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
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>

```xml
<stdyDescr>
  <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>
</stdyDescr>
```
Variable Level Solution (continued)

```xml
<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="valid">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]
    </catgry>
  </valrng>
  <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="valid">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>
  </valrng>
  <varFormat schema="other" type="numeric"/>
</var>
```
No DDI Solution at the Level of a Value Label

Small tweak to the DDI Codebook Schema would fix this.
Schematic of NCRN System
The challenge of ingesting disparate data

- Format disparity
- Schema disparity
- Sparceness of metadata
ETL Modular Approach – Building to Reuse

Standardizing Disparate Metadata Sources

SSB
- Stata Nestar Access CSV
- DDI 1.2 (Nestar)
  - DDI 1.2 - DDI 2.5

IPUMS USA
- DDI 2.1 - MySQL
  - DDI 2.1 - DDI 2.5

ACS 2009
- SA3 Zero Obs
  - DDI 2.1 - DDI 2.5

DDI 2.5

ETL Vargrps from Access
ETL Vargrps from CSV

- Merge Output
  - Merge Output

*This diagram omits the existing tools used to maneuver data into digestable formats.*
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$ced2ar/codeBook/dataDscr/var%20where%20starts-with(lower-case($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

CED²AR The Comprehensive Extensible Data Documentation and Access Repository

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

CED²AR The Comprehensive Extensible Data Documentation and Access Repository

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.

<table>
<thead>
<tr>
<th>Search Field</th>
<th>Search Term</th>
<th>Boolean Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable Name</td>
<td></td>
<td>And</td>
</tr>
<tr>
<td>Variable Name</td>
<td></td>
<td>And</td>
</tr>
<tr>
<td>Variable Name</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reset Form  Search
User Interface – Dataset-specific view
User Interface – Dataset-specific view

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.)

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

CED$^2$AR - The Comprehensive Extensible Data Documentation and Access Repository

You searched for "a", 86 results returned.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>birthdate</td>
<td>Date of Birth</td>
</tr>
<tr>
<td>cur_endmar</td>
<td></td>
</tr>
<tr>
<td>cur_endmar_flag</td>
<td></td>
</tr>
<tr>
<td>cur_endmar_reas</td>
<td></td>
</tr>
<tr>
<td>cur_startmar</td>
<td></td>
</tr>
<tr>
<td>current_enroll_col</td>
<td>Flag currently enrolled in college</td>
</tr>
<tr>
<td>current_enroll_hs</td>
<td>Flag currently enrolled in high school</td>
</tr>
<tr>
<td>db_pension</td>
<td>Defined Benefit Pension Plan</td>
</tr>
<tr>
<td>dc_pension</td>
<td>Defined Contribution Pension Plan</td>
</tr>
<tr>
<td>deathdate</td>
<td>Date of Death</td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 86 entries
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
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)