LD4L PRESENTATION TO MWG FORUM

May 15, 2015

USE CASE 2, CONVERSION, POST-PROCESSING, AND ENTITY RESOLUTION

Rebecca Younes
Semantic Web Developer, Mann Library

USE CASE 2 DEMO

Use Case 2

- *See and search on works by people to discover works of interest based on connection to people, and to understand people based on their relation to works*
- Links between catalog data and profile systems for the enhancement of each
- Catalog data: biblio-centric
- Profile system: people-centered

Use Case 2 Demo

- A demo on a very small scale of what we want to do for the full catalog
- Demonstrate links between CUL catalog and VIVO
- Round-trip from catalog to VIVO and back to catalog
- Sample data: Cornell thesis records

Online thesis catalog record

The Suggestive Influence Of Lineups On Memory

A Comparative Study Of Children And Adults

by Caisa Elizabeth Royer.

Author/Creator	Royer, Caisa Elizabeth.
Format	Thesis
Language	English
Published	c2014
Subject	Lineup Misinformation Effect Memory
Summary	This study was designed to test whether making a selection from a target-absent lineup could alter memory for the actual perpetrator. Three different lineup presentations were used: simultaneous, sequential, and elimination. In addition, three different memory tests were used in which participants saw two of the following three faces: the actual thief, a suspect from the original lineup, and a novel suspect. The sequential lineup resulted in the fewest initial false identifications, while the simultaneous lineup appeared to protect the most against memory alteration. Logistic regression models were utilized to estimate the likelihood of making a particular identification during the memory test. Participants were more likely to identify the innocent original suspect when they were older, non-white, and were shown the sequential lineup. Participants were more likely to make an accurate identification of the thief when they were younger and had made an identification from the initial lineup.
Description	59 pages.
Other author/creator	Ceci, Stephen John thesis advisor
Notes	Includes bibliographical references.

Availability

✔ Available online

Connect to full text. Access to electronic version of some theses may be restricted.

Link from thesis advisor to catalog search results

Home | About Us | Libraries and Hours | Courses | Research | Services | Ask a Librarian | Help | My Account

Search / **Catalog**

Selected Items | Search History | Borrow Direct | Interlibrary Loan

Refine your results

Access
- Online 2
- At the Library 1

Format
- Thesis 2

Author/Creator

Publication Year

Language

Subject/Genre

Fiction/Non-Fiction

Library Location

"Ceci, Stephen John" | Author/Creator | Search

Advanced Search | Search Tips

or Start over

Author/Creator: "Ceci, Stephen John" ×

Did you mean to type: ceri?

1 - 2 of 2

20 per page | Sort by relevance

1. Lie to me : compliant false accusations by children c2014

Hritz, Amelia Courtney.

Thesis c2014 English

Online

Multiple locations

Select

2. The Suggestive Influence Of Lineups On Memory : A Comparative Study Of Children And Adults c2014

Royer, Caisa Elizabeth.

Thesis c2014 English

Online

Select

Linking to richer data

VIVO Research & Expertise Across Cornell

Index | Log in

Search

Home | People | Organizations | Research | Events

Ceci, Stephen John | Helen L. Carr Professor of Developmental Psychology

Positions

- Professor, Human Development (HD), College of Human Ecology (CHE)

Currently, with Wendy Williams, I am in the throes of a major project examining sex differences in cognitive performance, which has culminated this year in one authored book, two chapters, and a review article (*Current Directions in Psychological Science*). We also created the Cornell Institute for Women in Science (CIWS) web page, and had an article accepted in the Proceedings of the National Academy of Sciences (PNAS) which was selected as a feature article ("This week in PNAS", or TWIP (... more)

Websites

Human Ecology profile
Cornell Institute for Women in Science
Cornell Library: Books authored
Cornell Library: Theses advised

Networks

Co-author Network

Map of Science

Affiliations | Extension | Research | Publications | Teaching | Service | Background | Other | View All

teaching activities

COMM-1840: Six Pretty Good Books: Explorations in Social Science – Fall 2012

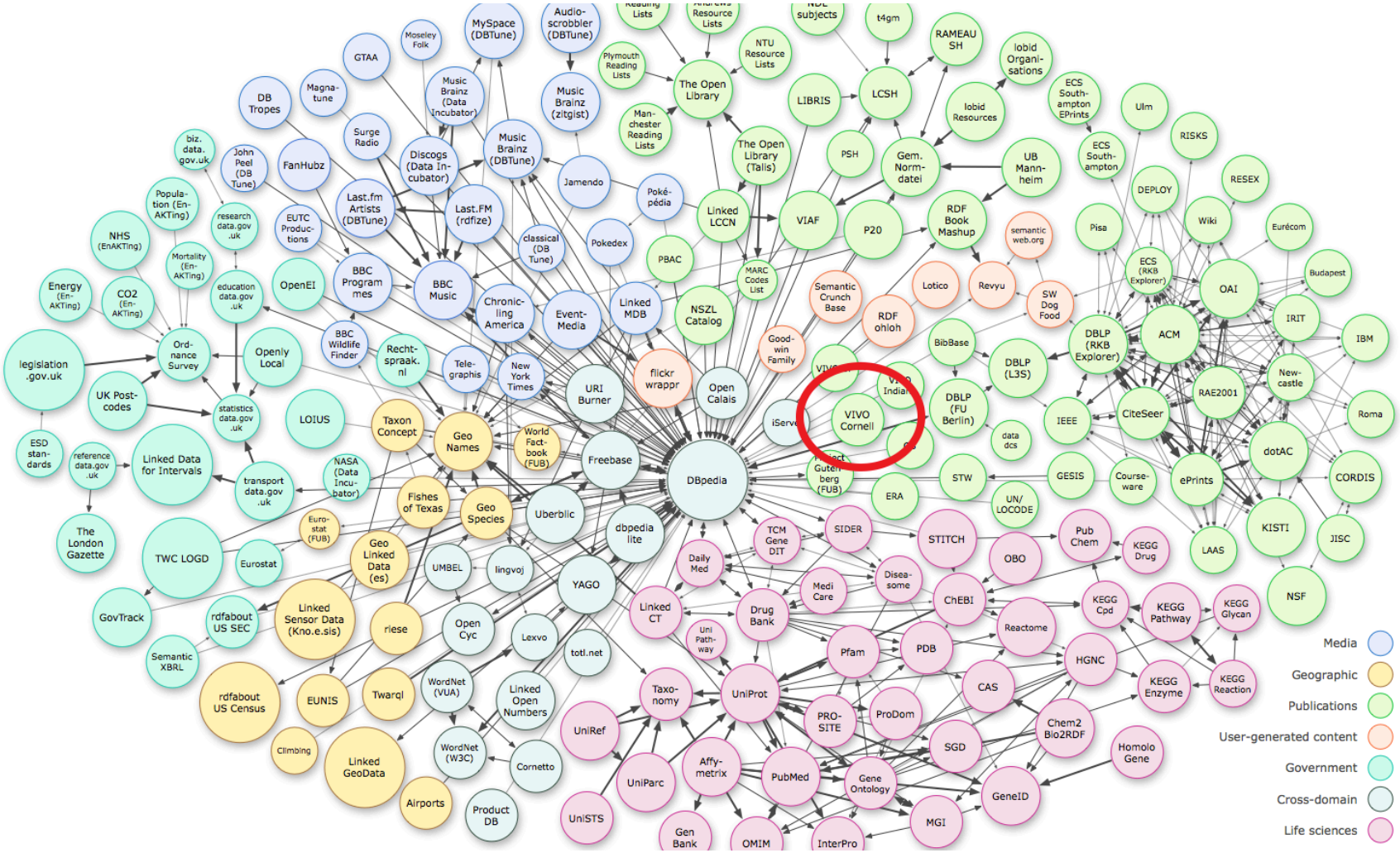
COMM-2580: Six Pretty Good Books: Explorations in Social Science – Fall 2014

COMM-4580: The Science of Social Behavior – Fall 2014

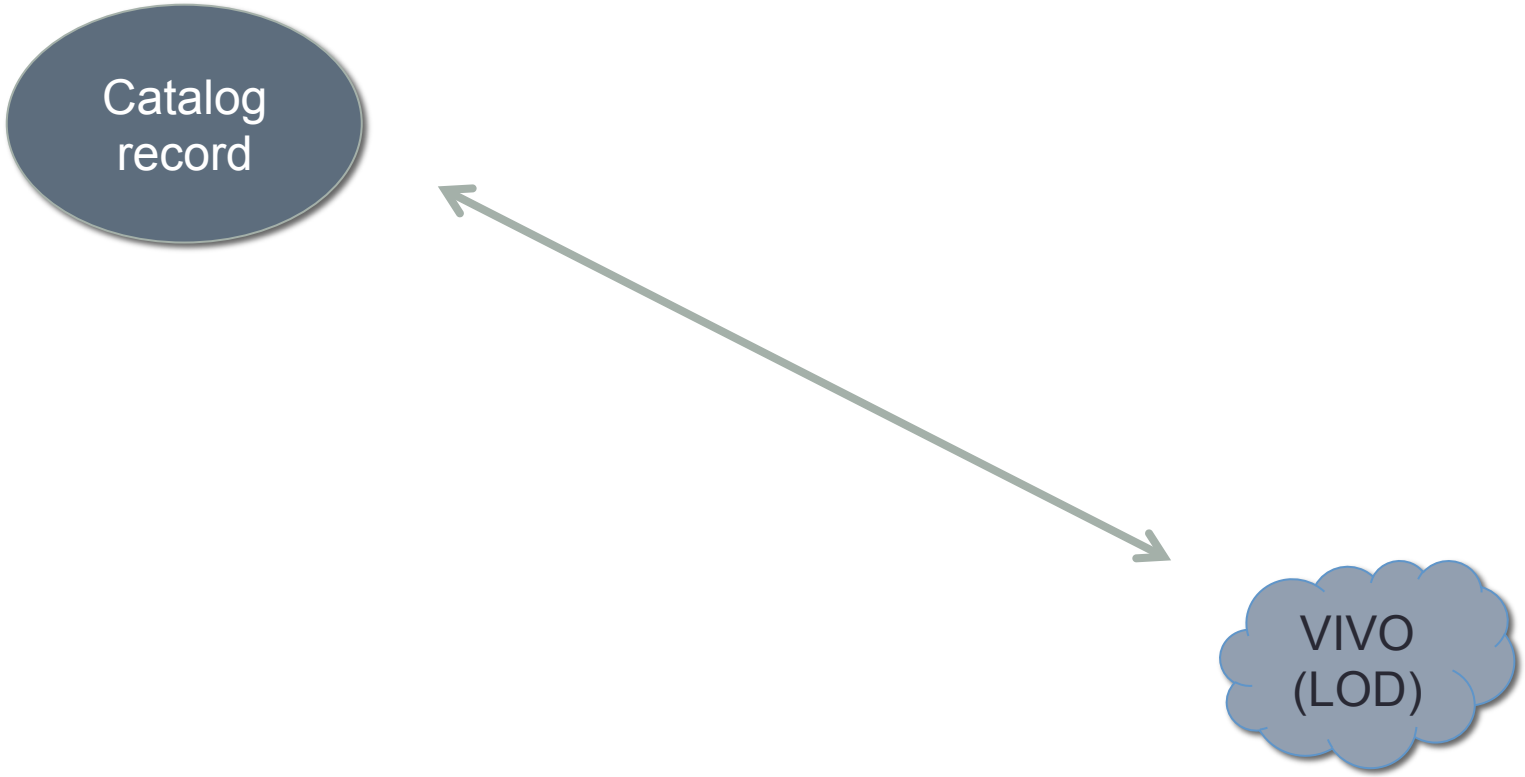
VIVO data is Linked Open Data


```
<rdf:Description rdf:about="http://vivo.cornell.edu/individual/individual23258">
  <hr:EndowCharCodeDesc>HELEN L. CARR PROFESSORSHIP IN HUMAN ECOLOGY</hr:EndowCharCodeDesc>
  <rdf:type rdf:resource="http://vivo.library.cornell.edu/ns/0.1#CornellEmployee"/>
  <hr:WorkingTitle>The Helen L. Carr Professor of</hr:WorkingTitle>
  <hr:PrefName>Ceci, Stephen J</hr:PrefName>
  <vivo:featuredIn rdf:resource="http://vivo.cornell.edu/individual/NIHawardsmorethan2milliontoCornellforstudyingwomeninsciences"/>
  <vivoc:memberOfGraduateField rdf:resource="http://vivo.cornell.edu/individual/individual32047"/>
  <vivo:featuredIn rdf:resource="http://vivo.cornell.edu/individual/TheMathematicsofSexassertsthatwomenoptoutofmathfieldsforflexibility"/>
  <rdf:type rdf:resource="http://vivoweb.org/ontology/core#FacultyMember"/>
  <vitro-public:mainImage rdf:resource="http://vivo.cornell.edu/individual/n37018"/>
  <obo:RO_0000053 rdf:resource="http://vivo.cornell.edu/individual/COMM-1840-FA12-2"/>
  <hr:preferredFirstName>Stephen</hr:preferredFirstName>
  <vivoc:affiliatedWithAsCornellFacultyMember rdf:resource="http://vivo.cornell.edu/individual/CornellPopulationProgramCPP"/>
  <vivoc:administrativeResponsibilities>I co-direct the NIH center: Cornell Institute for Women in Science (CIWS). I lead a large lab (chair o
  <hr:preferredLastName>Ceci</hr:preferredLastName>
  <rdf:type rdf:resource="http://vivo.library.cornell.edu/ns/0.1#CornellAcademicEmployee"/>
  <vivo:researchOverview>My current activities entail three lines of research, generating a  number of new publications during 2012. These thr
  <hr:Address1>M Van Rensselaer Hall, Room G80</hr:Address1>
  <rdf:type rdf:resource="http://vivo.cornell.edu/ns/mannadditions/0.1#CornellFaculty"/>
  <hr:Name>Ceci, Stephen John</hr:Name>
  <vivoc:memberOfGraduateField rdf:resource="http://vivo.cornell.edu/individual/individual31948"/>
  <vivoc:affiliatedWithAsCornellFacultyMember rdf:resource="http://vivo.cornell.edu/individual/individual11477"/>
  <cce:hasPrimarySpecializationArea rdf:resource="http://vivo.cornell.edu/individual/CCE_FPOW_5.1Y2014K256"/>
  <vivoc:CornellemailnetId>sjc9@cornell.edu</vivoc:CornellemailnetId>
  <hr:preferredNameString>Stephen J. Ceci</hr:preferredNameString>
</rdf:Description>
<rdf:Description rdf:about="http://vivo.cornell.edu/individual/n117889">
  <rdf:type rdf:resource="http://vivoweb.org/ontology/core#Relationship"/>
  <rdf:type rdf:resource="http://purl.obolibrary.org/obo/BFO_0000020"/>
  <rdf:type rdf:resource="http://purl.obolibrary.org/obo/BFO_0000001"/>
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
  <rdf:type rdf:resource="http://vivoweb.org/ontology/core#Authorship"/>
  <rdf:type rdf:resource="http://purl.obolibrary.org/obo/BFO_0000002"/>
  <vivo:relates rdf:resource="http://vivo.cornell.edu/individual/individual23258"/>
</rdf:Description>
<rdf:Description rdf:about="http://vivo.cornell.edu/individual/n1760489">
  <rdf:type rdf:resource="http://vivoweb.org/ontology/core#Relationship"/>
```

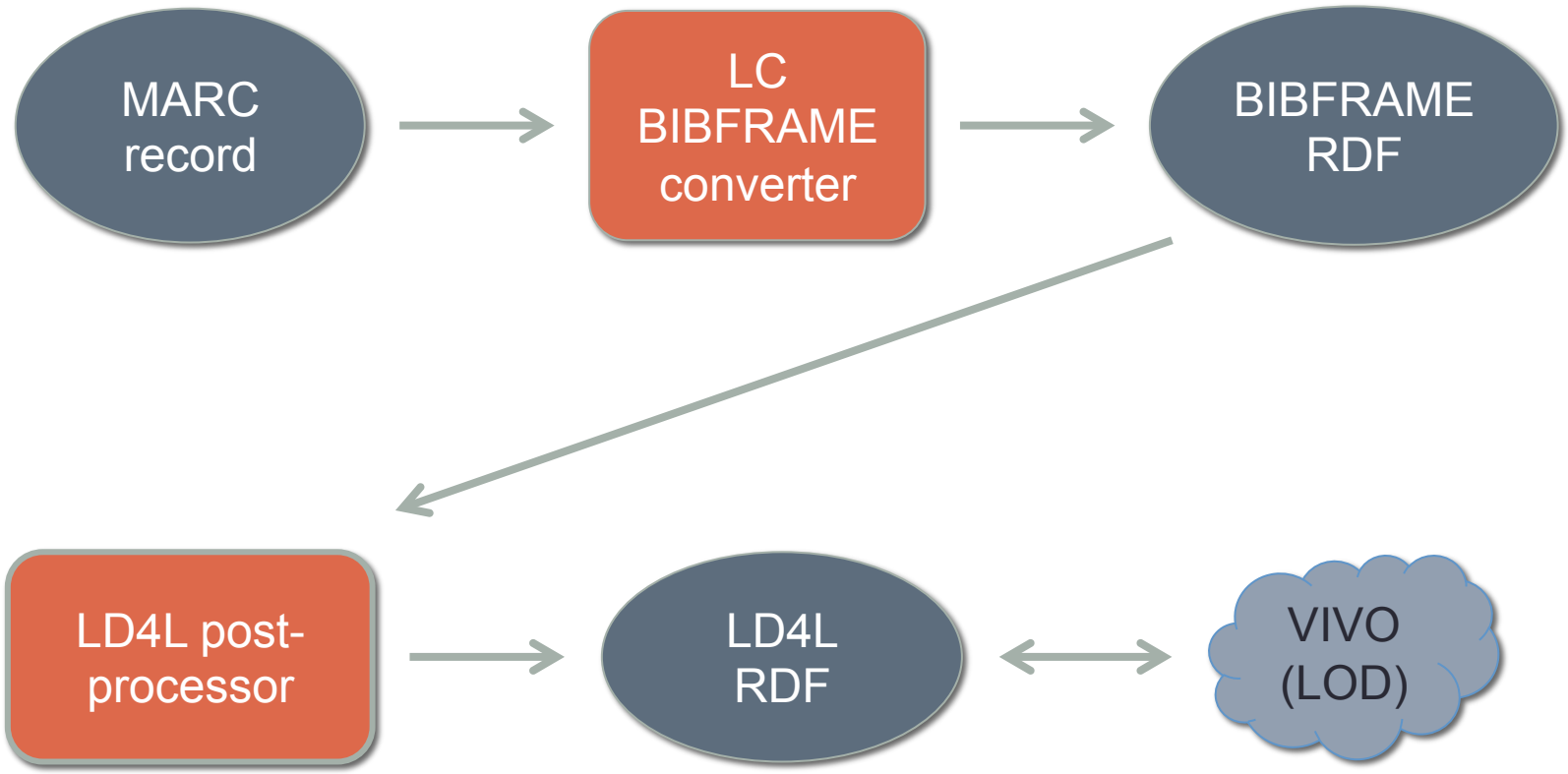
VIVO in the Linked Open Data cloud



Getting to Linked Open Data



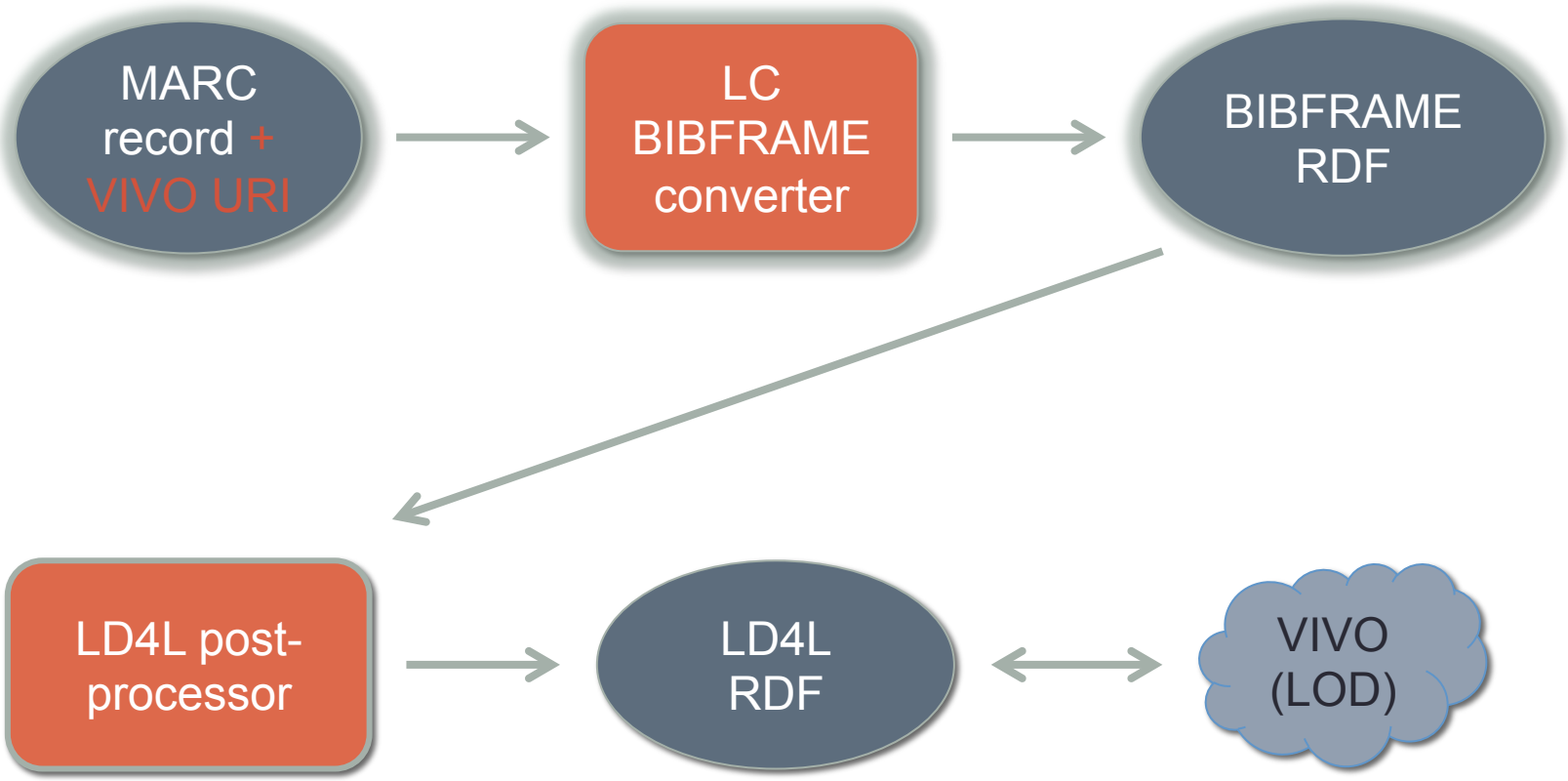
Stages of data transformation



Thesis record enhanced with VIVO URI

```
035        ‡a (CUThesis)181338774
035        ‡a (OCoLC)900927375
035        ‡a 8793268
040        ‡a NIC ‡c NIC
100 1      ‡a Royer, Caisa Elizabeth.
245 1  4   ‡a The Suggestive Influence Of Lineups On Memory ‡h [electronic resource] / ‡b A
           Comparative Study Of Children And Adults ‡c by Caisa Elizabeth Royer.
260        ‡c c2014
300        ‡a 59 pages:
502        ‡b M.A. ‡c Cornell University ‡d August, 2014
504        ‡a Includes bibliographical references.
520 3      ‡a This study was designed to test whether making a selection from a target-absent
           lineup could alter memory for the actual perpetrator. Three different lineup
           presentations were used: simultaneous, sequential, and elimination. In addition,
           three different memory tests were used in which participants saw two of the
           following three faces: the actual thief, a suspect from the original lineup, and a
           novel suspect. The sequential lineup resulted in the fewest initial false
           identifications, while the simultaneous lineup appeared to protect the most against
           memory alteration. Logistic regression models were utilized to estimate the
           likelihood of making a particular identification during the memory test.
           Participants were more likely to identify the innocent original suspect when they
           were older, non-white, and were shown the sequential lineup. Participants were more
           likely to make an accurate identification of the thief when they were younger and
           had made an identification from the initial lineup.
653        ‡a Lineup
653        ‡a Misinformation Effect
653        ‡a Memory
700 1      ‡a Ceci, Stephen John ‡e thesis advisor ‡0 http://vivo.cornell.edu/individual
           /individual23258
856 4  1   ‡u http://hdl.handle.net/1813/38774 ‡z Connect to full text. Access to electronic
           version of some theses may be restricted.
899        ‡a CUThesis
948 0      ‡a 20150128 ‡b i ‡d batch ‡e lts
948 1      ‡a 20150128 ‡b s ‡d batch ‡e lts ‡f ebk
```

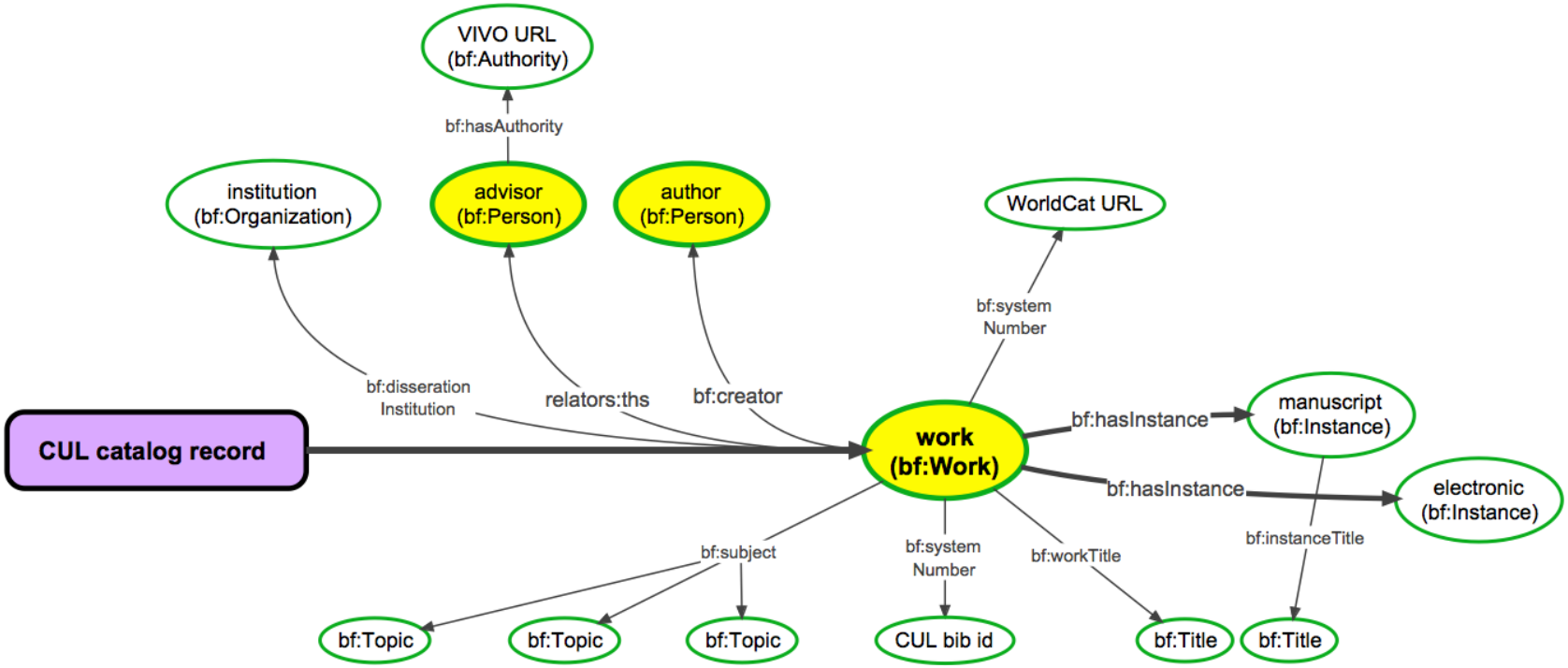
Stages of data transformation



BIBFRAME RDF


```xml
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:relators="http://id.loc.gov/vocabulary/relators/"
         xmlns:bf="http://bibframe.org/vocab/"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   <bf:Work rdf:about="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268">
      <rdf:type rdf:resource="http://bibframe.org/vocab/Text"/>
      <bf:dissertationInstitution rdf:resource="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268organization5"/>
      <bf:workTitle rdf:resource="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268title6"/>
      <bf:title xml:lang="x-bf-sort">Suggestive Influence Of Lineups On Memory</bf:title>
      <bf:creator rdf:resource="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268person8"/>
      <relators:ths rdf:resource="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268person9"/>
      <bf:dissertationDegree>M.A.</bf:dissertationDegree>
      <bf:dissertationYear>August, 2014</bf:dissertationYear>
      <bf:language rdf:resource="http://id.loc.gov/vocabulary/languages/eng"/>
      <bf:subject rdf:resource="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268topic14"/>
      <bf:subject rdf:resource="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268topic15"/>
   </bf:Work>
   <bf:Organization rdf:about="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268organization5">
      <bf:label>Cornell University</bf:label>
   </bf:Organization>
   <bf:Title rdf:about="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268title6">
      <bf:titleValue>The Suggestive Influence Of Lineups On Memory</bf:titleValue>
      <bf:subtitle>A Comparative Study Of Children And Adults</bf:subtitle>
   </bf:Title>
   <bf:Person rdf:about="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268person8">
      <bf:label>Royer, Caisa Elizabeth.</bf:label>
   </bf:Person>
   <bf:Person rdf:about="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268person9">
      <bf:label>Ceci, Stephen John</bf:label>
      <bf:hasAuthority rdf:resource="http://vivo.cornell.edu/individual/individual23258"/>
   </bf:Person>
   <bf:Topic rdf:about="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268topic14">
      <bf:label>Misinformation Effect</bf:label>
   </bf:Topic>
   <bf:Topic rdf:about="http://ld4l.library.cornell.edu/ld4l-vitro/individual/8793268topic15">
      <bf:label>Memory</bf:label>
   </bf:Topic>
</rdf:RDF>
```

BIBFRAME RDF



Limitations of BIBFRAME

- Limitations of the BIBFRAME ontology and the way records are converted to BIBFRAME create the need for a post-processor to get us to linked data.
- Caveat: BIBFRAME is still evolving, and may undergo significant changes based on LD4L recommendations
- Authorities vs. Real World Objects
- Strings vs. Things

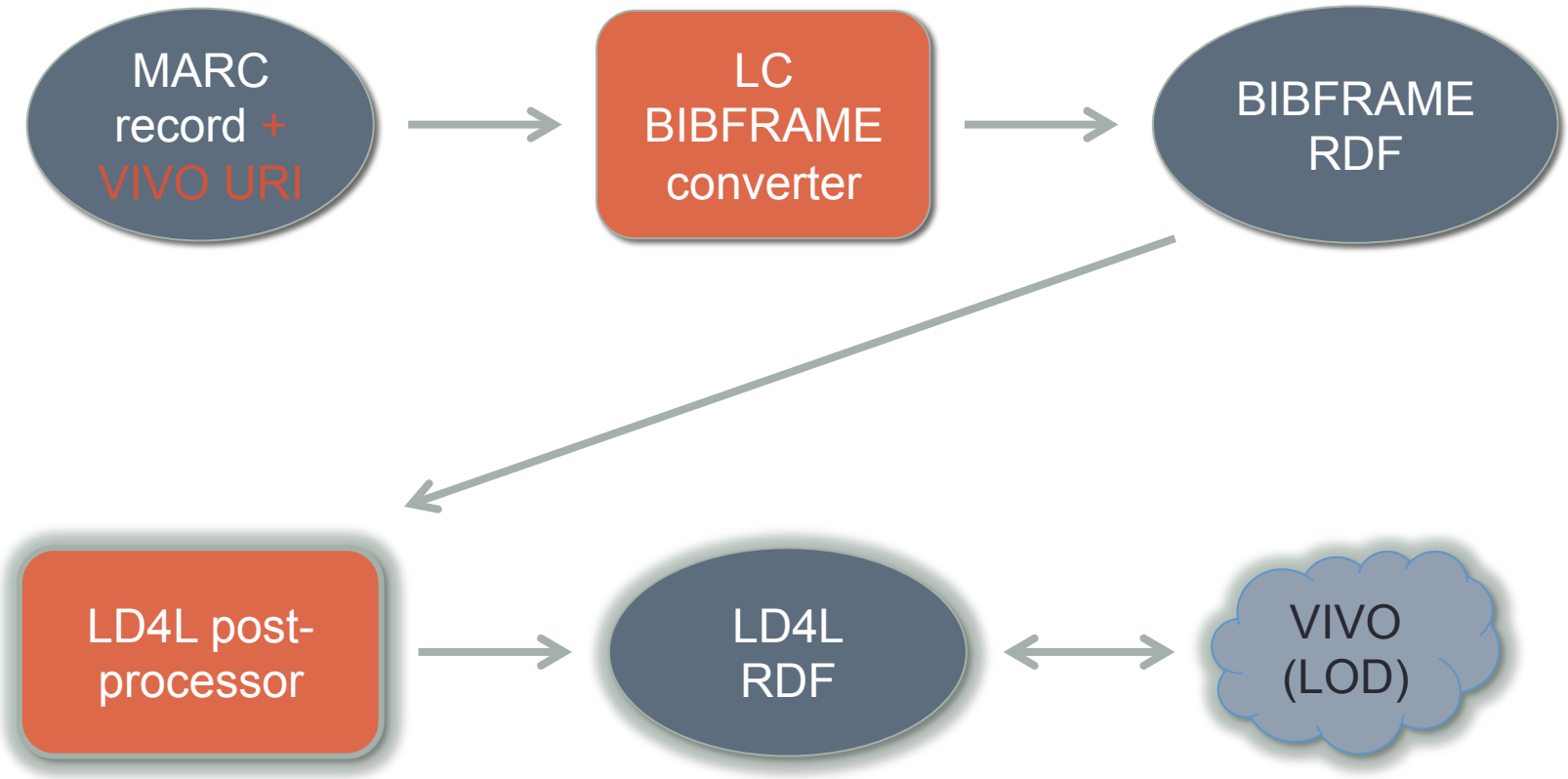
Authorities vs. Real World Objects

- Authorities vs. Real World Objects (RWOs)
- Properties of BIBFRAME Person: authorityAssigner, authoritySource, hasAuthority
- Properties and relationships of FOAF Person: name, knows, made, birthday, age
- BIBFRAME Authorities can’t connect to other Linked Open Data expressed as FOAF and other RWOs

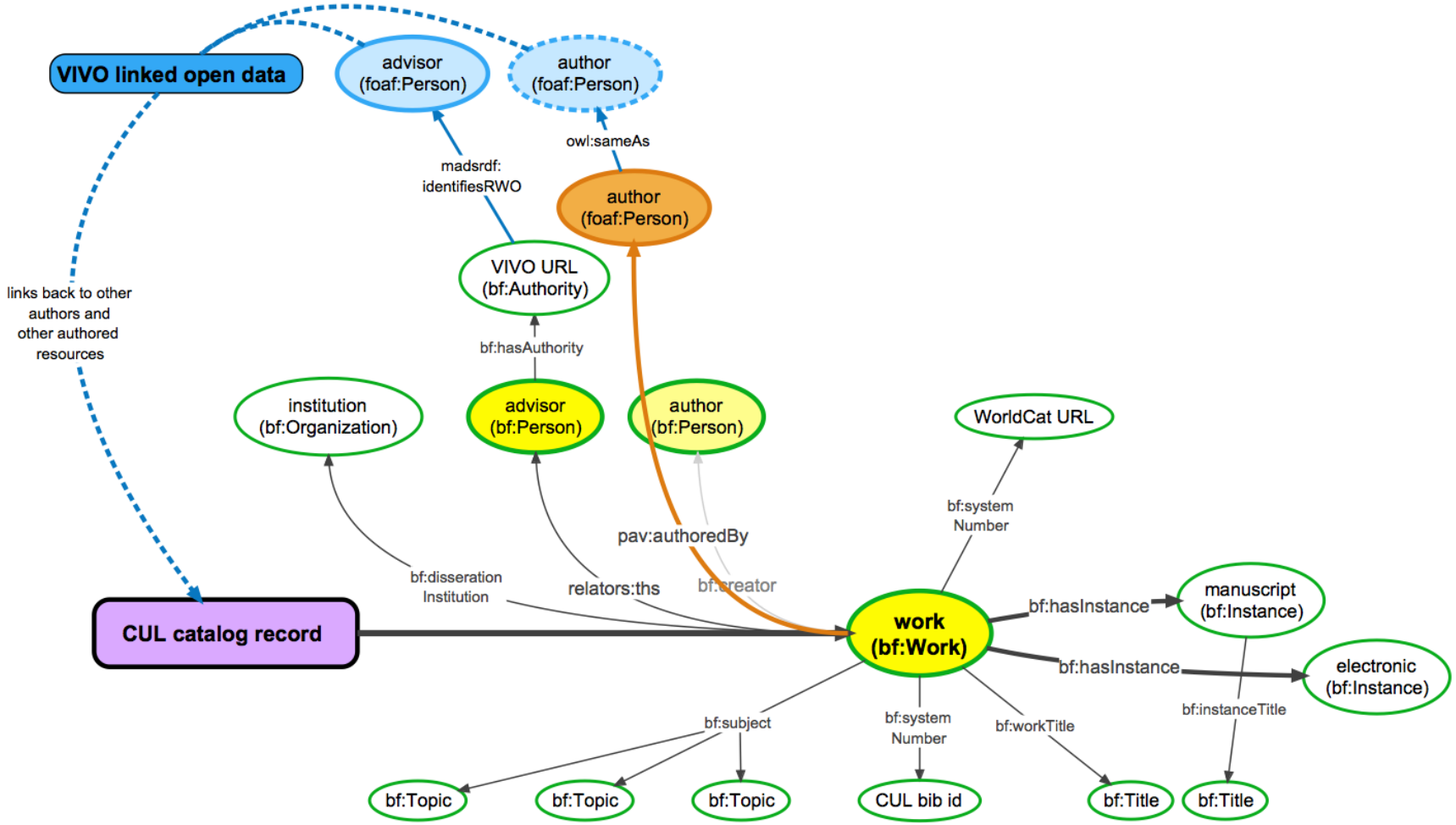
Strings vs. Things

- Strings:
  - “Kennedy, Charles W. (Charles William), 1882-1969.”
  - “Kennedy, Charles William.”
  - “Kennedy, Charles W., 1882-1969.”
- Thing: http://vivo.cornell.edu/individual/individual23258
- BIBFRAME RDF still makes heavy use of strings
- Strings are a dead-end from an LOD perspective
- The post-processor needs to turn these strings into Things that can be looked up and connected to other data in the entity resolution phase.

Stages of data transformation



LD4L Linked Data fragment



Examples of UC2 demo post-processing

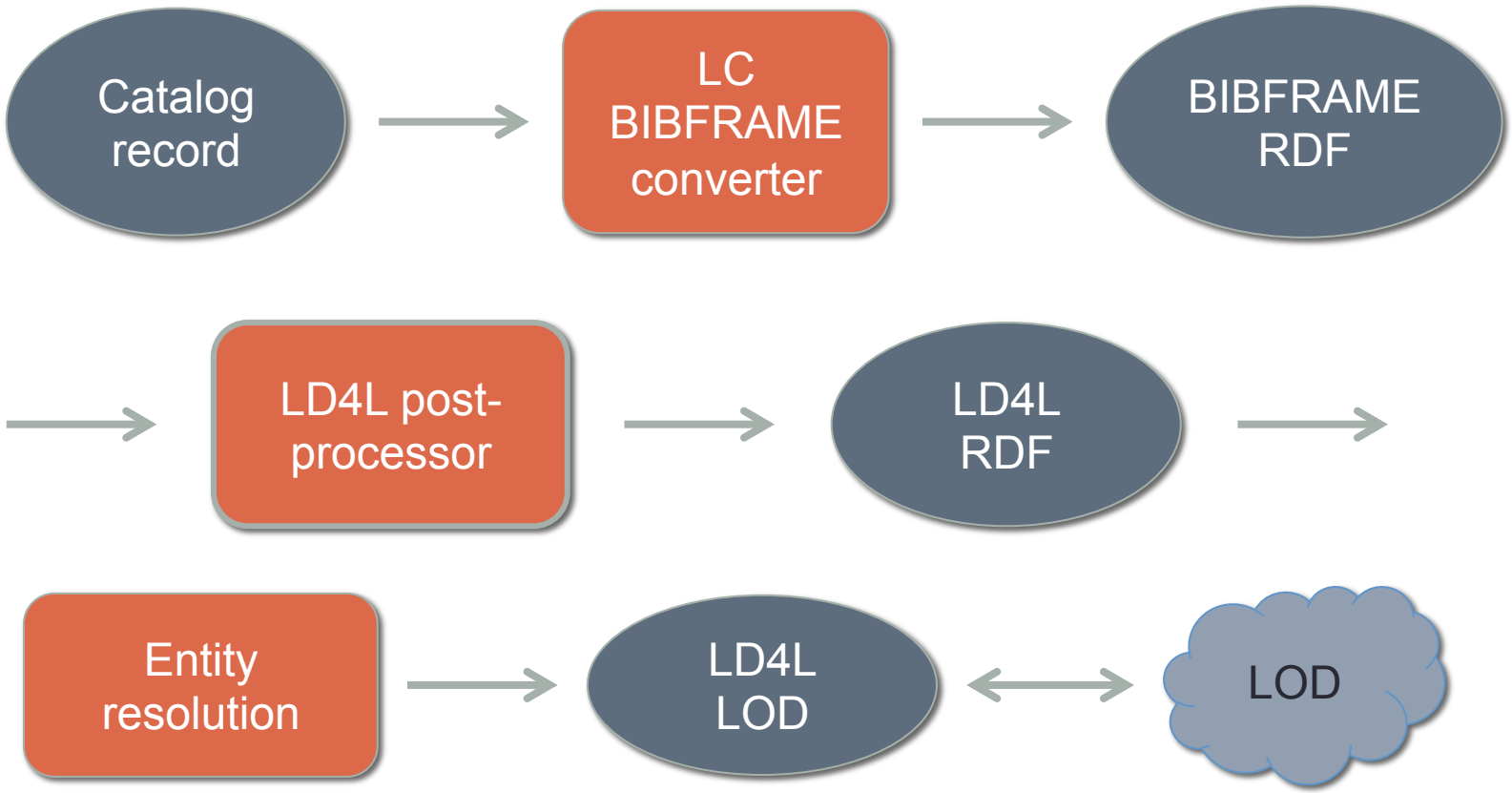
- Create FOAF Persons from BIBFRAME Persons
- Parse strings containing multiple pieces of data:
  - <bf:label>Darwin, Charles, 1809-1882.</bf:label>
- De-duping: LC converter creates duplicate works, persons, topics, etc. for each record.
  - E.g., if two theses have the same advisor, two distinct BIBFRAME Persons are generated by the converter
  - In this case de-duping is easy, since we have the VIVO URIs as identifiers
  - In the general case, it’s not easy

LD4L data in Vitro

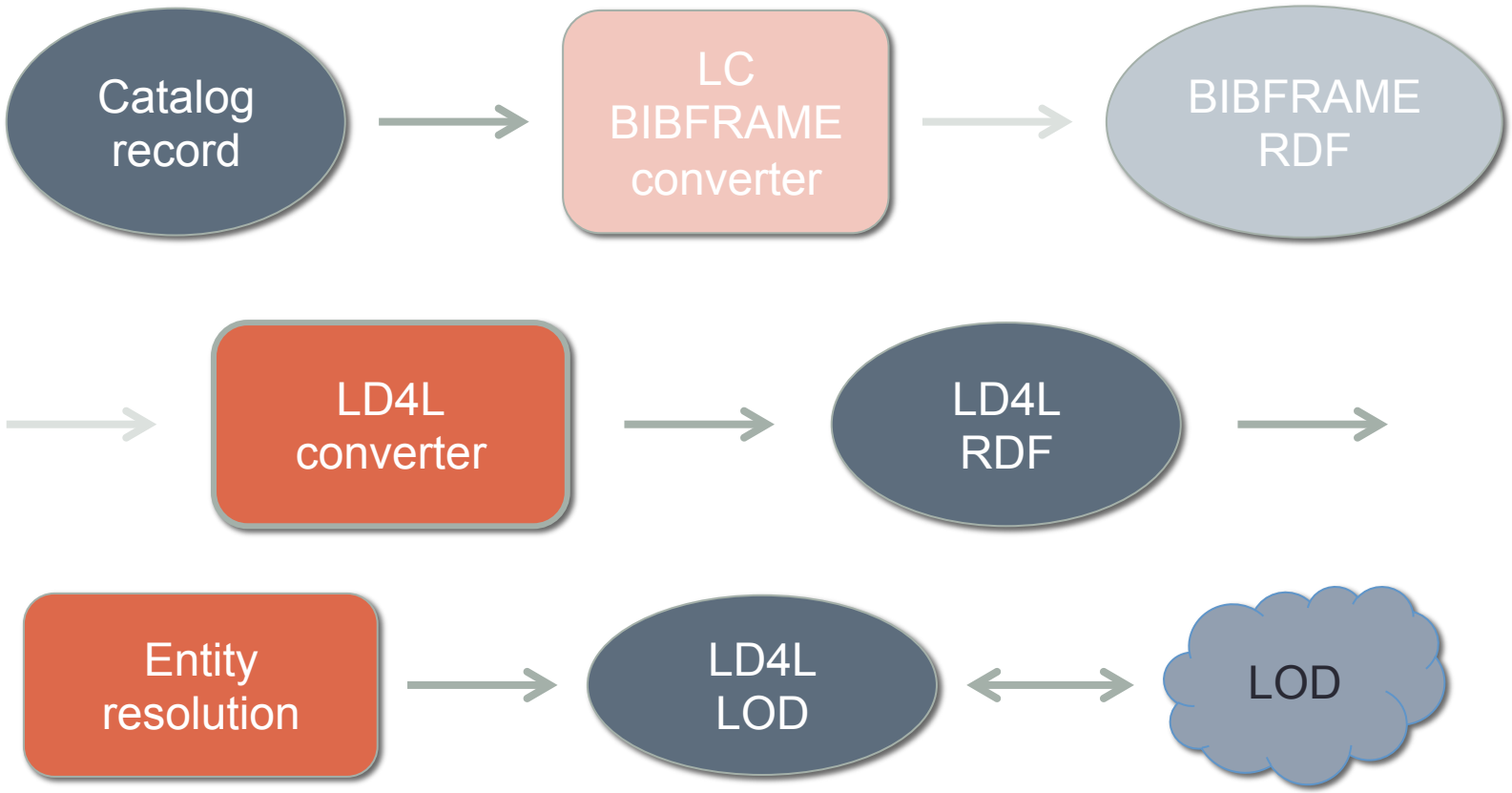
- Live at ld4l.library.cornell.edu/ld4l-vitro
- Simulates catalog records augmented with LOD
- Links to external LOD – in this case, VIVO
- Then VIVO brings us:
  - Back to additional catalog records
  - Out to the LOD cloud – for example:
    - Global identifiers
    - Co-author networks

FULL-SCALE CONVERSION, POST-PROCESSING, AND ENTITY RESOLUTION

From catalog to LOD



From catalog to LOD



Step 1. MARC to RDF conversion

- Harvard's 13 million records => 1 billion RDF triples
- ~ 8 million Cornell records => ~ 600 million triples
- VIVO: ~ 15 million triples
- Batch processing, multiple concurrent processors
- Data storage becomes an issue at this scale.
- LD4L (Jim) is investigating the performance of several triplestores against data of this magnitude
  - Data ingest
  - Queries
  - Building a Solr index
  - Inferencing

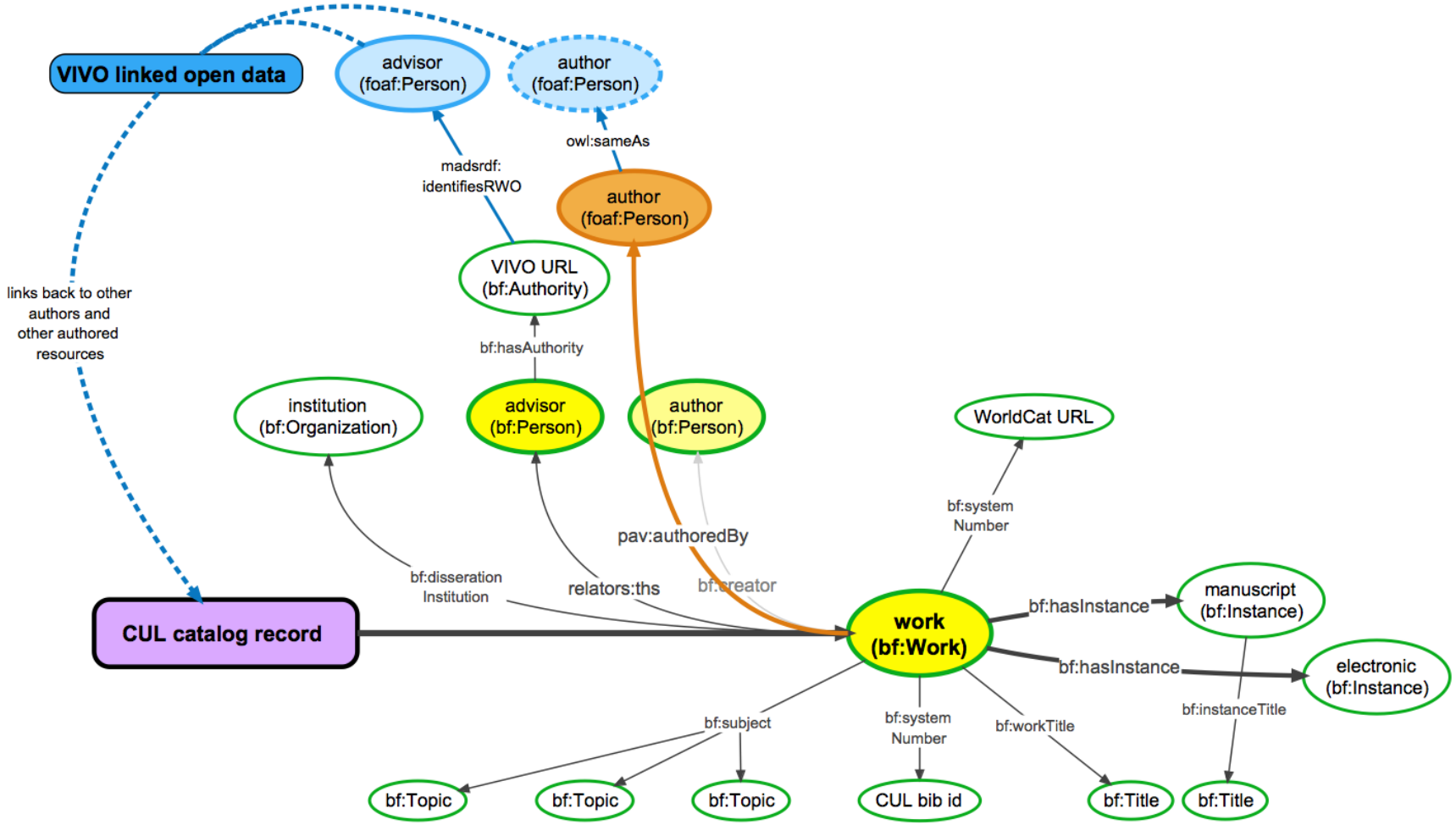
Step 2. Post-processing

- From simple post-processing on a limited set of homogeneous data to the full library catalog and the full variety of records
- Will use the final version of the LD4L ontology as a target
- Will use the BIBFRAME converter as an intermediate step
- Challenges:
  - Strings to Things requires parsing
    - <bf:label>Darwin, Charles, 1809-1882.</bf:label>
  - De-duping: LC converter creates duplicate works, persons, topics, etc. for each record

Step 3. Entity resolution (lookup)

- Once we have Things instead of strings, we want to connect those to other Things (or other URIs for the same Thing) in the LOD cloud for data aggregation and enrichment in both directions.
- Although the Cornell Library wants to add more URIs to records, it won't always be that simple.
- We need an entity resolution step to do lookups of global identifiers in data sources (OCLC, VIAF, ISNI, ORCID, etc.) to find more data about the entities in our data

LD4L Linked Data fragment



Additional connections to global identifiers for works, organizations, and people