Crowdsourcing Metadata – Challenges and Outlook

Montreal, 29 April 2016

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

Montreal, 29 April 2016

Lars Vilhuber (Cornell University)
Acknowledgements

Based on work with

• Benjamin Perry (formerly Cornell University)
• Venkata Kambhampaty (formerly Cornell University)
• Kyle Brumsted (McGill University)
• William Block (Cornell 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
What’s the problem?
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
Replicability is a mess...
Verification Is Important

• Falsifying data
  – Andrew Wakefield (autism and vaccines)
  – Yoshitaka Fujii (fabricated data in 172 out of 249 papers)
• “Believe it or not: how much can we rely on published data on potential drug targets?” doi:10.1038/nrd3439-c1
  – Drug maker cannot replicate more than 20-25% of findings
• “Why Most Published Research Findings Are False” Ioannidis JPA (2005) doi:10.1371/journal.pmed.0020124
Recent Replication Exercises

• Psychology:

OSC (2015) Replication Exercises

(100 studies)

“one-third to one-half of the original findings were also observed in the replication study”
Recent Replication Exercises

• Behavioral economics:

Camerer et al. (2016), “Evaluating replicability of laboratory experiments in economics,” Science 03 Mar 2016, DOI: 10.1126/science.aaf0918:
Camerer et al. (2016) Replication Exercises

(18 studies)

“significant effect in the same direction as the original study for 11 replications (61%)”
We took a different approach

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Kingi, Stanchi, Vilhuber (2016, unpublished)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)
  - Unsuccessful (28)
  - Successful (37)
- Non Confidential Data (65)

Kingi, Stanchi, Vilhuber (2016)
• New question:

Are the provided data and programs accessible?
### Table 3: Type of Confidential Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Admin local</th>
<th>Admin National</th>
<th>Admin Regional</th>
<th>Private Commercial</th>
<th>Private Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>2</td>
<td>8</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>2011</td>
<td>2</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>2013</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>19</td>
<td>5</td>
<td>9</td>
<td>5</td>
<td>44</td>
</tr>
</tbody>
</table>

### Table 4: Type of Access to Confidential Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Formal</th>
<th>Informal Commitment</th>
<th>Informal No Commitment</th>
<th>No Info</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>2011</td>
<td>3</td>
<td>0</td>
<td>10</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>2013</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>5</td>
<td>27</td>
<td>6</td>
<td>44</td>
</tr>
</tbody>
</table>
What’s part of the problem?
A: Lack of proper deposit
The old source of knowledge – and data!
It’s mental...
Where do your university’s Ph.D. theses go?
Or more likely
Where do your researchers’ data go?

• Their computer?
• Dropbox?
Where can researchers go?
Options are available
Options are available

- Earth and natural sciences

DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad has integrated data submission for a growing list of journals; submission of data from other publications is also welcome.

Browse for data
- Recently published
- Popular
- By author
- By journal

Search for data
- Submit data now
- How and why?

Advanced search

Latest from @datadryad

Tweets by @datadryad
Options are available

• Social and behavioral sciences

Three easy steps:
1. Name your project
2. Upload and describe files
3. Publish your data

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

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researchers</td>
<td>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?</td>
</tr>
<tr>
<td>Journals</td>
<td>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?</td>
</tr>
<tr>
<td>Developers</td>
<td>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?</td>
</tr>
<tr>
<td>Institutions</td>
<td>Establish a research data management solution for your community. Federate 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?</td>
</tr>
</tbody>
</table>
Three main problems
Three main problems

• Researchers don’t use them
  – When researchers use them, usage is suboptimal (but better than nothing!)

• Journals don’t point to them

• Don’t work for many researchers at all
Researchers don’t use them

Training? Incentives? Ease of use?
Some Case Studies
Self-archiving

https://sites.google.com/site/pastpart/home/programs

Page not found

We're sorry, but we were unable to locate the page you requested.

Here are some similar pages from this site:

- CV_suggested.pdf

(c) 2016 John M. Abowd and Lars Vilhuber
Or…

- “… to respond to the request, I had to turn on the old computer, quickly find the files on the HD, before it overheated and shut down.”
Problems When There Is a Will

• Storage on Google, Dropbox, etc. relies on personal payment of recurring cost by researcher

• Inadvertent reorganization leads to retrieval failure
A Good Example
Gentzkow, Shapiro, Sinkinson (2014)

  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>9:09 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>1:38 PM</td>
<td></td>
</tr>
</tbody>
</table>


Persistent URL: [http://doi.org/10.3886/E1361V3](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 .dats. 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”
A Self-serving Example
Abowd and Vilhuber (2012)

- Article:

- Appendix
  - Description at
    (note: no DOI!)
  - Tried to be careful about referencing data, but no DOIs available on any of the data
    - Even our own data (National QWI, 38MB compressed)
  - Only generic programs
  - Final dataset was too large – not accepted.
Abowd and Vilhuber (2012)

No confidential data were used in this paper. All public-use Quarterly Workforce Indicators data can be accessed from http://www.vrdc.cornell.edu/news/data/qwi-public-use-data/. The national indicators developed in this paper can be accessed from http://www.vrdc.cornell.edu/news/data/qwi-national-data/. We are grateful for the comments and suggestions of many of our colleagues, past and present, too numerous to list here and thus listed at the website above and in the working paper version of this article. The opinions expressed in this paper are those of the authors and not the U.S. Census Bureau nor any of the research sponsors.
Abowd and Vilhuber (2012)

• No citation of own data


National Estimates of Gross Employment and Job Flows from the Quarterly Workforce Indicators with Demographic and Industry Detail

John M. Abowd and Lars Vilhuber

Abstract

The Quarterly Workforce Indicators (QWI) are local labor market data produced and released every quarter by...
Abowd and Vilhuber (2011)

• Later went back and added a proper replication archive
• Done after the fact
• No way to link article to data archive
Replication data for: National estimates of gross employment and job flows from the Quarterly Workforce Indicators with demographic and industry detail

Abowd, John M., Vilhuber, Lars, 2014, "Replication data for: National estimates of gross employment and job flows from the Quarterly Workforce Indicators with demographic and industry detail", http://dx.doi.org/10.7910/DVN/27928, Harvard Dataverse, V2

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description

The Quarterly Workforce Indicators are local labor market data produced and released by the Bureau. Unlike any other local labor market series produced in the U.S., or the rest of the NBER series for workers (accession and separations), jobs (creations and destructions) and earnings, the Quarterly Workforce Indicators provide economic data for industry groups, and detailed geography (counties, Workforce Investment Areas). As well as experimental, unreleased block-level estimates. To compile the public-use data (and only those public-use data) to construct the first national important enhancement to existing series because they include demographic and industry compiled from data that have been integrated at the micro-level by the Longitudinal Emplo
Abowd and Vilhuber (2011)

• Replication archive is linked to the article

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Employment, Dynamics, Economics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic Classification</td>
<td>Economics</td>
</tr>
<tr>
<td>Related Publication</td>
<td></td>
</tr>
<tr>
<td>Producer</td>
<td>Labor Dynamics Institute (Cornell University) (LDI) [<a href="http://www2.irocd.cornell.edu/news/data/qwi-national-data/">http://www2.irocd.cornell.edu/news/data/qwi-national-data/</a>]</td>
</tr>
</tbody>
</table>

(c) 2016 John M. Abowd and Lars Vilhuber
Abowd and Vilhuber (2011)

But no way to link the article back to the data (post-publication)
Journals don’t use them

… but that might be changing
Journals are starting to use them.
Journals are starting to use them
Journals are starting to use them
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?
The problem cannot be solved by researchers

- The data owners must be part of the solution
  - Statistical agencies
  - Government department and ministries
  - Private companies
- Must be able to designate a trusted data custodian
Some are quite commendable
Even detailed information

<table>
<thead>
<tr>
<th>Full Title</th>
<th>General Social Survey, 2011: Cycle 25, Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtitle</td>
<td>Cycle 25, Family</td>
</tr>
<tr>
<td>Alternative Title</td>
<td>GSS 2011: Family</td>
</tr>
<tr>
<td>Parallel Title</td>
<td>Encuesta social general, 2011: Cycle 25, Famille</td>
</tr>
<tr>
<td>Identification Number</td>
<td>ca-statcan-56195</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Authoring Entity</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistics Canada</td>
</tr>
<tr>
<td></td>
<td>StatCan</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Producer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statistics Canada</td>
</tr>
<tr>
<td></td>
<td>StatCan</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Copyright</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Copyright © Statistics Canada, 2012</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date of Distribution</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2012-07-18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Series Information</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>General Social Survey - Family (GSS) [4501]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Version</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15769.6</td>
</tr>
</tbody>
</table>
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

LITERAL QUESTION
In what year did your parents separate?

<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>

SUMMARY STATISTICS
This variable is numeric

UNIVERSE
Respondents who answered: PA_Q230 = 1.

NOTES
This variable is suppressed on the public use microdata file.
Let me take stock
Problems

- Users don’t curate data
  - When they do, it’s not very good or reliable
- Data providers don’t always curate data
  - Or don’t expose their metadata
  - Or don’t have metadata
- Once out there, it’s as if it were on paper – immutable, and cannot be improved
  - Links between articles and data
  - Improvements to documentation of data
What can data librarians do?
Don’t (just) liberate the data!

Liberate the data users!
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
Provide tools for users to link data to articles
RD Switchboard is under development
But relies on existing metadata
OpenAIRE attempts to do so

• … allowing users to establish the link
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
NBER CES
National QWI
SSB
SynLBD

Search
Searching all codebooks. No filters active.

Advanced Search
Show 10 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

Internal Metadata → Datasets
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
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: 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 ... 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

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Label</th>
<th>Top Access Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>afdc_MN</td>
<td>afdcamt_MN</td>
<td>released</td>
</tr>
<tr>
<td>afdcamt_MN</td>
<td></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
## Versioning

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>0f4a515</td>
<td><code>{ssbv601,lars@vilhuber.com,cover}</code></td>
<td>37 minutes ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>5c824de</td>
<td><code>{ssbv601,lars@vilhuber.com,cover}</code></td>
<td>an hour ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>c0358f</td>
<td>Committing codebooks retrieved directly from BaseX</td>
<td>4 days ago</td>
</tr>
<tr>
<td>venkata</td>
<td>a61abe3</td>
<td><code>{testbdv1,anonymous,edit}</code></td>
<td>4 days ago</td>
</tr>
<tr>
<td>venkata</td>
<td>501e51e</td>
<td><code>{testbdv1,anonymous,edit}</code></td>
<td>4 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>5edbff9</td>
<td><code>{acs2009,bap63@cornell.edu,edit}</code></td>
<td>5 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>d66d3c4</td>
<td><code>{ssbv601,lorraineer@ymail.com,reg,plus_ssn,benefit_totamt_k_24375}</code></td>
<td>5 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>1f845c1</td>
<td><code>{siabv1,warren.brown48@gmail.com,cover}</code></td>
<td>5 days ago</td>
</tr>
<tr>
<td>tomcat7</td>
<td>cf77f31</td>
<td><code>{siabv1,warren.brown48@gmail.com,cover,bild}</code></td>
<td>2015-11-17</td>
</tr>
<tr>
<td>tomcat7</td>
<td>b34a11b</td>
<td><code>{siabv1,warren.brown48@gmail.com,var,persnr}</code></td>
<td>2015-11-17</td>
</tr>
<tr>
<td>venkata</td>
<td>2cb6d7d</td>
<td><code>{lbdv2,anonymous,cover}</code></td>
<td>2015-11-17</td>
</tr>
<tr>
<td>tomcat7</td>
<td>1250bf</td>
<td><code>{siabv1,warren.brown48@gmail.com,reg,plus_ssn,benefit_totamt_k_24375}</code></td>
<td>2015-11-17</td>
</tr>
<tr>
<td>tomcat7</td>
<td>aaf94f1</td>
<td><code>{siabv1,bap63@cornell.edu,edit}</code></td>
<td>2015-11-17</td>
</tr>
<tr>
<td>tomcat7</td>
<td>6e52a6e</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

Crowdsourced Metadata

Datasets

Internal Metadata
Basic Information Flow

**Staging Area**
- Datasets
- Internal Metadata

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

*User switches*
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 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/

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\(^2\)AR handles identity management
Editing made easy
SIPP Synthetic Beta v6.02

View Variables (123 variables)
View codebook score
Last update to metadata: 2016-01-26 14:36:26 (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/


Variable Name  
totearn_ser_YYYY

Top Access Level  
released

Label  
SER: Capped Earnings from all FICA-covered jobs
Access Level: released

Codebook  
SIPP Synthetic Beta v6.02

Concept  

Type  
numeric

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.

Full Description  

Files  
ssb_v6_0_synthetic1_1.sas7bdat  http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html (SAS)
ssb_v6_0_synthetic1_1.dta  http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html (Stata)
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, 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, 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).
Basic Information Flow

Staging Area

- Datasets
- Internal Metadata

Public Facing

- Official Metadata
- Crowdsourced Metadata

User switches
Everybody can see changes

Remote

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>toetarn_ser_YYYY</th>
</tr>
</thead>
<tbody>
<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></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>

- sub_x_v6.0_2_syntheticK_M.sas7dat
  - SAS

- sub_x_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)

Current

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>toetarn_ser_YYYY</th>
</tr>
</thead>
<tbody>
<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></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>

- sub_x_v6.0_synthetic1.1.sas7dat
  - SAS

- sub_x_v6.0_synthetic1.1.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 (1 total)

#1

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
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 158, 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

*Use crownsourced* □ | *Use original* □

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
Combining Knowledge: Citations

- 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:lorireeder@gmail.com">lorireeder@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:lorireeder@gmail.com">lorireeder@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:lorireeder@gmail.com">lorireeder@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:lorireeder@gmail.com">lorireeder@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:lorireeder@gmail.com">lorireeder@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:lorireeder@gmail.com">lorireeder@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:lorireeder@gmail.com">lorireeder@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:lorireeder@gmail.com">lorireeder@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

> 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
Preferred source
Demo

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

Search Variables  Browse Variables   Browse by Codebook  Documentation  About

You are viewing the official metadata. View crowdsourced contributions.

CED²AR / 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://ciser.cornell.edu

Citation

Please cite this codebook as:

Please cite this dataset as:

Try for yourself: http://demo.ncrn.cornell.edu
Where does this leave us?
Making metadata easier to create

• For researchers:
  – Training
  – Better knowledge dissemination

• For data providers
  – Faster provision of data
  – Possibility of letting users document data in structured fashion

• For data curators
  – Delegation of work in a productive way to data experts
  – Control over workflow of enhancements
Things we didn’t solve

- Catch-22 of data providers who cannot release data to archives (role for libraries, RDCs)
- Making better programmers out of social scientists (on average)
- What to have for lunch
Thank you!

Questions?

ced2ar-devs-l@cornell.edu