Crowdsourcing Codebook Enhancements
A DDI-based Approach

FCSM, December 2nd 2015

Lars Vilhuber (Cornell University)
Benjamin Perry (Cornell University)
Venkata Kambhampaty (Cornell University)
Kyle Brumsted (McGill University)
William Block (Cornell University)
Issues

• Data curators (Agencies) lack a mechanism to obtain structured feedback for their metadata

• Metadata standards for the social science community are difficult to navigate, even with complex tools

• Metadata curation is a labor intensive process
Our Approach

• Provide easy-to-use tools and interfaces to structured metadata

• Rely on open standards, namely the Data Documentation Initiative (DDI) schema

• Build infrastructure that enables data curators to leverage community-driven input to official documentation
How?

CED\textsuperscript{2}AR

The Comprehensive Extensible Data Documentation and Access Repository
What is CED$^2$AR?

- Metadata curation software
- Designed for documenting existing datasets
- Funded by NSF grant #1131848
- Online at www2.ncrn.cornell.edu/ced2ar-web
What is CED$^2$AR?
Basic Information Flow

Datasets → Internal Metadata → Official Metadata

Public Facing

Crowdsourced Metadata

User switches

Staging Area
Basic Information Flow

Staging Area

Datasets → Internal Metadata

Public Facing

Official Metadata

Crowdsourced Metadata

User switches
Internal Processing

1. Creation of skeletal metadata
   - Assuming data is already curated
   - Converting data into standardized metadata
     • Tools included (for SAS, Stata, SPSS, CSV), not discussed here [appendix slides]

2. Hand editing and subsetting
   - Adding verbose descriptions
   - Applying disclosure limitation
Internal Processing

• Simple editing interface
  – Web-based, with limited rich text features
  – Math allowed (LaTeX)

• Feedback
  – Completeness of codebook?
  – Without technical jargon!
  – Can be tuned
Internal Processing: Hand Editing

Abstract

The SIPP Synthetic Beta (SSB) is a Census Bureau product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the Survey of Income and Program Participation (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Form W-2 records and SSA records of retirement and disability benefit receipt, and were produced by Census Bureau staff economists and statisticians in collaboration with researchers at Cornell University, the SSA and the IRS. The purpose of the SSB is to provide access to linked data that are usually not publicly available due to confidentiality concerns.

To overcome these concerns, Census has synthesized, or modeled, all the variables in a way that changes the record of each individual in a manner designed to preserve the underlying covariate relationships between the variables. The only variables that were not altered by the synthesis process and still contain their original values are gender and a link to the first reported marital partner in the survey. Seven SIPP panels (1990, 1991, 1992, 1993, 1996, 2001, 2004) form the basis for the SSB, with a large subset of variables available across all the panels selected for inclusion and harmonization across the years. Administrative data were added and some editing was done to correct for logical inconsistencies in the IRS/SSA earnings and benefits data.

This field supports ASCII math. See FAQ for details.
Internal Processing: Access Control

- Marking elements with different restrictions

```
Select what sub-elements to mark
- Select All
  - Mean
  - Median
  - Mode
  - Valid
  - Invalid
  - Min
  - Max
  - Standard Deviation
  - Other Summary Statistics

Select what access level to apply, then check which variables to apply to. Finally, click changes levels.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Label</th>
<th>Top Access Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>afdc_MN</td>
<td>Indicator for receipt of AFDC or TANF benefits</td>
<td>released</td>
</tr>
<tr>
<td>afdcamt_MN</td>
<td>Amount of AFDC received</td>
<td>released</td>
</tr>
<tr>
<td>birthdate</td>
<td>Date of Birth</td>
<td>released</td>
</tr>
<tr>
<td>current_enroll_coll</td>
<td>Currently Enrolled in College</td>
<td>released</td>
</tr>
<tr>
<td>current_enroll_hs</td>
<td>Currently Enrolled in HS (or less)</td>
<td>released</td>
</tr>
</tbody>
</table>
```
Internal Processing: Scoring

• Provide feedback to improve sparse documentation

Codebook Score

Variables

100.0% of variables have labels
85.1% of variables have significant full descriptions
Variables without significant full descriptions ...

43.0% of variables have values
Variables without values ...

0.0% of variables have summary statistics

Title Page

Missing related studies
Missing access conditions
Missing bibliographic citation
Missing related publications

Overall Score

80.3%
Workflow control

- **Ability to view additions/subtractions**
  - Between versions
  - Between crowd-sourced information and official information

- **Ability to control access**
  - Editing versus viewing
  - Authentication and reputation
Versioning

- Uses Git, a distributed version control system
- Every aspect of the system is configurable
  - Scheduled tasks check for changes
  - Once changes exceed threshold, they are pushed
  - Remote instances pull changes on demand
Combining Knowledge

Viewing changes made

<table>
<thead>
<tr>
<th>Codebook</th>
<th>Git Status</th>
<th>Last Local Update</th>
<th>Last Remote Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssb.v602.xml</td>
<td>UP_TO_DATE</td>
<td>September 10, 2015 at 9:25 AM {ssbv602,<a href="mailto:fs379@cornell.edu">fs379@cornell.edu</a>,cover}</td>
<td>September 10, 2015 at 9:25 AM {ssbv602,<a href="mailto:fs379@cornell.edu">fs379@cornell.edu</a>,cover}</td>
</tr>
<tr>
<td>ssb.v6.xml</td>
<td>UP_TO_DATE</td>
<td>November 12, 2015 at 4:16 PM {ssbv601,<a href="mailto:lorireeder@gmail.com">lorireeder@gmail.com</a>,var,layoff_M}</td>
<td>November 12, 2015 at 4:16 PM {ssbv601,<a href="mailto:lorireeder@gmail.com">lorireeder@gmail.com</a>,var,layoff_M}</td>
</tr>
</tbody>
</table>
All changes are logged externally via Git

<table>
<thead>
<tr>
<th>Author</th>
<th>Commit</th>
<th>Message</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>tomcat7</td>
<td>afa515</td>
<td>{ssbv601,<a href="mailto:lars@vilhuber.com">lars@vilhuber.com</a>,cover}</td>
<td>37 minutes ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>5c824de</td>
<td>{ssbv601,<a href="mailto:lars@vilhuber.com">lars@vilhuber.com</a>,cover}{ssbv601,<a href="mailto:lars@vilhuber.com">lars@vilhuber.com</a>,var,fl...</td>
<td>an hour ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>c03c58f</td>
<td>Committing codebooks retrieved directly from BaseX</td>
<td>4 days ago</td>
</tr>
<tr>
<td>venkata</td>
<td>a61abe3</td>
<td>{testbdv1,anonymous,edit}</td>
<td>4 days ago</td>
</tr>
<tr>
<td>venkata</td>
<td>5b1e51e</td>
<td>{testbdv1,anonymous,edit}</td>
<td>4 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>5edbf9</td>
<td>{acs2009,<a href="mailto:bap63@cornell.edu">bap63@cornell.edu</a>,edit}</td>
<td>5 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>d66d3b4</td>
<td>{ssbv601,<a href="mailto:lorireeder@gmail.com.var">lorireeder@gmail.com.var</a>,plus_sdil_benefit_totamt_k}{ss...</td>
<td>5 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>1f854c1</td>
<td>{siabv1,<a href="mailto:warren.brown48@gmail.com">warren.brown48@gmail.com</a>,cover}{siabv1,<a href="mailto:bap63@cornell.e">bap63@cornell.e</a>...</td>
<td>5 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>c277f31</td>
<td>{siabv1,<a href="mailto:warren.brown48@gmail.com">warren.brown48@gmail.com</a>,var,bild}{siabv1,warren.brown4...</td>
<td>2015-11-17</td>
</tr>
<tr>
<td>tomcat7</td>
<td>b34a118</td>
<td>{siabv1,<a href="mailto:warren.brown48@gmail.com">warren.brown48@gmail.com</a>,var,persnr}{siabv1,warren.brow...</td>
<td>2015-11-17</td>
</tr>
<tr>
<td>venkata</td>
<td>2cb6d7d</td>
<td>{ltbdv2,anonymous,cover}</td>
<td>2015-11-17</td>
</tr>
<tr>
<td>tomcat7</td>
<td>12650cf</td>
<td>{siabv1,<a href="mailto:warren.brown48@gmail.com">warren.brown48@gmail.com</a>,var,nn}{siabv1,warren.brown4...</td>
<td>2015-11-17</td>
</tr>
<tr>
<td>tomcat7</td>
<td>aaf94f1</td>
<td>{siabv1,<a href="mailto:bap63@cornell.edu">bap63@cornell.edu</a>,edit}{blss2011,<a href="mailto:bap63@cornell.edu">bap63@cornell.edu</a>,edit}{...</td>
<td>2015-11-17</td>
</tr>
<tr>
<td>tomcat7</td>
<td>0e52a6e</td>
<td>Committing codebooks retrieved directly from BaseX</td>
<td>2015-11-17</td>
</tr>
</tbody>
</table>
Basic Information Flow

Staging Area

Datasets → Internal Metadata

Public Facing

Official Metadata

User switches

Crowdsourced Metadata
Official view

CED²AR
Official Server - The Comprehensive Extensible Data Documentation and Access Repository

You are viewing the official CED²AR View Variables (123 variables)
Last update to metadata: 2015-11-24 10:05:15 (upload date)
Document Date: November 12, 2015

Codebook prepared by: Cornell NSF-Census Research Network

Data Distributed by:
Labor Dynamics Institute
http://www2.vrdc.cornell.edu/news/data/sipp-synthetic-beta-file/
Crowdsourced view

CED²AR
Community Development Server (Beta) - The Comprehensive Extensible Data Documentation and Access Repository

You are viewing crowdsourced metadata. View the official version.

SIPP Synthetic Beta
v6.02

View Variables (123 variables)
Last update to metadata: 2015-11-24 09:59:07 (auto-generated)
Document Date: November 12, 2015

Codebook prepared by: Cornell NSF-Census Research Network

Data Distributed by:

Labor Dynamics Institute
http://www2.vrdc.cornell.edu/news/data/sipp-synthetic-beta-file/

Basic Information Flow

Staging Area

Datasets → Internal Metadata

Public Facing

Official Metadata → Crowdsourced Metadata

User switches
Authentication and Attribution

• When opening up contributions to a wide audience, how to triage between “rants” and meaningful contributions?
• Use of ORCID (academic network) for authentication
• Public attribution with link to (verified) academic ID is key for positive feedback (your effort is recognized) and prevention of negative contribution (your rant is traceable to you!)
Authentication

• Supports OpenID and OAuth2
  – Currently using Google and ORCID with OAuth2
  – Developing connectors to work with additional providers
• CED²AR handles identity management
Basic Information Flow

Staging Area

Datasets → Internal Metadata

Public Facing

Official Metadata

Crowdsourced Metadata

User switches
Combining Knowledge: Merging

• Curators are given an interface to merge crowdsourced documentation with official
Combining Knowledge: Merging

current_enroll_coll

### Crowdsourced Documentation

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>current_enroll_coll</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
<td>Currently Enrolled in College</td>
</tr>
<tr>
<td>Codebook</td>
<td>SIPP Synthetic Beta v6</td>
</tr>
<tr>
<td>Concept</td>
<td></td>
</tr>
<tr>
<td>Concept Vocabulary</td>
<td></td>
</tr>
<tr>
<td>Concept Vocabulary URI</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>numeric</td>
</tr>
<tr>
<td>Files</td>
<td></td>
</tr>
</tbody>
</table>

### Official Documentation

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>current_enroll_coll</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label</td>
<td>Currently Enrolled</td>
</tr>
<tr>
<td>Codebook</td>
<td>SIPP Synthetic Beta v6</td>
</tr>
<tr>
<td>Concept</td>
<td></td>
</tr>
<tr>
<td>Concept Vocabulary</td>
<td></td>
</tr>
<tr>
<td>Concept Vocabulary URI</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>numeric</td>
</tr>
<tr>
<td>Files</td>
<td></td>
</tr>
</tbody>
</table>
Combining Knowledge: Merging

Crowdsourced Documentation

Last update to metadata: 2015-08-18 08:43:01 (upload date)
Document Date: June 15th, 2014

Citation

Please cite this codebook as:

Please cite this dataset as:
U.S. Census Bureau. SIPP Synthetic Beta: Version 5.1 [Computer file]. Washington DC; Cornell University, Synthetic Data Server [distributor], Ithaca, NY, 2013

Abstract

The SIPP Synthetic Beta (SSB) is a Census Bureau product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the Survey of Income and Program Participation (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Earned Income and SSA

Official Documentation

Last update to metadata: 2015-10-23 11:12:44 (auto-generated)
Document Date: June 15, 2014

Citation

Please cite this codebook as:

Please cite this dataset as:
U.S. Census Bureau. SIPP Synthetic Beta: Version 5.1 [Computer file]. Washington DC; Cornell University, Synthetic Data Server [distributor], Ithaca, NY, 2013

Abstract

The SIPP Synthetic Beta (SSB) is a Census Bureau product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the Survey of Income and Program Participation (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Earned Income and SSA
Combining Knowledge: Citations

• Contributors can be tracked for each of their changes

**Modified Variables**

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Date Changed</th>
<th>Commit Message</th>
<th>User</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015 at 11:15 AM</td>
<td>View commit</td>
<td><a href="mailto:lorireeder@gmail.com">lorireeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td>pos_phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015 at 11:15 AM</td>
<td>View commit</td>
<td><a href="mailto:lorireeder@gmail.com">lorireeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td>phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015 at 11:15 AM</td>
<td>View commit</td>
<td><a href="mailto:lorireeder@gmail.com">lorireeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td>pos_phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015 at 11:15 AM</td>
<td>View commit</td>
<td><a href="mailto:lorireeder@gmail.com">lorireeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td>phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015 at 11:15 AM</td>
<td>View commit</td>
<td><a href="mailto:lorireeder@gmail.com">lorireeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td>pos_phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015 at 11:15 AM</td>
<td>View commit</td>
<td><a href="mailto:lorireeder@gmail.com">lorireeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td>phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015 at 11:15 AM</td>
<td>View commit</td>
<td><a href="mailto:lorireeder@gmail.com">lorireeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td>pos_phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015 at 11:15 AM</td>
<td>View commit</td>
<td><a href="mailto:lorireeder@gmail.com">lorireeder@gmail.com</a></td>
<td>Remote</td>
</tr>
</tbody>
</table>
Combining Knowledge: Citations

Lars Vilhuber

ORCID ID
id: orcid.org/0000-0001-5733-8932

Education (3)
Employment (1)
Funding (7)
Works (29)

CED²AR: The Comprehensive Extensible Data Documentation and Access Repository
IEEE/ACM Joint Conference on Digital Libraries
2014-09 | conference-paper
DOI: 10.1109/jcdl.2014.6970178
Source: CrossRef Metadata Search
You are viewing the official metadata. View crowdsourced contributions.

CED$^2$AR / CNSS 2012

CNSS 2012

View Variables (123 variables)
Last update to metadata: 2015-11-23 11:38:10 (upload date)
Document Date: 2015-01-27 11:59:45

Codebook prepared by: Cornell Institute for Social and Economic Research
Data prepared by: Cornell Survey Research Institute

Data Distributed by:
Cornell Institute for Social and Economic Research
http://cisar.cornell.edu

Citation
Please cite the codebook as:

Please cite this dataset as:

Try for yourself: http://demo.ncrn.cornell.edu
Future Directions

• Where are we taking the project?
• RePEC integrations?
  – Additional authentication tied to existing (document) metadata
  – “Claiming” user-created datasets
  – Entry-point to discover datasets (data provenance) and therefore need for documentation
Thank you!
Questions?
ced2ar-devs-I@cornell.edu
Extra Slides
What technologies are used to build CED²AR?

- Server: Apache/Tomcat
- Databases: BaseX and Neo4j
- Primary Languages/Frameworks:
  - Java
  - Spring MVC
  - Bootstrap, JQuery, LESS
Generating Metadata

- Convert data to metadata
- Online or offline conversion
- Can start with Stata, SPSS, SAS, R, ASCII or a relational DB
- We generate DDI 2.5 (codebook)