CISER

Cornell Institute for Social and Economic Research
A LEADER IN SOCIAL SCIENCE DATA AND COMPUTING

Improving User Access to Metadata for Public and Restricted Use US Federal Statistical Files

William C. Block
Jeremy Williams
Lars Vilhuber
Carl Lagoze
Warren Brown
John Abowd

Cornell University

CISER

Outline

- Overview of Cornell NSF-Census (NCRN) Node
- Comprehensive Extensive Data Documentation and Access Repository (CED^{2}AR)
  - Extract, Transform and Load (ETL)
  - Application Programming Interface (API)
  - User Interface (UI)
  - Coming Features of CED^{2}AR
- The Problem of Provenance
- Future NCRN Work

CISER

Summary of NCRN Problem

- Inadequate curation of secure datasets
- Inconsistent or non-existent identification
- Need for selective hiding of data and metadata
- Scientific Method demands solution

“Those inside can’t see out, and those outside can’t see in!”

CISER

NCRN DDI Solution at the Variable Level: <dataAccs>

```
<stdyDscr>
    <citation> [8 lines]
    <dataAccs ID="A1">
        <useStmt>
            <conditions>Public</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A2">
        <useStmt>
            <confDec>To download this dataset, the user must obtain Special Sworn Status from the United States Census Bureau.</confDec>
            <conditions>Confidential</conditions>
        </useStmt>
    </dataAccs>
    <dataAccs ID="A3">
        <useStmt>
            <confDec>You're never gonna see this data.</confDec>
            <conditions>Need to know</conditions>
        </useStmt>
    </dataAccs>
</stdyDscr>
```

Variable Level Solution (continued)

```
<var ID="V1500" dcml="0" files="F3" intrvl="discrete" name="totfam_kids" access="A1">
    <location width="12"/>
    <labl>Total Number of Children in Family</labl>
    <valrng> [2 lines]
    <sumStat type="vald">1000</sumStat>
    <sumStat type="invd">0</sumStat>
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <catgry> [3 lines]
    <varFormat schema="other" type="numeric"/>
</var>
<var ID="V1588" dcml="0" files="F3" intrvl="contin" name="totinc" access="A2">
    <location width="12"/>
    <labl>Total Personal Income</labl>
    <valrng> [2 lines]
    <sumStat type="vald">240</sumStat>
    <sumStat type="invd">760</sumStat>
    <sumStat type="min">-278.739</sumStat>
    <sumStat type="max">39515.631</sumStat>
    <sumStat type="mean">1861.779</sumStat>
    <sumStat type="stdev">4015.033</sumStat>
    <varFormat schema="other" type="numeric"/>
</var>
```

CISER

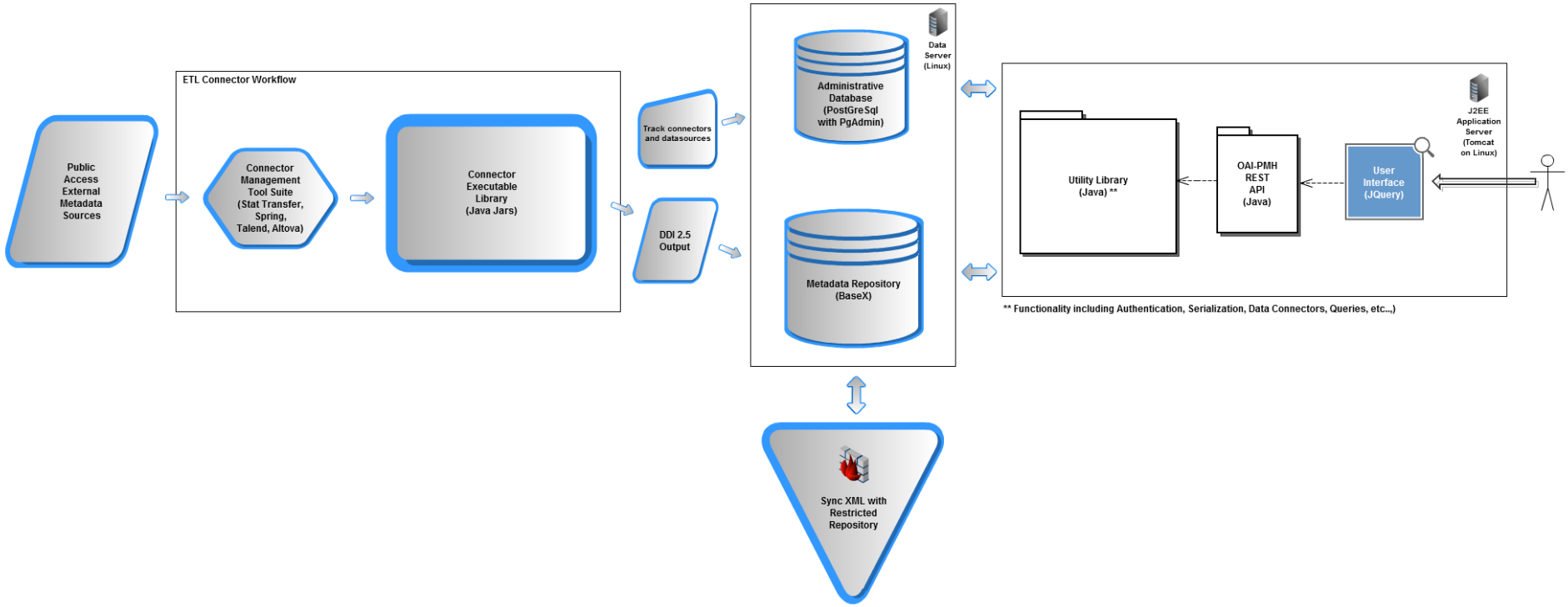
No DDI Solution at the Level of a Value Label

```
<var ID="V1588" dcml="0" files="F3" intrvl="contin" name="totinc" access="A1">
    <location width="12"/>
    <labl>Total Personal Income</labl>
    <catgry>
        <catValu>0</catValu>
        <labl>5-25k</labl>
    </catgry>
    <catgry>
        <catValu>1</catValu>
        <labl>25-75k</labl>
    </catgry>
    <catgry>
        <catValu>2</catValu>
        <labl>75-125k</labl>
    </catgry>
    <catgry>
        <catValu>3</catValu>
        <labl>125-250k</labl>
    </catgry>
    <catgry access="A2">
        <catValu>4</catValu>
        <labl>250k+</labl>
    </catgry>
    <varFormat schema="other" type="numeric"/>
</var>
```

Small tweak to the DDI Codebook Schema would fix this.

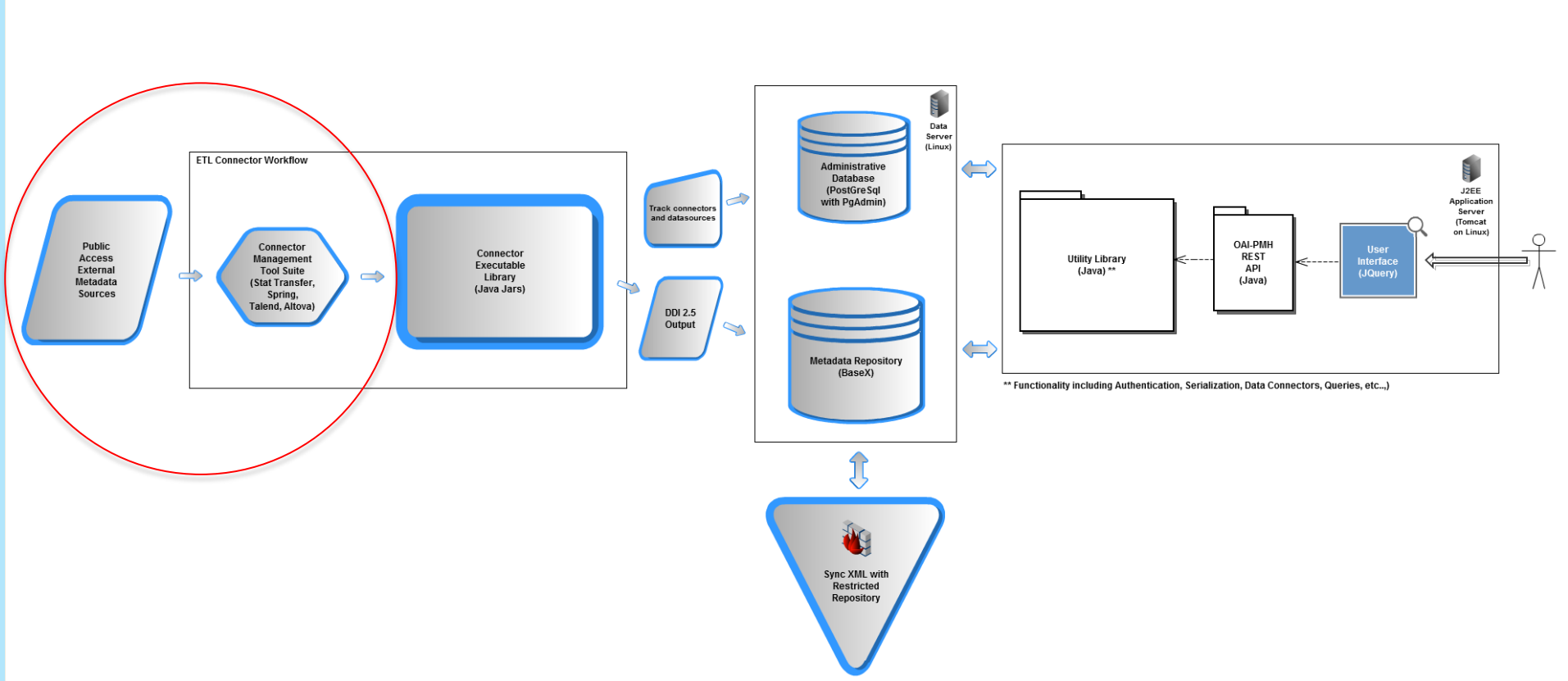
CISER

Schematic of NCRN System



CISER

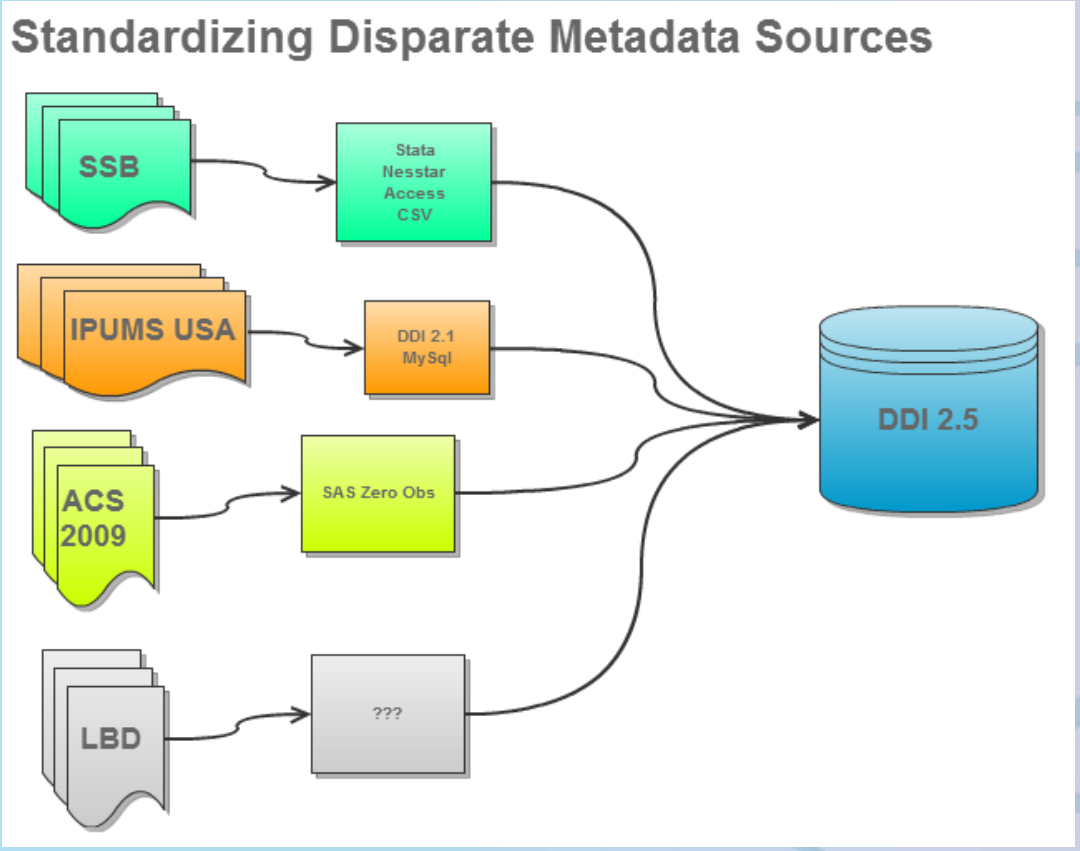
Schematic of NCRN System



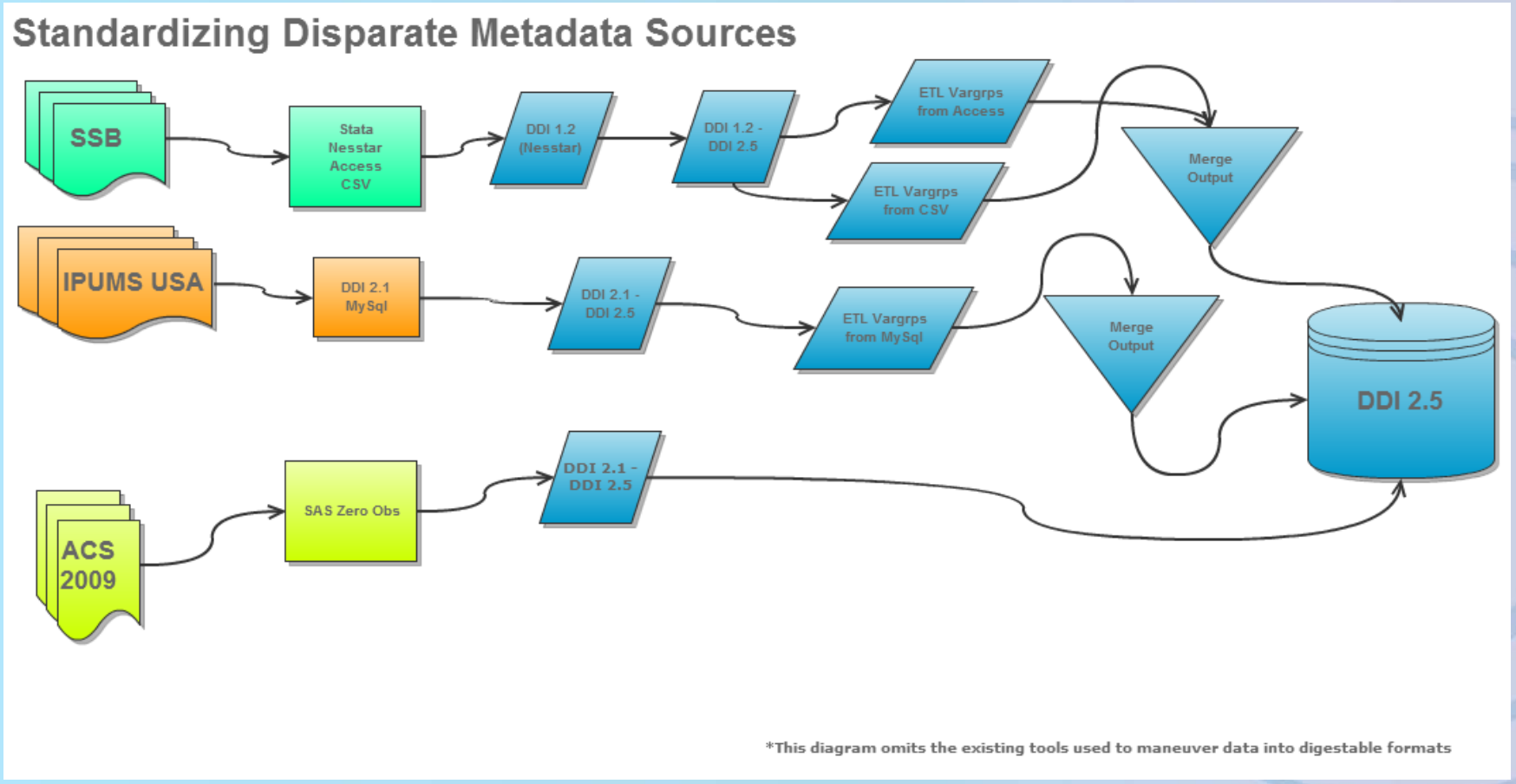
CISER

The challenge of ingesting disparate data

- Format disparity
- Schema disparity
- Sparceness of metadata

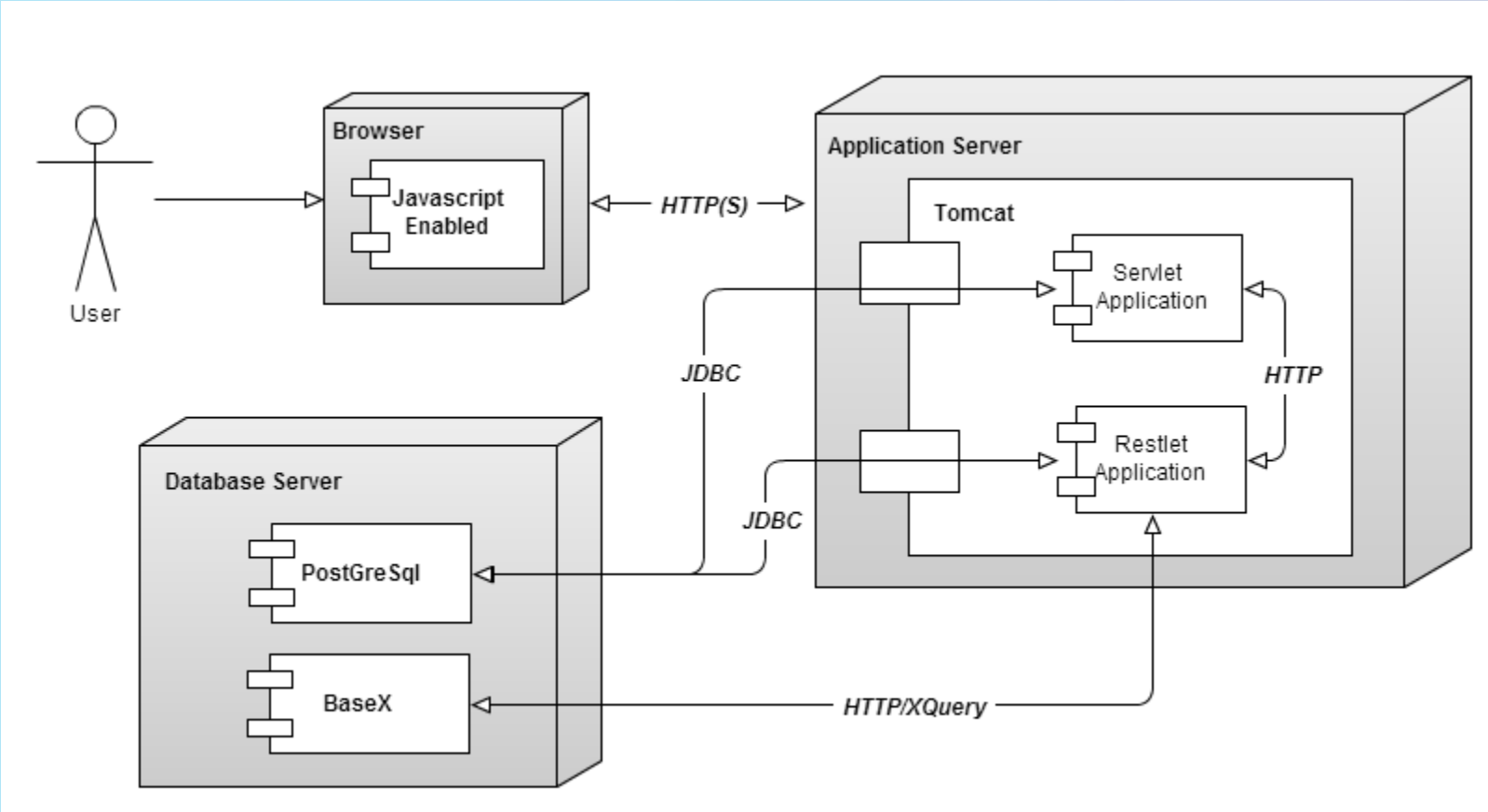


ETL Modular Approach – Building to Reuse



CISER

Component View of System



Application Programming Interface (API)

The benefits of REST

For our purposes, REST is a set of software architecture principles that leverage the intrinsic architecture of the Web.

Five Key Principles:

- Give everything (resource) an ID
- Link things together
- Uses standard methods
- Represent things (resources) in multiple ways
- Communicate so that the client and server are independent of one another

CISER

Application Programming Interface (API)

Motivation – Simplicity for greater utility

From this:

{baseurl}/rest?query=let%20$ced2ar%20:=%20collection('CED2AR')%20for%20$var%20in%20$ced2ar/codeBook/dataDscr/var%20where%20%20starts-with(lower-case($var/@name),%20%22a%22)%20return%20$var

To this:

{baseurl}/search?return=variables&where=variablename=a*

CISER

Application Programming Interface (API)

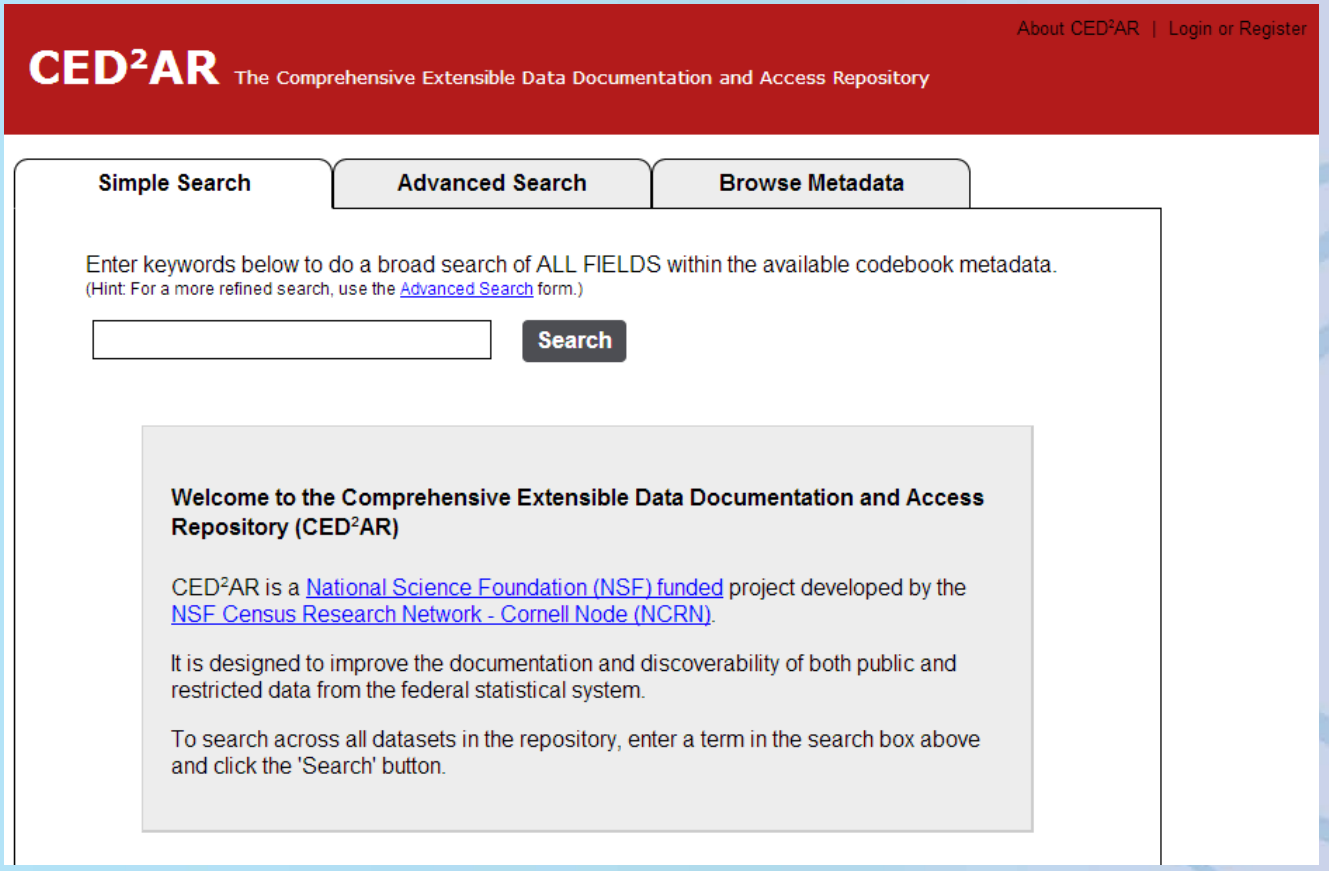
The API currently supports the following query parts:

- **Return**
  - A chosen set of fields within the DDI schema
- **Where**
  - A chosen set of supported DDI fields to filter your query by
  - And, or, 'and not'
  - Contains, starts-with, ends-with
- **Sort**
  - Descending, ascending
- **Limit**
  - (i.e.: give me results 10-50 from each codebook)

The API makes interacting with the repository easier because it abstracts away the underlying Xquery necessary to perform the query.

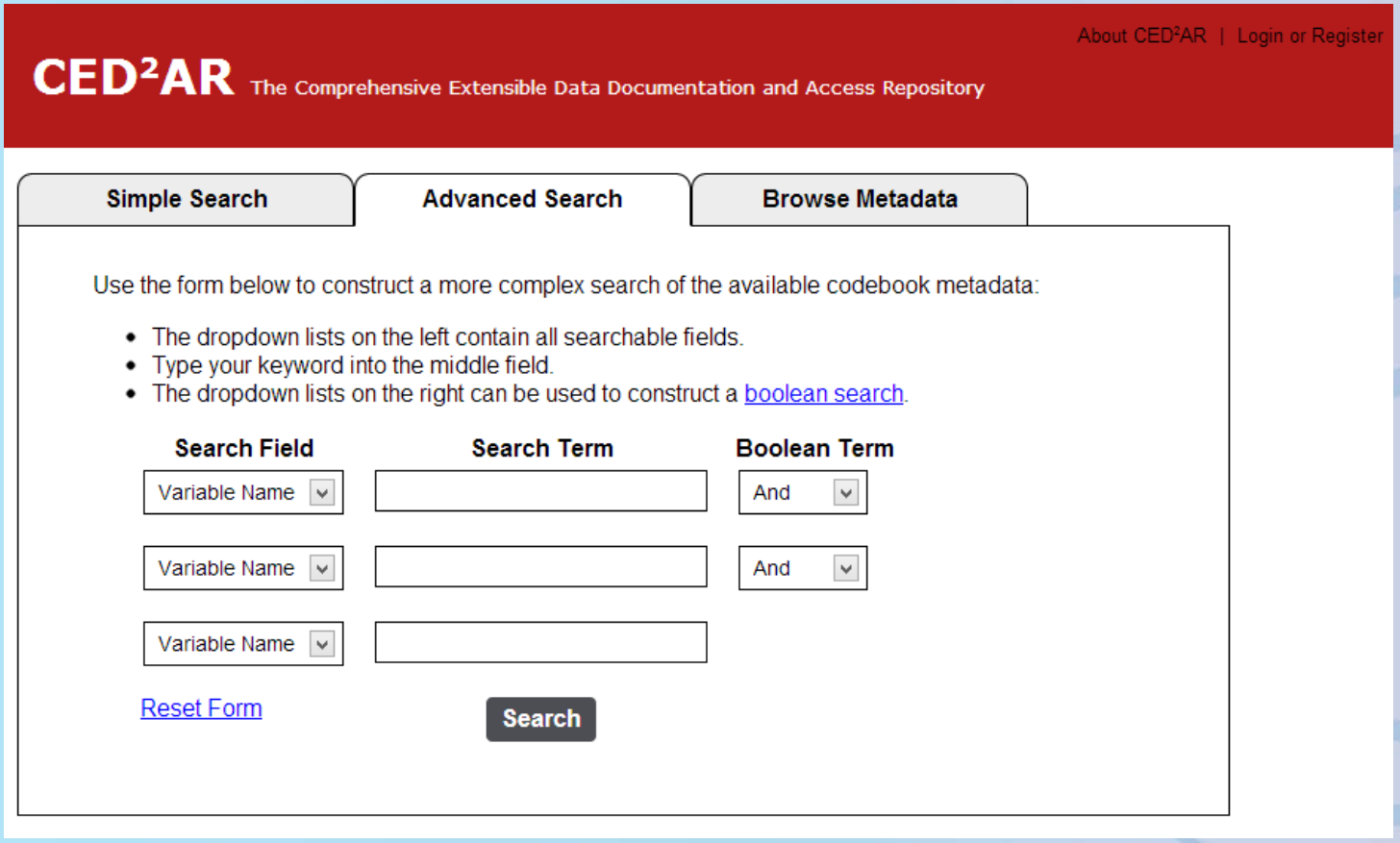
CISER

User Interface – Simple search



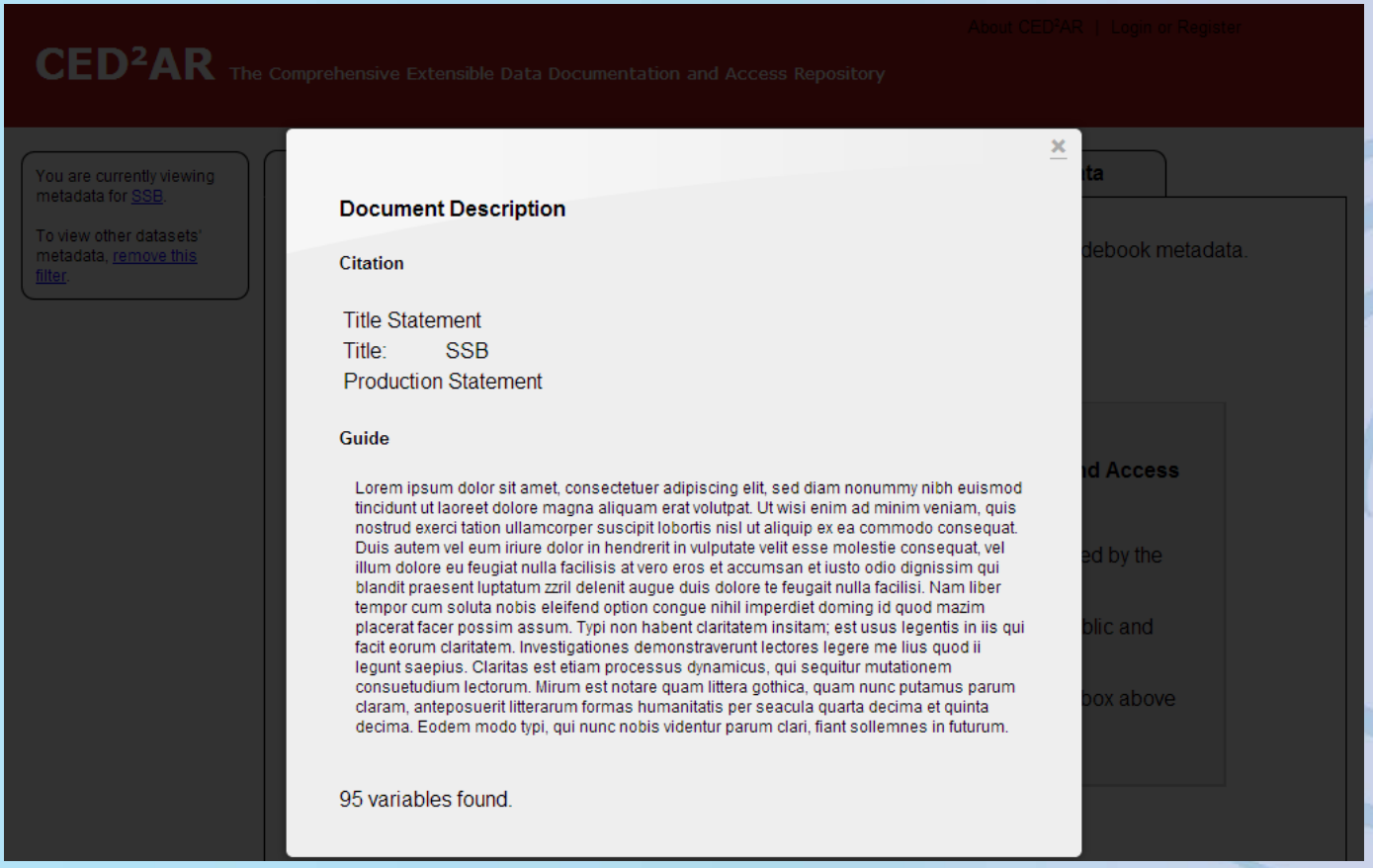
CISER

User Interface – Advanced Search



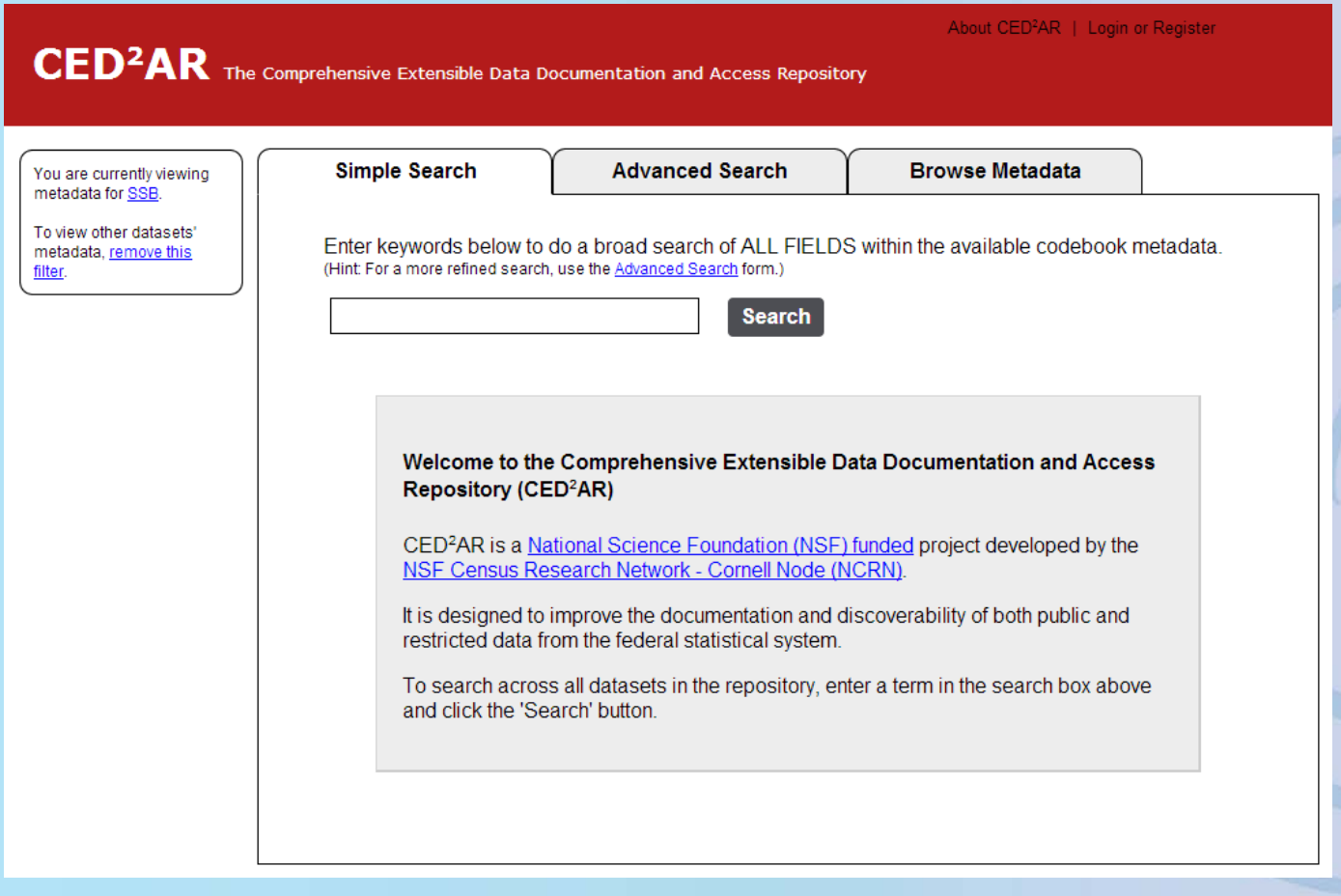
CISER

User Interface – Dataset-specific view



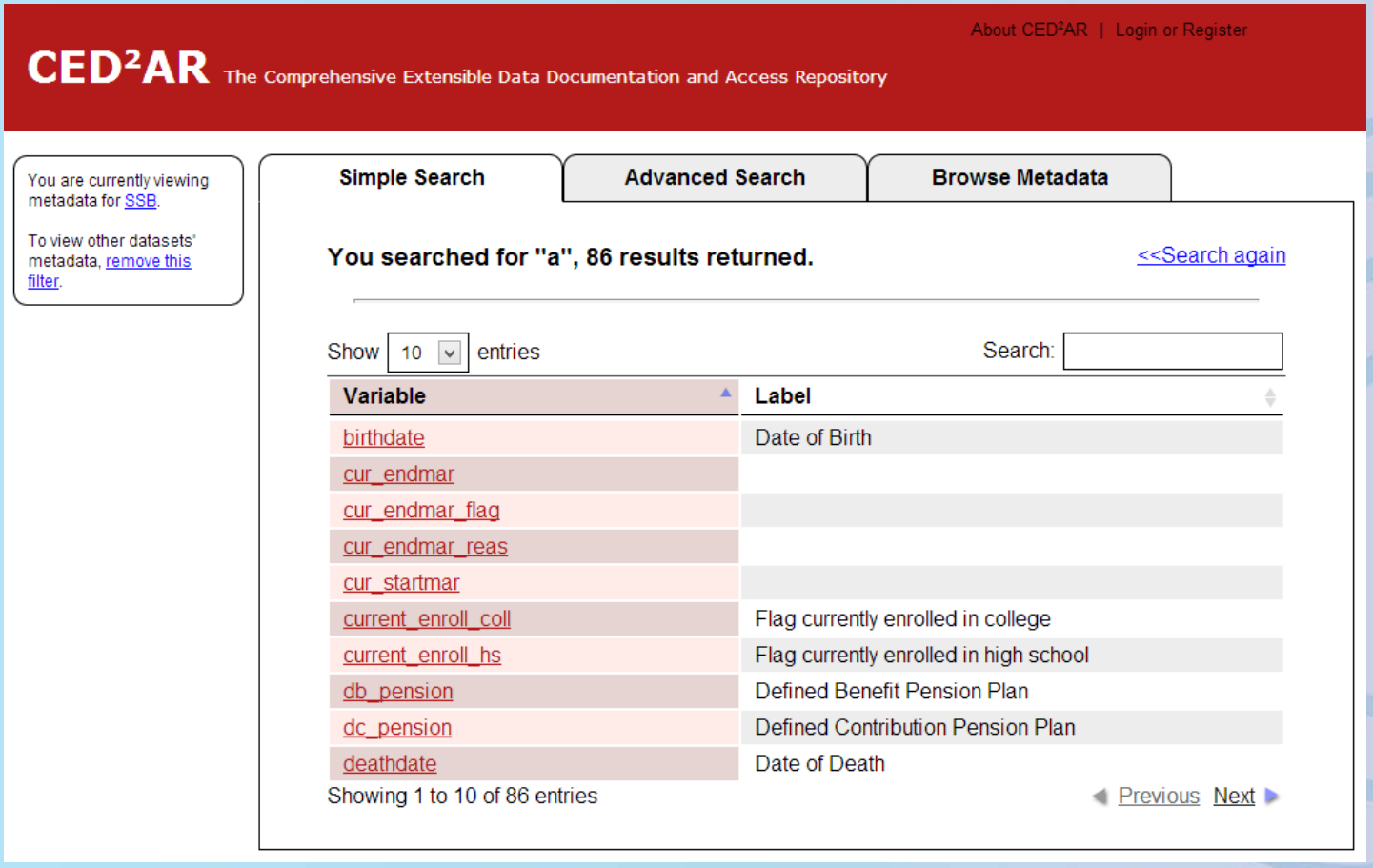
CISER

User Interface – Dataset-specific view



CISER

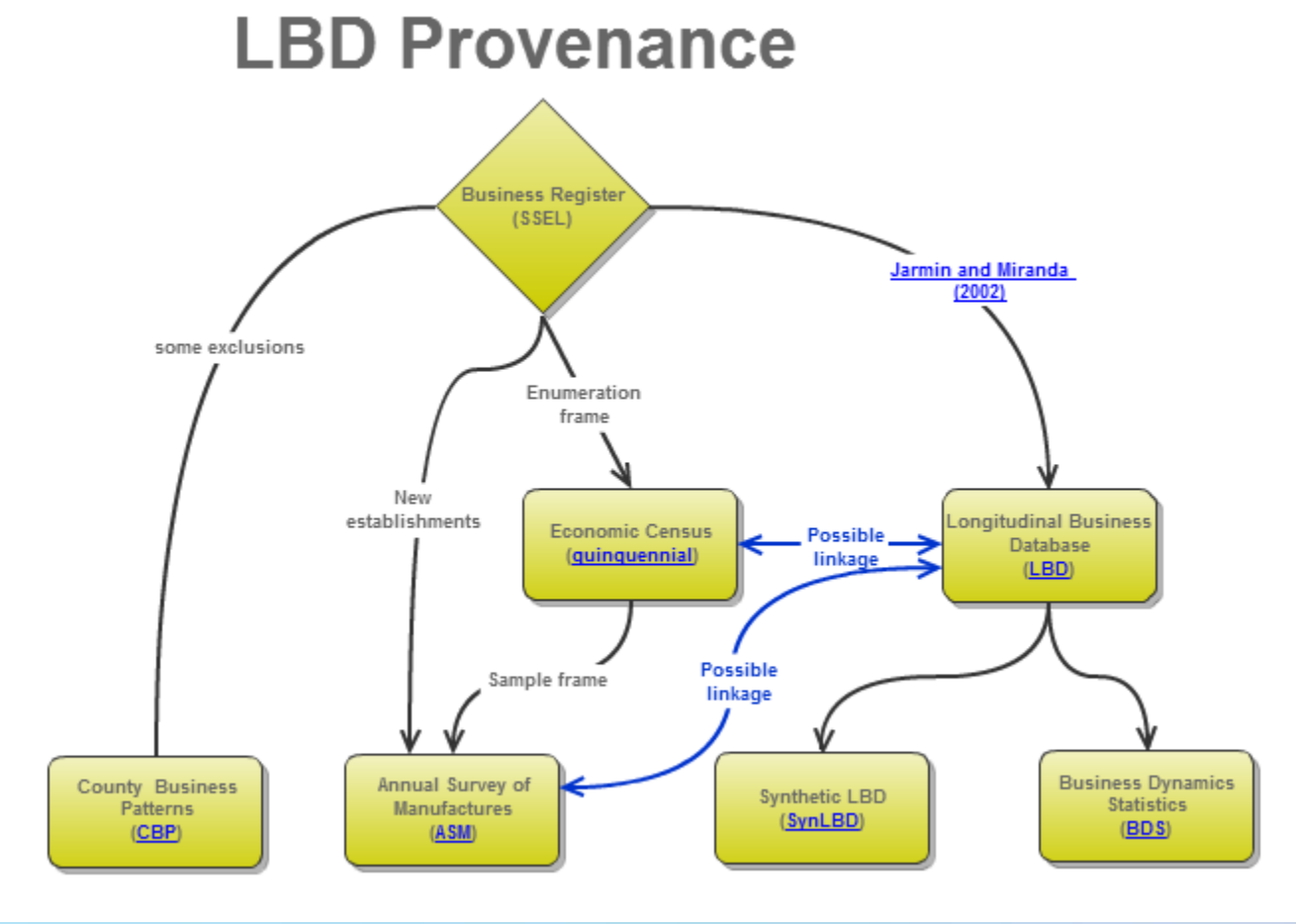
User Interface – Search results



Future Features

- Backend
  - Additional data sources
  - Generic metadata ingest mechanism from finite set of statistical applications
- API
  - OAI-PMH compliance
  - Support for more of the DDI schema
- Web Application
  - More robust filtering (i.e.: by variable group, date, etc..,)
  - Variable groups added to search results
  - Export search results
- System Wide
  - Features to support the storage, exploration and discover of the provenance of variables, datasets, etc..,

The Problem of Provenance

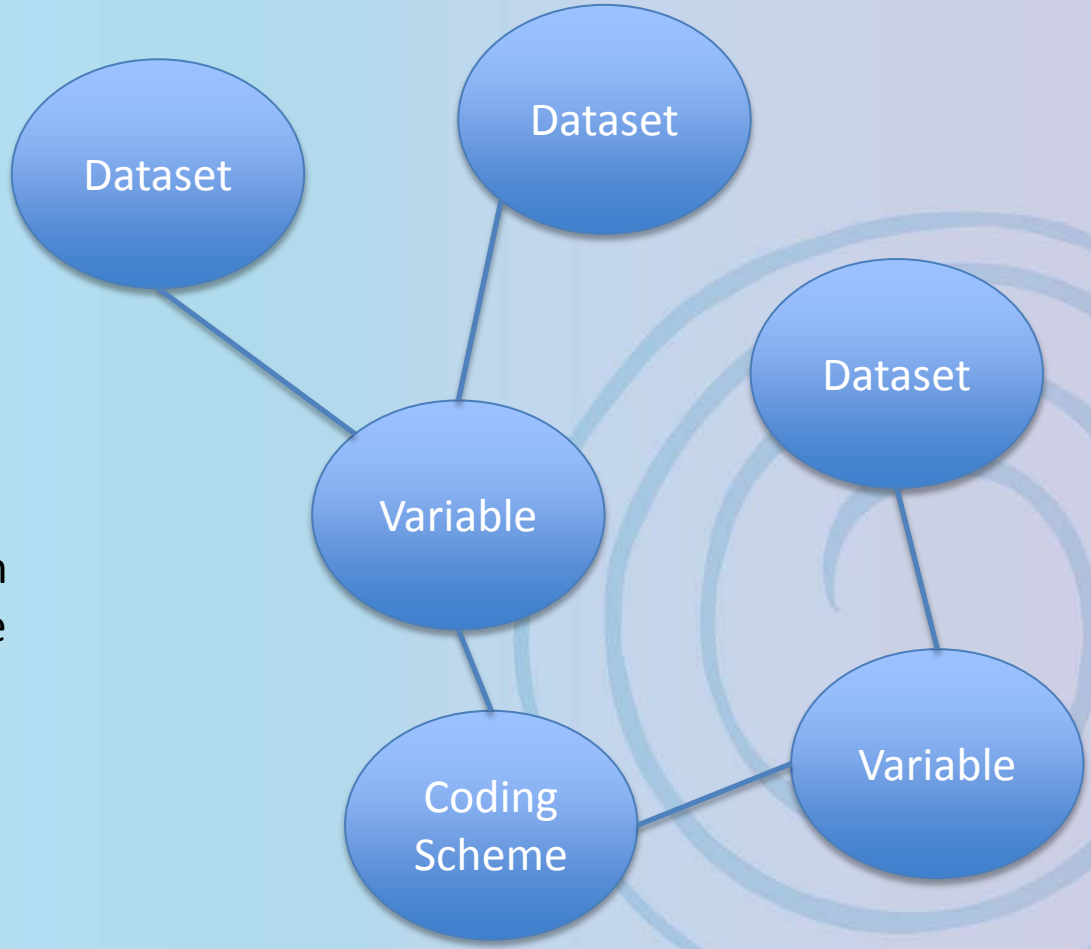


CISER

Controlled Provenance Vocabulary; RDF Triples

Graph structure lends itself well to managing and querying provenance.

Developing a controlled vocabulary to describe these relationships, then storing the data as RDF triples is one possible approach.



Future NCRN Work

- Deploy to a Census Research Data Center
  - Automate where possible
- Provide process by which metadata can be enhanced
  - Existing tools
  - Home grown programs
- Provenance solution
  - Embed known relationships within metadata
  - Algorithmically determine similarity (dataset level; variable level)