Crowdsourcing Metadata – Challenges and Outlook

Washington, 10 May 2016

Lars Vilhuber, William Block (Cornell University)
Crowdsourcing Metadata – Challenges and Outlook

Lars Vilhuber, William Block (Cornell University)

Washington, 10 May 2016
Acknowledgements

Based on work with

• Benjamin Perry (formerly Cornell University)
• Venkata Kambhampaty (formerly Cornell University)
• Kyle Brumsted (McGill University)
• Jeremy Williams (Cornell University)
• Carl Lagoze (University of Michigan)
• John Abowd (Cornell University)

and materials presented in INFO 7470, all of that with funding by NSF Grant #1131848
I’m going to argue that...

- **Replicability** is a problem...
  - and (A) easier *deposit* methods could alleviate it
  - but progress is slow
I’m going to argue that...

- Having replicable archives *shifts* the problem...
  - in time: (B) older articles cannot be linked to data
  - in scope: (C) curators need *expert help* in documenting the data
A test

LDI “reproducibility” project:
Kingi, Stanchi, Vilhuber (2016, unpublished)
“The Reproducibility of Economics Research”
Kingi, Stanchi, Vilhuber (2016)

- Simpler test:

Do the provided data and programs yield the published results?
Figure 1: A Breakdown of the Articles
Total Articles (109)

- Confidential Data (44)
- Non Confidential Data (65)

- Unsuccessful (28)
- Successful (37)
The old source of knowledge – and data!
Options are available

• Social and behavioral sciences

Share your social and behavioral science research data

Get started now »

Maximize Access
Be recognized and cited

Store Safely
Store your data with confidence

Protect Confidentiality
Ensure confidentiality and privacy
Options are available

• “Research data”

Open source research data repository software

Researchers
Enjoy full control over your data. Receive web visibility, academic credit, and increased citation counts. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

Journals
Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an unbreakable link between articles in your journal and associated data. Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?

Developers
Participate in a vibrant and growing community that is helping to drive the norms for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. Integrate research analysis, visualization and exploration tools, or other research and data archival systems with Dataverse. Want to contribute?

Institutions
Establish a research data management solution for your community. Federation with a growing list of Dataverse repositories worldwide for increased discoverability of your community’s data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?
Researchers don’t use them

Training? Incentives? Ease of use?
A Good Example
Gentzkow, Shapiro, Sinkinson (2014)

  American Economic Review, 104(10): 3073-3114. DOI: 10.1257/aer.104.10.3073
- Data at http://doi.org/10.3886/E1361V3

Principal Investigator(s): Gentzkow, Matthew (University of Chicago, Booth School of Business); Shapiro, Jesse (University of Chicago, Booth School of Business); Sinkinson, Michael

<table>
<thead>
<tr>
<th>Title</th>
<th>Date Entered</th>
<th>File Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>codebook</td>
<td>2014-04-03</td>
<td>.txt</td>
</tr>
<tr>
<td>Gentzkow, Matthew; Shapiro, Jesse; Sinkinson, Michael</td>
<td>9:08 PM</td>
<td></td>
</tr>
<tr>
<td>data</td>
<td>2014-04-03</td>
<td>.txt</td>
</tr>
<tr>
<td>Gentzkow, Matthew; Shapiro, Jesse; Sinkinson, Michael</td>
<td>9:08 PM</td>
<td></td>
</tr>
<tr>
<td>Orig</td>
<td>2014-04-07</td>
<td>.txt</td>
</tr>
<tr>
<td>Gentzkow, Matthew; Shapiro, Jesse; Sinkinson, Michael</td>
<td>1:28 PM</td>
<td></td>
</tr>
</tbody>
</table>


Persistent URL: http://doi.org/10.3886/E1361V3

Project Description

Summary
The focus of this data collection was the historical circulation and subscription prices of US daily newspapers in 1924. These data are obtained from audit reports obtained from the Audit Bureau of Circulations, an independent organization created to verify circulation. They include circulation by town and delivery channel for each newspaper. The sample is all audited daily newspapers by the Audit Bureau of Circulations.

All pdfs and extracted data. Copyright belongs to the Audit Bureau of Circulations. We have obtained written permission from the Audit Bureau of Circulations to post the PDFs and data files.
What’s good about this?

• Permanent URL
• Availability of
  – Original data
  – Transformed data
  – Open availability
• Easy online inspection
Not Perfect

- Archive at openICPSR not actually tied to article and vice-versa

- Conversely, “online appendix” just a “blob”
Journals are starting to use them

The American Journal of Political Science is committed to significant advances in knowledge and understanding of citizenship, governance, and politics, and to the public value of political science research.

1 to 10 of 228 Results

Replication Data for: Navigating the Range of Statistical Tools for Inferential Network Analysis
Apr 21, 2016
Cramer, Skyler; Leifeld, Philip; McGing, Scott; Rolfe, Meredith. 2016, "Replication Data for Navigating the Range of Statistical Tools for Inferential Network Analysis", http://dx.doi.org/10.7910/DVN/3PFR67, Harvard Dataverse, V1

The last decade has seen substantial advances in statistical techniques for the analysis of network data, and a major increase in the frequency with which these tools are used. These techniques are designed to accomplish the same broad goal, statistically valid inference in the presence of dependent data.

Replication Data for: Constitutional Qualms or Politics as Usual? The Factors Shaping Public Support for Unilateral Action
Apr 13, 2016
Kinder, Douglas; Christensen, Dino. 2016, "Replication Data for: Constitutional Qualms or Politics as Usual? The Factors Shaping Public Support for Unilateral Action", http://dx.doi.org/10.7910/DVN/DDSH05, Harvard Dataverse, V1

The formal institutional constraints that Congress and the courts impose on presidential unilateral action are real. As a result, recent scholarship suggests that public opinion may be the strongest check against executive overreach. However, little is known about how the public...

Replication Data for: Can Political Inequalities be Educated Away? Evidence from a Large-scale Reform
Apr 13, 2016
Lindgren, Karl-Oskar; Oskarsson, Sven; Dawes, Christopher. 2016, "Replication Data for: Can Political Inequalities be Educated Away? Evidence from a Large-scale Reform", http://dx.doi.org/10.7910/DVN/3TO01Q, Harvard Dataverse, V1

Over the years, many suggestions have been made on how to reduce the importance of family background in political recruitment. In this...
But the biggest problem...

Figure 1: A Breakdown of the Articles
Total Articles (109)

Confidential Data (44)  Non Confidential Data (65)

Unsuccessful (28)  Successful (37)
• Articles using confidential data are (weakly) more cited than others
But: for confidential data…

- Data is not available
- Metadata is not available
- Programs? So-so…
Should We Just Trust These Guys?
Some are quite commendable
**Even detailed information**

| **Full Title** | General Social Survey, 2011: Cycle 25, Family |
| **Subtitle**   | Cycle 25, Family |
| **Alternative Title** | GSS 2011: Family |
| **Parallel Title** | Encuesta social generale, 2011: Cycle 25, Famille |
| **Identification Number** | ca-statcan-58195 |

| **Authoring Entity** | |
| **Name** | **Affiliation** |
| Statistics Canada | StatCan |

| **Producer** | |
| **Name** | **Affiliation** | **Abbreviation** | **Role** |
| Statistics Canada | | StatCan |

**Copyright**
Copyright © Statistics Canada, 2012

**Date of Distribution**
2012-07-18

**Series Information**
General Social Survey - Family (GSS)  [4501]

**Version**
15769.6
How many users actually use that?
Data documentation is dry

• How reliable is that question?

Dataset: General Social Survey, 2011: Cycle 25, Family

Cycle 25, Family

Variable PA_Q240: Year parents separated

<table>
<thead>
<tr>
<th>Values</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>9997</td>
<td>Not asked</td>
</tr>
<tr>
<td>9998</td>
<td>Not stated</td>
</tr>
<tr>
<td>9999</td>
<td>Don't know</td>
</tr>
</tbody>
</table>

LITERAL QUESTION
In what year did your parents separate?

SUMMARY STATISTICS
This variable is numeric

UNIVERSE
Respondents who answered: PA_Q230 = 1.

NOTES
This variable is suppressed on the public use microdata file.
Don’t (just) liberate the data!

Liberate the data users!
Our contribution

Leverage researcher knowledge
Our Approach

- Rely on open standards, namely the Data Documentation Initiative (DDI) schema
- Provide easy-to-use tools and interfaces to structured metadata
- Build infrastructure that enables data curators to leverage community-driven input to official documentation
How?

CED$^2$AR

The Comprehensive Extensible Data Documentation and Access Repository
What is CED²AR?

- 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?

CED\(^2\)AR

Official Server - The Comprehensive Extensible Data Documentation and Access Repository

Search Variables  Browse Variables  Browse by Codebook  Documentation  About

Filter
Codebooks

Search

Searching all codebooks. No filters active.

Advanced Search

Show 10 \(\downarrow\) variables

Compare Variables

No variables selected

© 2012-2015, Cornell Institute for Social and Economic Research

Report a Bug  Email us  Copyright Information  NCRN GitHub
Basic Information Flow

Staging Area

Datasets → Internal Metadata

Public Facing

Official Metadata → Crowdsourced Metadata

User switches

Crowdsourced Metadata → Official Metadata
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

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

3. User accessible
   - These tools can be manipulated by normal users
   - They could be incorporated into existing workflows
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
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.
• 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 ... more*
- 43.0% of variables have values
  *Variables without values ... more*
- 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%
Fine-grained access controls

Important when working with confidential (meta)data
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>
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
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>0FeAsis51e</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>c83c58f9</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>5p1e51e</td>
<td>{testbdv1,anonymous,edit}</td>
<td>4 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>5edbf9f9</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>d66d3d04</td>
<td>{ssbv601,<a href="mailto:lorreeder@gmail.com">lorreeder@gmail.com</a>,var,phus_ssd_local_benefit_totamt_k}</td>
<td>5 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>1f845c1</td>
<td>{siabv1,<a href="mailto:warren.brown48@gmail.com">warren.brown48@gmail.com</a>,cover}{siabv1,<a href="mailto:bap63@cornell.edu">bap63@cornell.edu</a>,edit}</td>
<td>5 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>cb77f31</td>
<td>{siabv1,<a href="mailto:warren.brown48@gmail.com">warren.brown48@gmail.com</a>,var,bild}{siabv1,warren.brown4...}</td>
<td>cestesting</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.brown...}</td>
<td>cestesting</td>
</tr>
<tr>
<td>venkata</td>
<td>2cb6de7d</td>
<td>{lbdv2,anonymous,cover}</td>
<td>vrk4</td>
</tr>
<tr>
<td>tomcat7</td>
<td>12630cf</td>
<td>{siabv1,<a href="mailto:warren.brown48@gmail.com">warren.brown48@gmail.com</a>,var,bnn}{siabv1,warren.brown4...}</td>
<td>cestesting</td>
</tr>
<tr>
<td>tomcat7</td>
<td>af94f3</td>
<td>{siabv1,<a href="mailto:bap63@cornell.edu">bap63@cornell.edu</a>,edit}{bliss2011,<a href="mailto:bap63@cornell.edu">bap63@cornell.edu</a>,edit}</td>
<td>cestesting</td>
</tr>
<tr>
<td>tomcat7</td>
<td>0e52a6e</td>
<td>Committing codebooks retrieved directly from BaseX</td>
<td>cestesting</td>
</tr>
</tbody>
</table>
Basic Information Flow

**Staging Area**
- Datasets
- Internal Metadata

**Public Facing**
- Official Metadata
- Crowdsourced Metadata

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

You are viewing the official CED²AR crowdsourced contributions.

SIPP Synthetic Beta
v6.02

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 prepared by: United States Department of Commerce, Bureau of the Census.

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

United States Department of Commerce, Bureau of the Census.
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 prepared by: United States Department of Commerce, Bureau of the Census.

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

United States Department of Commerce, Bureau of the Census.
Authentication and Attribution

• When opening up contributions to a wide audience, how to triage between “rants” and meaningful contributions?

• Here: 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
• CED\textsuperscript{2}AR handles identity management
Editing made easy
You are viewing crowdsourced metadata. View the official version.

CED2AR / SIPP Synthetic Beta v6.02

SIPP Synthetic Beta v6.02

View Variables (123 variables)
View codebook score
Last update to metadata: 2016-01-26 14:35:26 (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/


<table>
<thead>
<tr>
<th>Variable Name</th>
<th>totearn_ser_YYYY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Access Level</td>
<td>released</td>
</tr>
<tr>
<td>Label</td>
<td>SER: Capped Earnings from all FICA-covered jobs</td>
</tr>
<tr>
<td>Access Level</td>
<td>released</td>
</tr>
<tr>
<td>Codebook</td>
<td>SIPP Synthetic Beta v6.02</td>
</tr>
<tr>
<td>Concept</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>numeric</td>
</tr>
<tr>
<td>Question Text</td>
<td></td>
</tr>
<tr>
<td>Full Description</td>
<td>Person-level annual earnings that were taxed by FICA; these variables include earnings only up to the FICA taxable maximum and cover the years 1951-2011. These earnings are the inputs for calculating the OASDI benefit a person and his or her spouse will receive upon retirement or disability.</td>
</tr>
<tr>
<td>Files</td>
<td></td>
</tr>
</tbody>
</table>
Note #1 - Access Level: released

Having a 0 value on toetarn_ser_yyyy could mean a couple of things: 1) this individual had no FICA-covered earnings in that year; 2) this individual had no labor income at all in this tax year; 3) this individual worked for an employer that failed to report earnings in this year (that is to say, this has nothing to do with whether a person filed taxes because the earnings are reported by the employer, not the employee). Prior to 1978, if a person has $0 earnings on the Summary Earnings Record (SER), there's really no way of knowing whether they had no earnings or whether they had non-FICA earnings because the SER only reports FICA-covered earnings reported by employers. For years 1978 and later, you can compare the SER to the Detailed Earnings Record (DER). The DER captures all earnings subject to income tax, so both FICA and non-FICA earnings are reported on the DER.

If you are looking at earnings in earlier years, particular the 1960s and earlier, there will be more people with $0 earnings because many jobs were not FICA-taxable then. Even today, there are some instances of legitimate non-FICA earnings that would not be reflected on the SER. One example of this is that graduate student stipends are not taxed for FICA or Medicare, so these earnings would not be reflected on the SER (https://www.irs.gov/Charities–Non-Profits/Student-Exception-to-FICA-Tax).
Having a 0 value on `totearn_ser_yyyy` could mean a couple of things: 1) this individual had no FICA-covered earnings in that year; 2) this individual had no labor income at all in this tax year; 3) this individual worked for an employer that failed to report earnings in this year (that is to say, this has nothing to do with whether a person filed taxes because the earnings are reported by the employer, not the employee). Prior to 1978, if a person has $0 earnings on the Summary Earnings Record (SER), there’s really no way of knowing whether they had no earnings or whether they had non-FICA earnings because the SER only reports FICA-covered earnings reported by employers. For years 1978 and later, you can compare the SER to the Detailed Earnings Record (DER). The DER captures all earnings subject to income tax, so both FICA and non-FICA earnings are reported on the DER.

If you are looking at earnings in earlier years, particularly the 1960s and earlier, there will be more people with $0 earnings because many jobs were not FICA-taxable then. Even today, there are some instances of legitimate non-FICA earnings that would not be reflected on the SER. One example of this is that graduate student stipends are not taxed for FICA or Medicare, so these earnings would not be reflected on the SER (https://www.irs.gov/Charities--Non-Profits/Student-Exception-to-FICA-Tax).
Basic Information Flow

**Staging Area**
- Datasets
- Internal Metadata

**Public Facing**
- Official Metadata
- Crowdsourced Metadata

*User switches*
Everybody can see changes

<table>
<thead>
<tr>
<th>Remote</th>
<th>Current</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable Name</td>
<td>totearn_ser_YYYY</td>
</tr>
<tr>
<td>Label</td>
<td>SER: Capped Earnings from all FICA-covered jobs</td>
</tr>
<tr>
<td>Codebook</td>
<td>SIPP Synthetic Beta v6.02</td>
</tr>
<tr>
<td>Concept</td>
<td>Concept</td>
</tr>
<tr>
<td>Concept Vocabulary</td>
<td>Concept Vocabulary</td>
</tr>
<tr>
<td>Concept Vocabulary URI</td>
<td>Concept Vocabulary URI</td>
</tr>
<tr>
<td>Type</td>
<td>numeric</td>
</tr>
<tr>
<td>Files</td>
<td>Files</td>
</tr>
</tbody>
</table>

```
sub_v6.0.2_syntheticK_M.sas7bdst
SAS

sub_v6.0.2_syntheticK_M.dta
Stata
```

Question Text

Person-level annual earnings that were taxed by FICA; these variables include earnings only up to the FICA taxable maximum and cover the years 1951-2011. These earnings are the inputs for calculating the OASDI benefit a person and his or her spouse will receive upon retirement or disability.

Notes (0 total)

1. Having a 0 value on totearn_ser_YYYY could mean a couple of things: 1) this individual had no FICA-covered earnings in that year; 2) this individual had no labor income at all in this tax year; 3) this individual worked for an employer that failed to report earnings in this year (that is to say, this has nothing to do with whether a person filed taxes because the earnings are reported by the employer, not the employee). Prior to 1978, if a person has 0 earnings on the Summary Earnings Record (SER), there's really no way of knowing whether they had no earnings or whether they had non-FICA earnings because the SER only reports FICA-covered earnings reported by employers. For years 1978 and later, you can...
Combining Knowledge: Merging

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

current_enroll_coll

<table>
<thead>
<tr>
<th>Crowdsourced Documentation</th>
<th>Official Documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable Name</strong></td>
<td>current_enroll_coll</td>
</tr>
<tr>
<td><strong>Label</strong></td>
<td>Currently Enrolled in College</td>
</tr>
<tr>
<td><strong>Codebook</strong></td>
<td>SIPP Synthetic Beta v6</td>
</tr>
<tr>
<td><strong>Concept</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Concept Vocabulary</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Concept Vocabulary URI</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>numeric</td>
</tr>
<tr>
<td><strong>Files</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Variable Name</strong></td>
<td>current_enroll_coll</td>
</tr>
<tr>
<td><strong>Label</strong></td>
<td>Currently Enrolled</td>
</tr>
<tr>
<td><strong>Codebook</strong></td>
<td>SIPP Synthetic Beta v6</td>
</tr>
<tr>
<td><strong>Concept</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Concept Vocabulary</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Concept Vocabulary URI</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>numeric</td>
</tr>
<tr>
<td><strong>Files</strong></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 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) Form W-2 records 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...
• Contributors can be tracked for each of their changes

<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</td>
<td>View commit</td>
<td><a href="mailto:lori.reeder@gmail.com">lori.reeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td></td>
<td>AM</td>
<td></td>
<td></td>
<td>Change</td>
</tr>
<tr>
<td>phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015</td>
<td>View commit</td>
<td><a href="mailto:lori.reeder@gmail.com">lori.reeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td></td>
<td>AM</td>
<td></td>
<td></td>
<td>Change</td>
</tr>
<tr>
<td>pos_phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015</td>
<td>View commit</td>
<td><a href="mailto:lori.reeder@gmail.com">lori.reeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td></td>
<td>AM</td>
<td></td>
<td></td>
<td>Change</td>
</tr>
<tr>
<td>phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015</td>
<td>View commit</td>
<td><a href="mailto:lori.reeder@gmail.com">lori.reeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td></td>
<td>AM</td>
<td></td>
<td></td>
<td>Change</td>
</tr>
<tr>
<td>pos_phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015</td>
<td>View commit</td>
<td><a href="mailto:lori.reeder@gmail.com">lori.reeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td></td>
<td>AM</td>
<td></td>
<td></td>
<td>Change</td>
</tr>
<tr>
<td>pos_phus_retire_benefit_totamt</td>
<td>November 18, 2015</td>
<td>View commit</td>
<td><a href="mailto:lori.reeder@gmail.com">lori.reeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td></td>
<td>AM</td>
<td></td>
<td></td>
<td>Change</td>
</tr>
<tr>
<td>phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015</td>
<td>View commit</td>
<td><a href="mailto:lori.reeder@gmail.com">lori.reeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td></td>
<td>AM</td>
<td></td>
<td></td>
<td>Change</td>
</tr>
<tr>
<td>pos_phus_ssdii_benefit_totamt_k</td>
<td>November 18, 2015</td>
<td>View commit</td>
<td><a href="mailto:lori.reeder@gmail.com">lori.reeder@gmail.com</a></td>
<td>Remote</td>
</tr>
<tr>
<td></td>
<td>AM</td>
<td></td>
<td></td>
<td>Change</td>
</tr>
</tbody>
</table>
Combining Knowledge: Citations

Lars Vilhuber

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

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
Next steps
Making CED²AR v2 robust

• Addition of UTF-8 editing support (Spanish, French, Portuguese, etc.)
• Additional fields (link to survey questions, anything within DDI-C)
• Bug fixes
Making CED²AR scalable in V3

• Current implementation of CED²AR is packaged for a single server (=portable)
  – Already a scalable archive backend (git)
  – Could be fronted by a load balancer
  – But live server is not scalable

• Current implementation of CED²AR is tied to a limited schema – adding schema is hard
Thank you!
Questions?

ced2ar-devs-I@cornell.edu