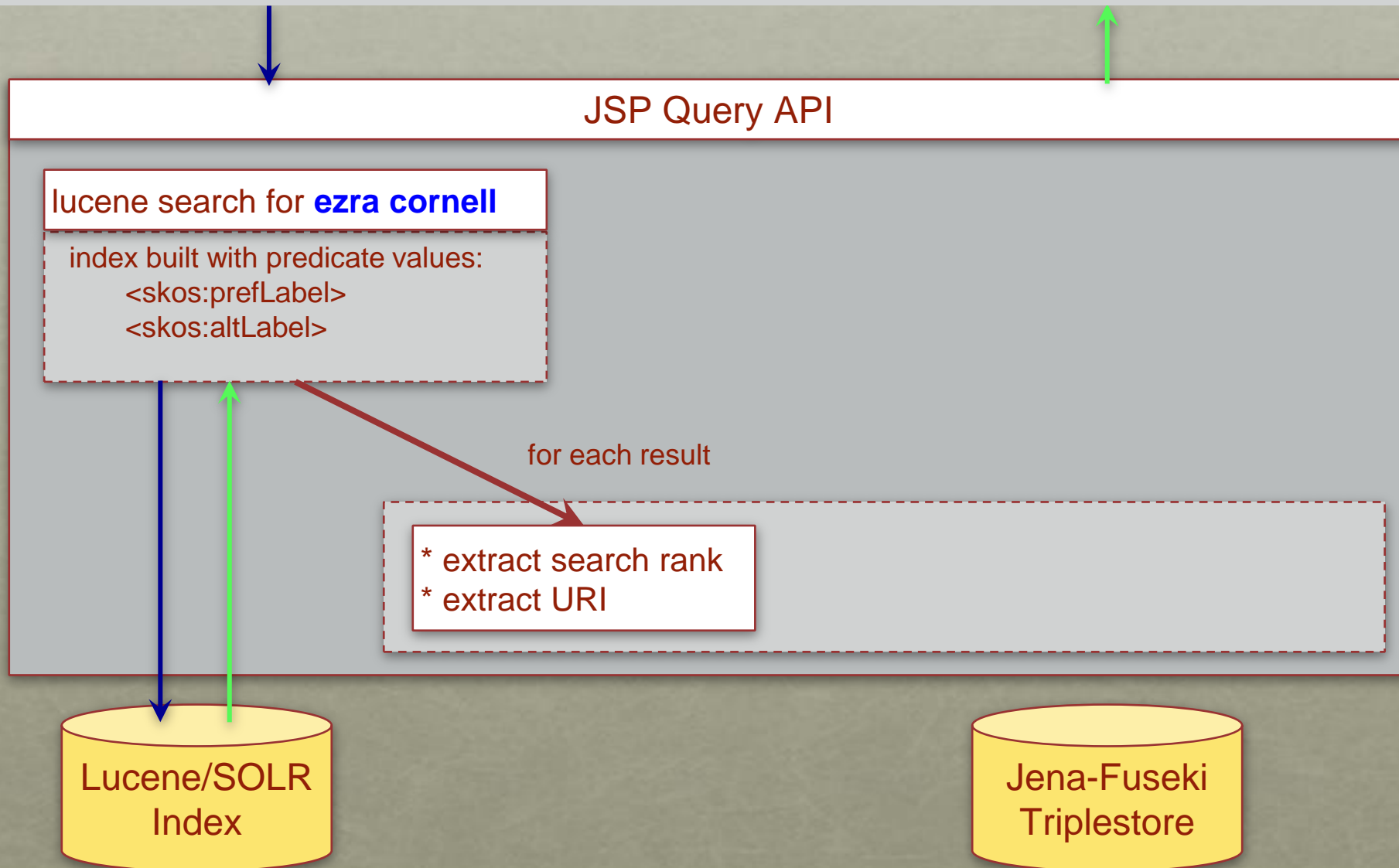
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



Cache Server Query Process

One full setup per authority

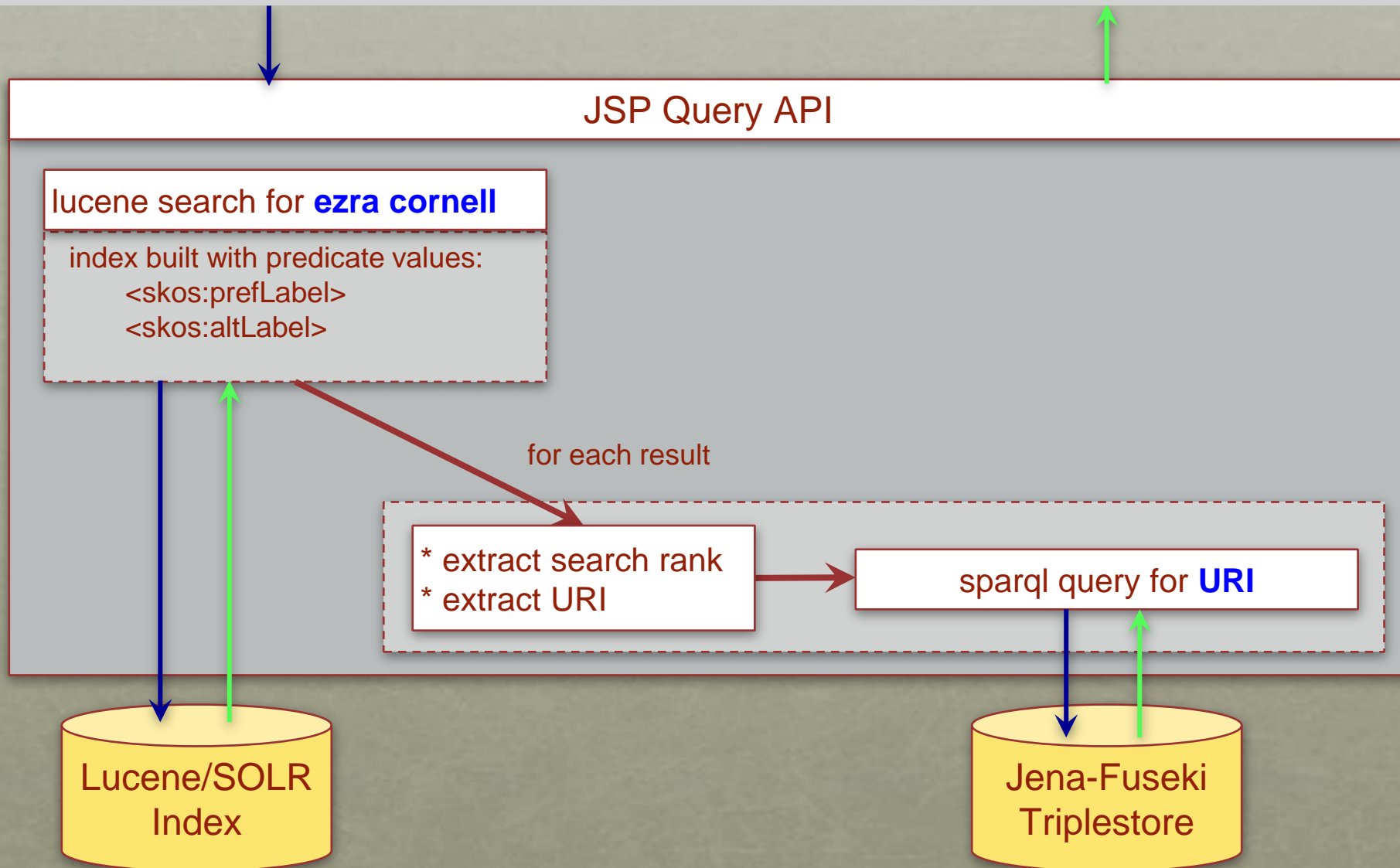
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



Cache Server Query Process

One full setup per authority

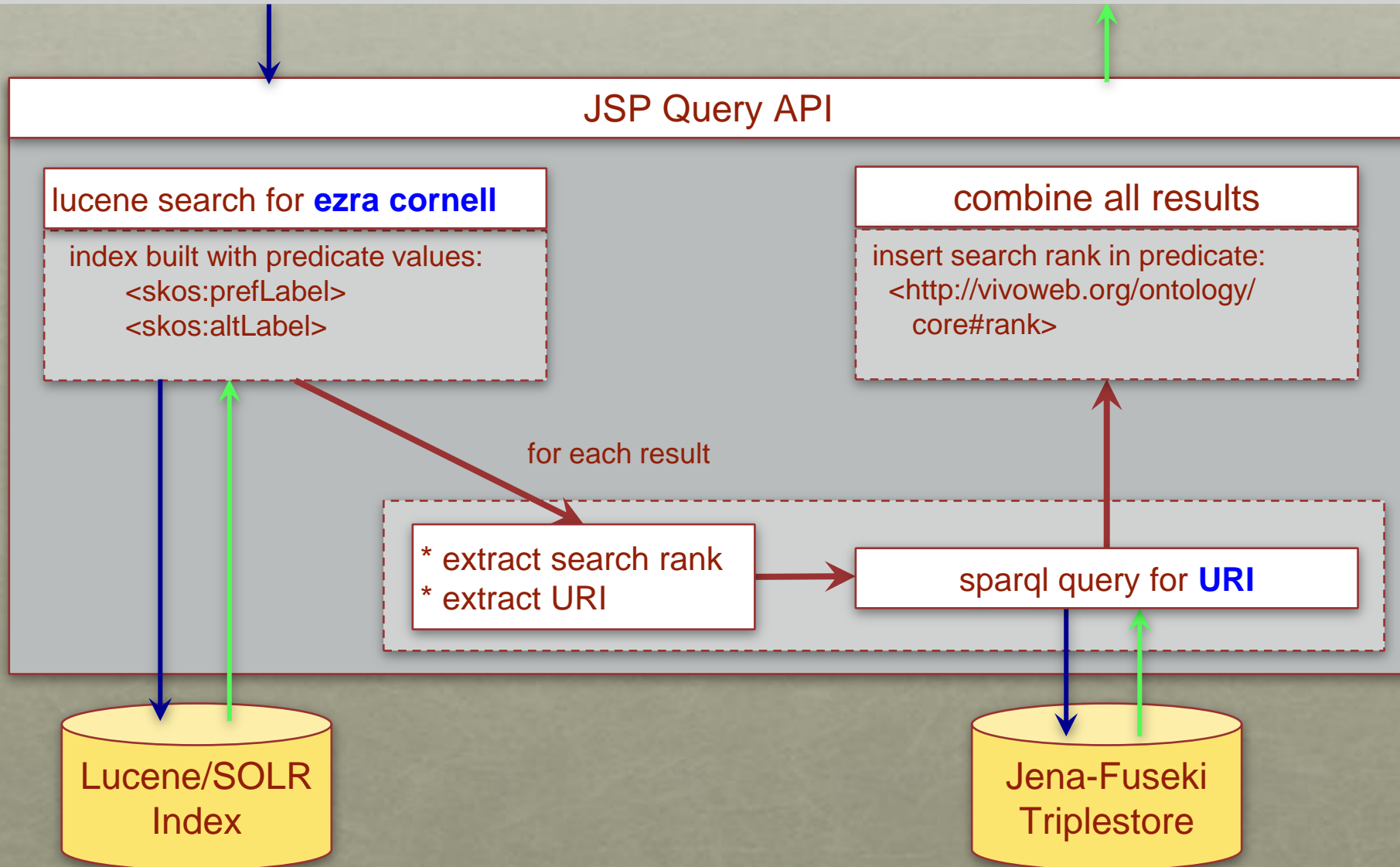
http://services.ld4l.org/ld4l_services/loc_name_batch.jsp?query=ezra%20cornell&maxRecords=10



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

Author *
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

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

What's next?

Addressing Architectural Challenges

- Generalize process for accessing context on the cache server and in the normalization layer
- Multi-authority search and reconciliation
- Address the need for cache refresh
- Mirrored cache servers

User Experience and Design

- User-centered Design
 - Observe, listen, learn, design, evaluate, iterate
- Iteratively design and evaluate UI for lookup/authorities with catalogers
- Search result ranking/ordering/filtering for catalogers
- Additional UI platforms, e.g. FOLIO

Questions?

<http://tinyurl.com/ld4l-auth-access>

Appendix for Challenges 1-4

Challenge 1: Documentation

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

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=rd	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

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
- [Wikidata](#) – I've heard that they are working on Reconciliation issues but haven't yet explored in depth
 - [Intro Video](#) (3hrs)
 - [API Access](#)
 - SPARQL – [user manual](#)
 - [federated queries with other authorities](#)

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

Links to Code & More

- [QA Server](#) - Code for a small app that provides the Questioning Authority normalization layer
- [Linked Data Authorities](#) - Configurations that can be used with QA Server
- [LD4L Services](#) - UI access to Cache Server
- [VitroLib](#) - Code for the VitroLib cataloging tool