

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

Outline

- Overview of Cornell NSF-Census (NCRN) Node
 - Comprehensive Extensive Data Documentation and Access Repository (CED²AR)
 - Extract, Transform and Load (ETL)
 - Application Programming Interface (API)
 - User Interface (UI)
 - Coming Features of CED²AR
 - The Problem of Provenance
 - Future NCRN Work
- 

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!”



NCRN DDI Solution at the Variable Level: <dataAccs>

```
<studyDscr>
  <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>
</studyDscr>
```

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]
  <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>
```

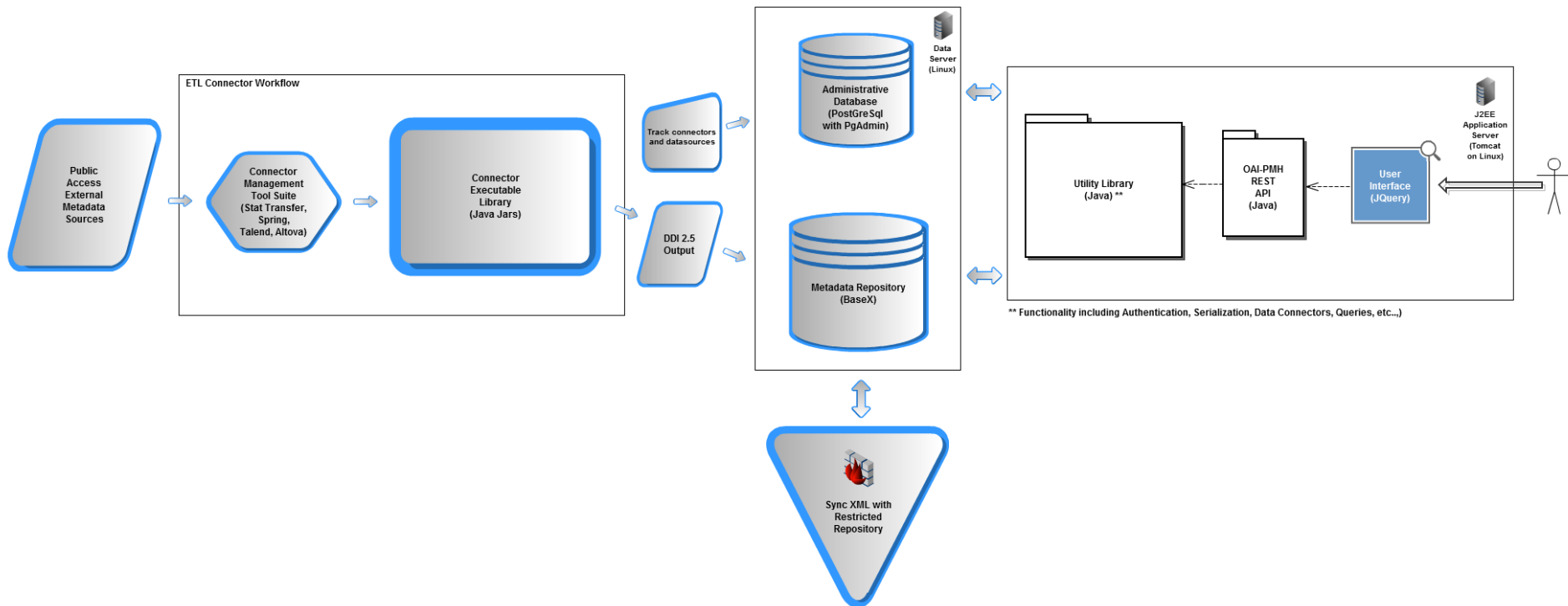
No DDI Solution at the Level of a Value Label

```
<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.

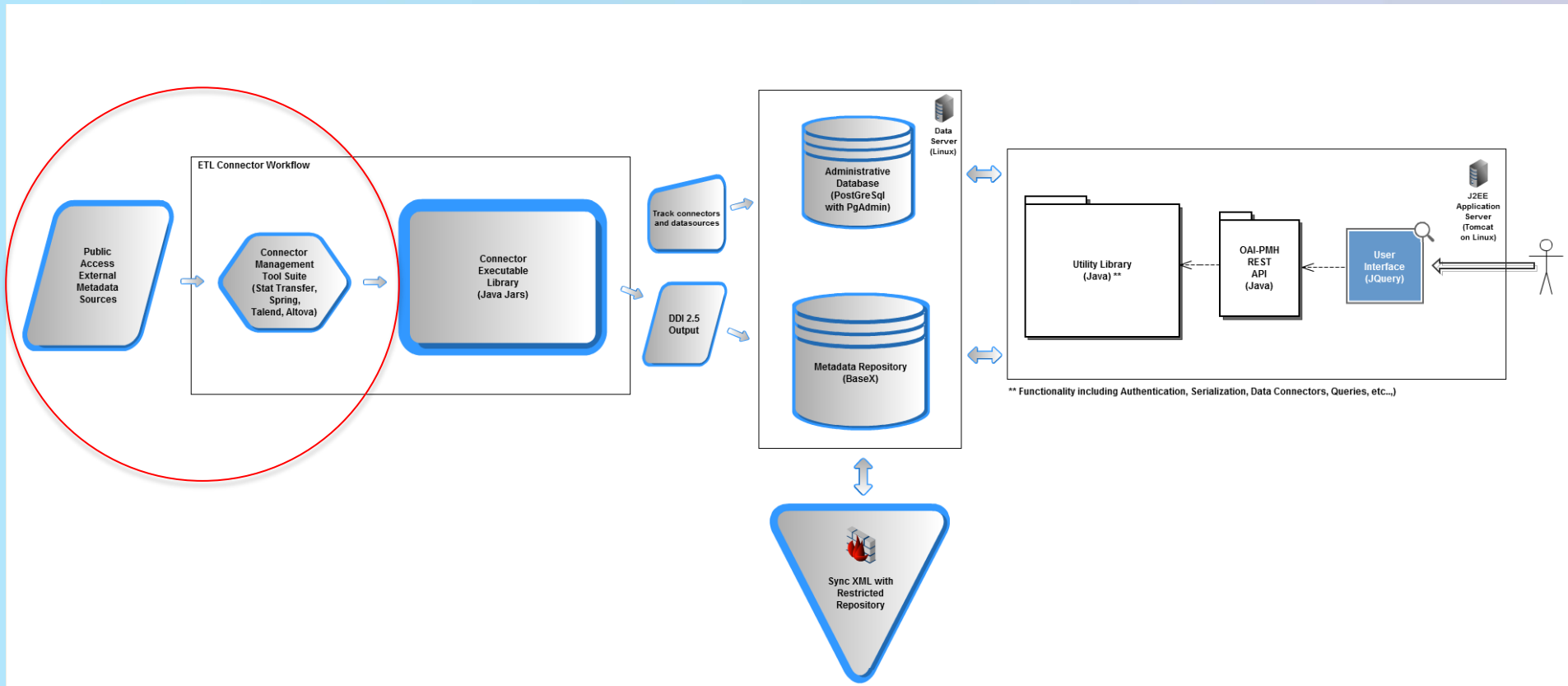
CISER

Schematic of NCRN System



CISER

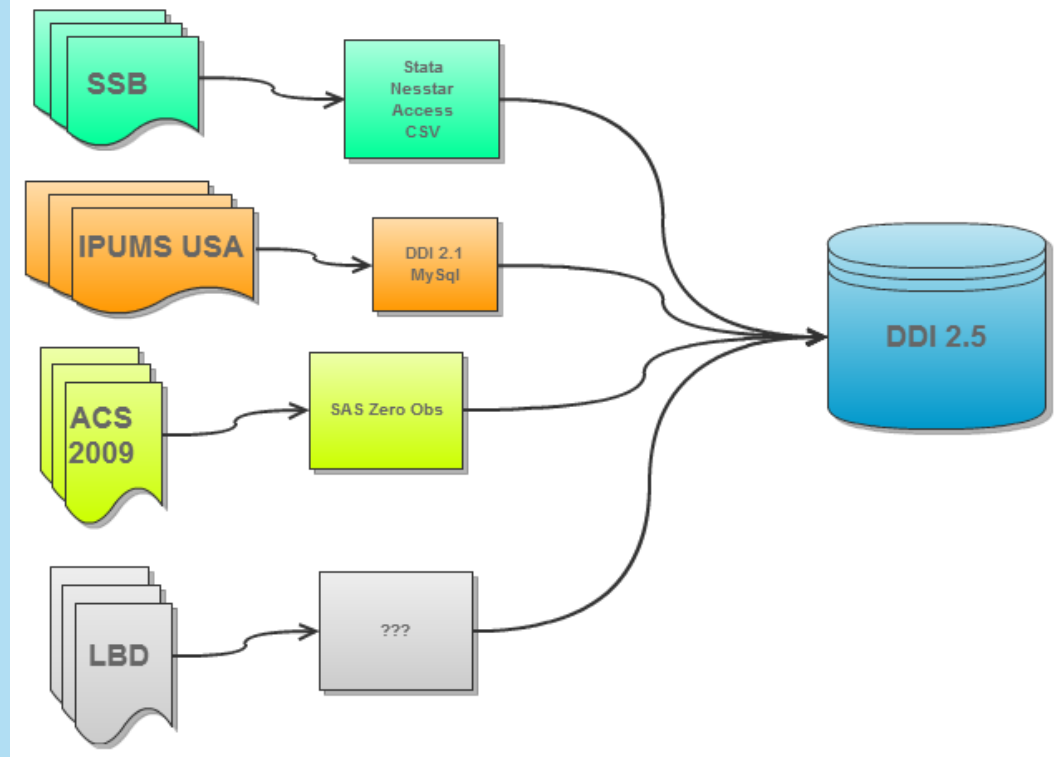
Schematic of NCRN System



The challenge of ingesting disparate data

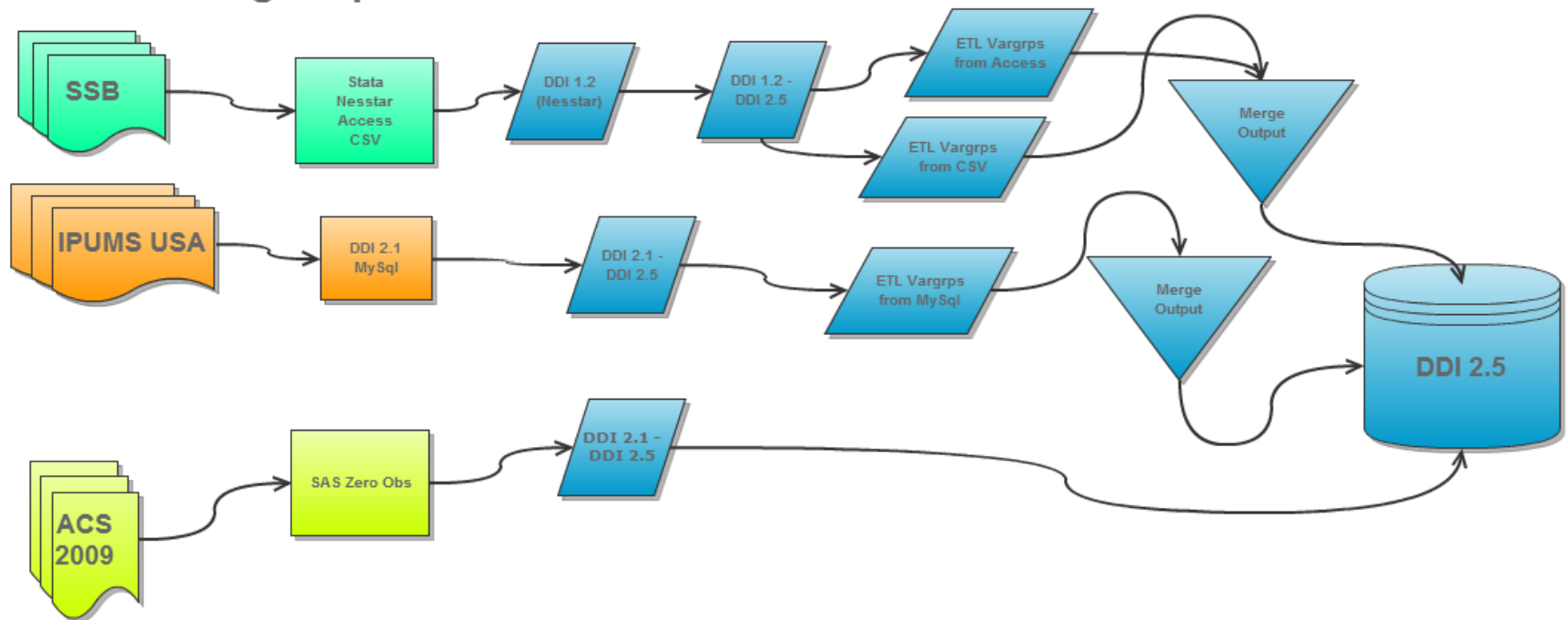
- Format disparity
- Schema disparity
- Sparseness of metadata

Standardizing Disparate Metadata Sources



ETL Modular Approach – Building to Reuse

Standardizing Disparate Metadata Sources

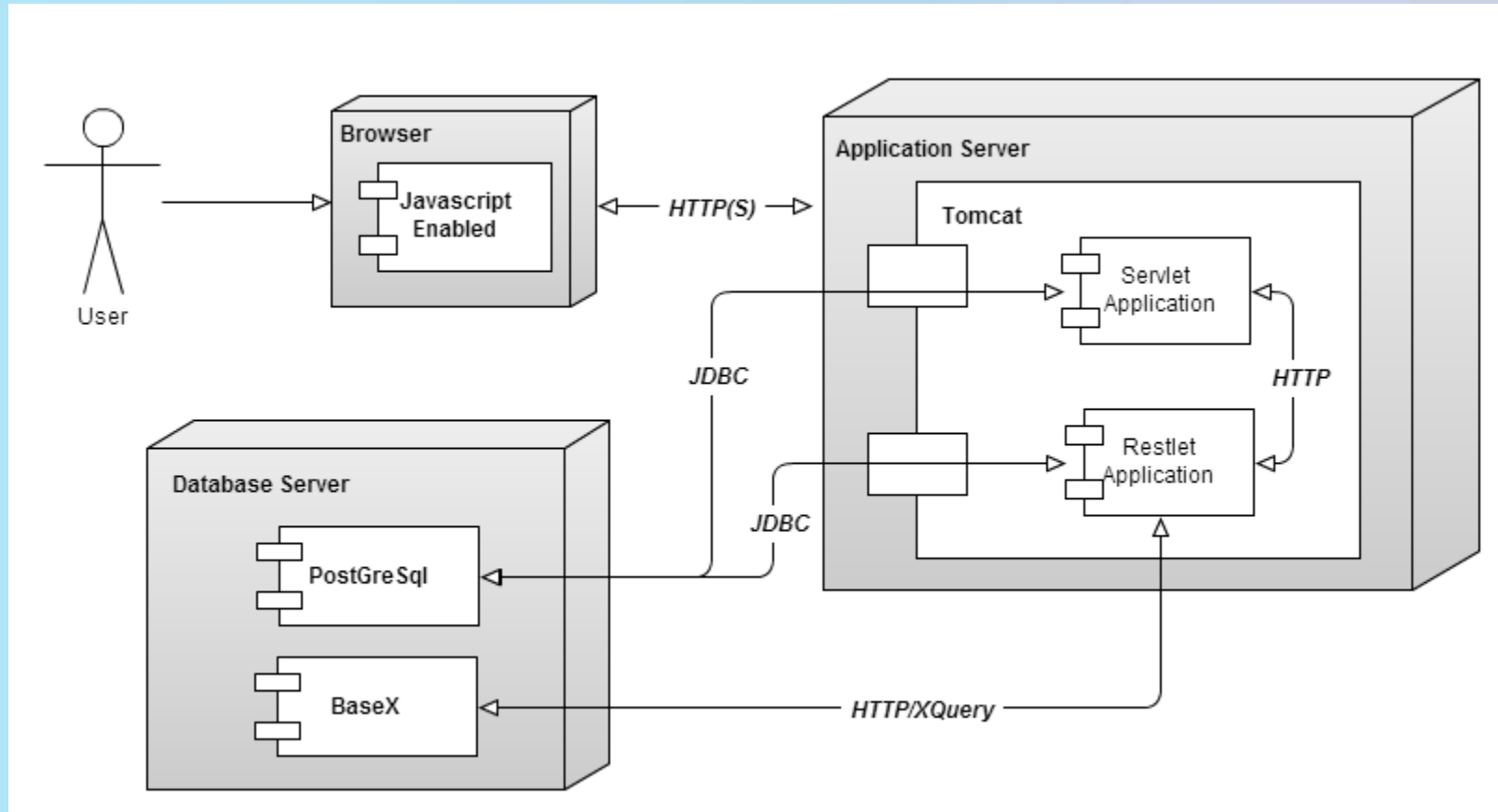


*This diagram omits the existing tools used to maneuver data into digestable formats



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



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*
```



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.



User Interface – Simple search

About CED²AR | Login or Register

CED²AR The Comprehensive Extensible Data Documentation and Access Repository

Simple Search **Advanced Search** **Browse Metadata**

Enter keywords below to do a broad search of ALL FIELDS within the available codebook metadata.
(Hint: For a more refined search, use the [Advanced Search](#) form.)

Welcome to the Comprehensive Extensible Data Documentation and Access Repository (CED²AR)

CED²AR is a [National Science Foundation \(NSF\) funded](#) project developed by the [NSF Census Research Network - Cornell Node \(NCRN\)](#).

It is designed to improve the documentation and discoverability of both public and restricted data from the federal statistical system.

To search across all datasets in the repository, enter a term in the search box above and click the 'Search' button.

User Interface – Advanced Search

About CED²AR | [Login](#) or [Register](#)

CED²AR

The Comprehensive Extensible Data Documentation and Access Repository

Simple Search **Advanced Search** **Browse Metadata**

Use the form below to construct a more complex search of the available codebook metadata:

- The dropdown lists on the left contain all searchable fields.
- Type your keyword into the middle field.
- The dropdown lists on the right can be used to construct a [boolean search](#).

Search Field	Search Term	Boolean Term
Variable Name <input type="text"/>	<input type="text"/>	And <input type="text"/>
Variable Name <input type="text"/>	<input type="text"/>	And <input type="text"/>
Variable Name <input type="text"/>	<input type="text"/>	

[Reset Form](#)



CISER

User Interface – Dataset-specific view

The screenshot displays the CED²AR website interface. At the top left, the logo 'CED²AR' is followed by the tagline 'The Comprehensive Extensible Data Documentation and Access Repository'. In the top right corner, there are links for 'About CED²AR' and 'Login or Register'. A dark sidebar on the left contains a message: 'You are currently viewing metadata for [SSB](#). To view other datasets' metadata, [remove this filter](#).' The main content area is partially obscured by a white modal window titled 'Document Description' with a close button (x) in the top right corner. The modal contains the following sections: 'Citation', 'Title Statement' (with 'Title: SSB' and 'Production Statement'), and 'Guide'. The 'Guide' section contains a large block of placeholder text (Lorem ipsum). At the bottom of the modal, it states '95 variables found.' The background of the main page shows sections for 'Data', 'Debook metadata', and 'Access'.



User Interface – Dataset-specific view

The screenshot displays the CED²AR website interface. At the top, a red header contains the logo 'CED²AR' and the tagline 'The Comprehensive Extensible Data Documentation and Access Repository'. Navigation links for 'About CED²AR' and 'Login or Register' are visible in the top right. Below the header, three tabs are present: 'Simple Search' (selected), 'Advanced Search', and 'Browse Metadata'. A search box with a 'Search' button is provided. A central grey box contains a welcome message and instructions on how to use the search function. A sidebar on the left provides information about the current dataset's metadata and offers a link to view other datasets' metadata.

CED²AR The Comprehensive Extensible Data Documentation and Access Repository

About CED²AR | Login or Register

You are currently viewing metadata for [SSB](#).

To view other datasets' metadata, [remove this filter](#).

Simple Search | **Advanced Search** | **Browse Metadata**

Enter keywords below to do a broad search of ALL FIELDS within the available codebook metadata.
(Hint: For a more refined search, use the [Advanced Search](#) form.)

Search

Welcome to the Comprehensive Extensible Data Documentation and Access Repository (CED²AR)

CED²AR is a [National Science Foundation \(NSF\) funded](#) project developed by the [NSF Census Research Network - Cornell Node \(NCRN\)](#).

It is designed to improve the documentation and discoverability of both public and restricted data from the federal statistical system.

To search across all datasets in the repository, enter a term in the search box above and click the 'Search' button.

User Interface – Search results

About CED²AR | [Login](#) or [Register](#)

CED²AR

The Comprehensive Extensible Data Documentation and Access Repository

You are currently viewing metadata for [SSE](#).

To view other datasets' metadata, [remove this filter](#).

Simple Search | **Advanced Search** | **Browse Metadata**

You searched for "a", 86 results returned. [<<Search again](#)

Show entries Search:

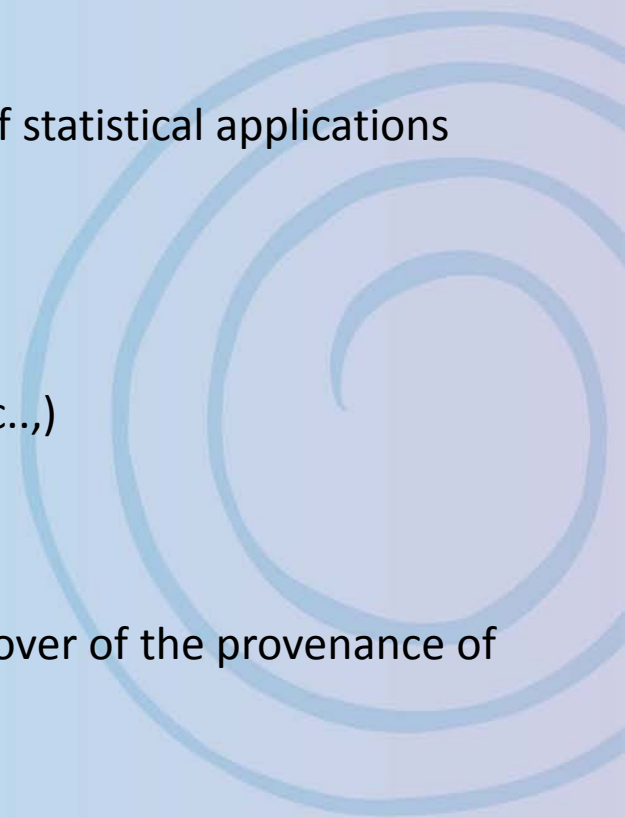
Variable	Label
birthdate	Date of Birth
cur_endmar	
cur_endmar_flag	
cur_endmar_reas	
cur_startmar	
current_enroll_coll	Flag currently enrolled in college
current_enroll_hs	Flag currently enrolled in high school
db_pension	Defined Benefit Pension Plan
dc_pension	Defined Contribution Pension Plan
deathdate	Date of Death

Showing 1 to 10 of 86 entries [Previous](#) [Next](#)

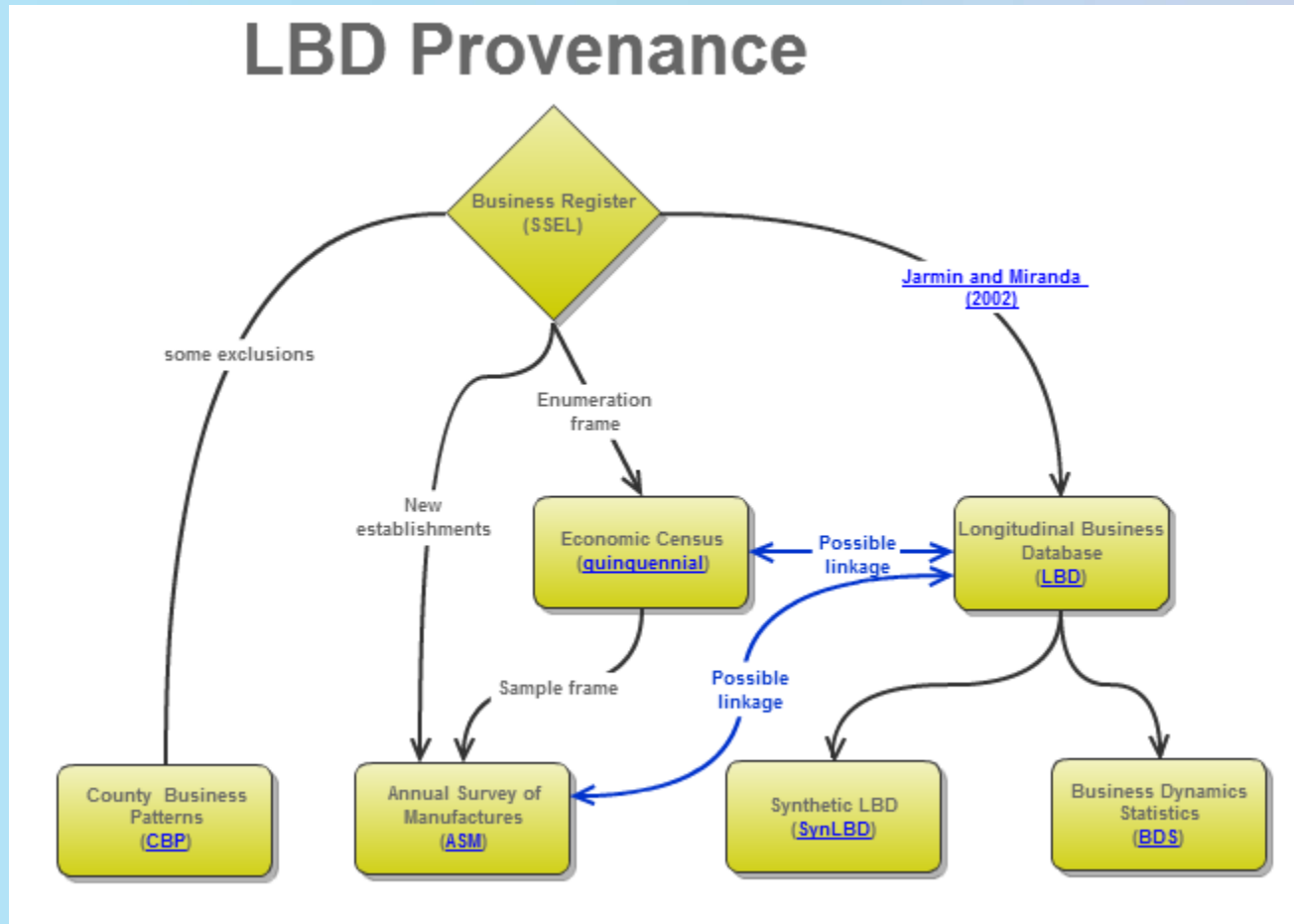


Future Features

- 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



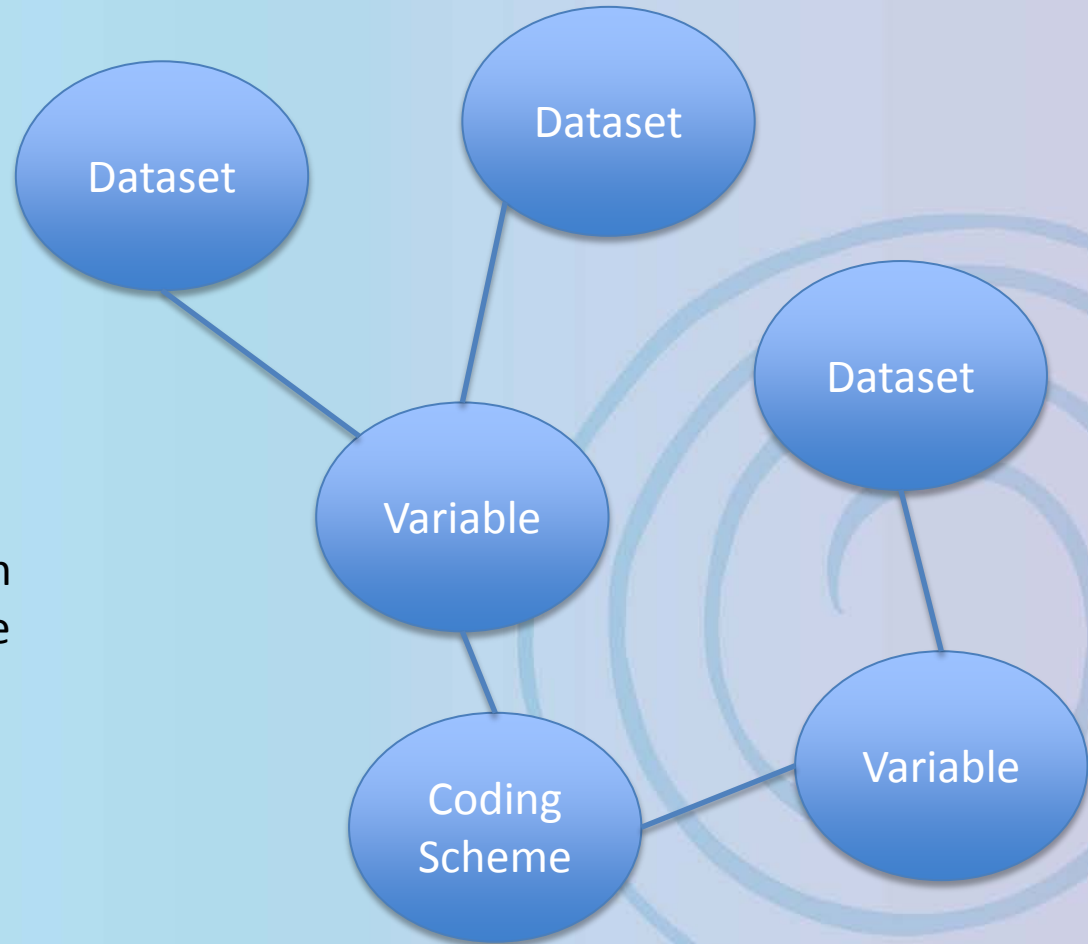


CISER

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)

