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?

CED²AR

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 CED²AR?

CED²AR

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

Compare Variables

No variables selected

Search

Searching all codebooks. No filters active.

Search

Advanced Search

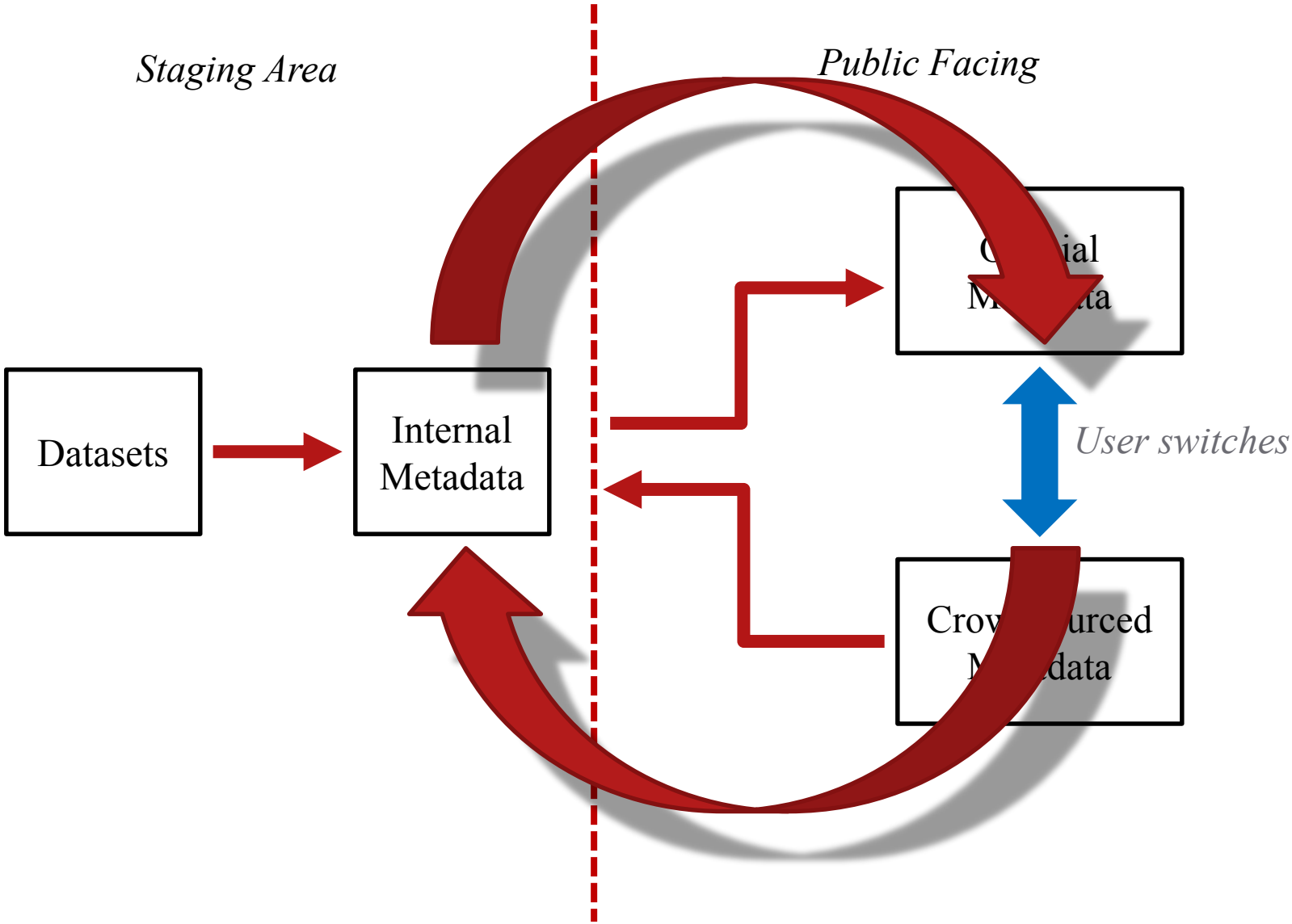
Show 10 variables

(cc) BY-NC-SA

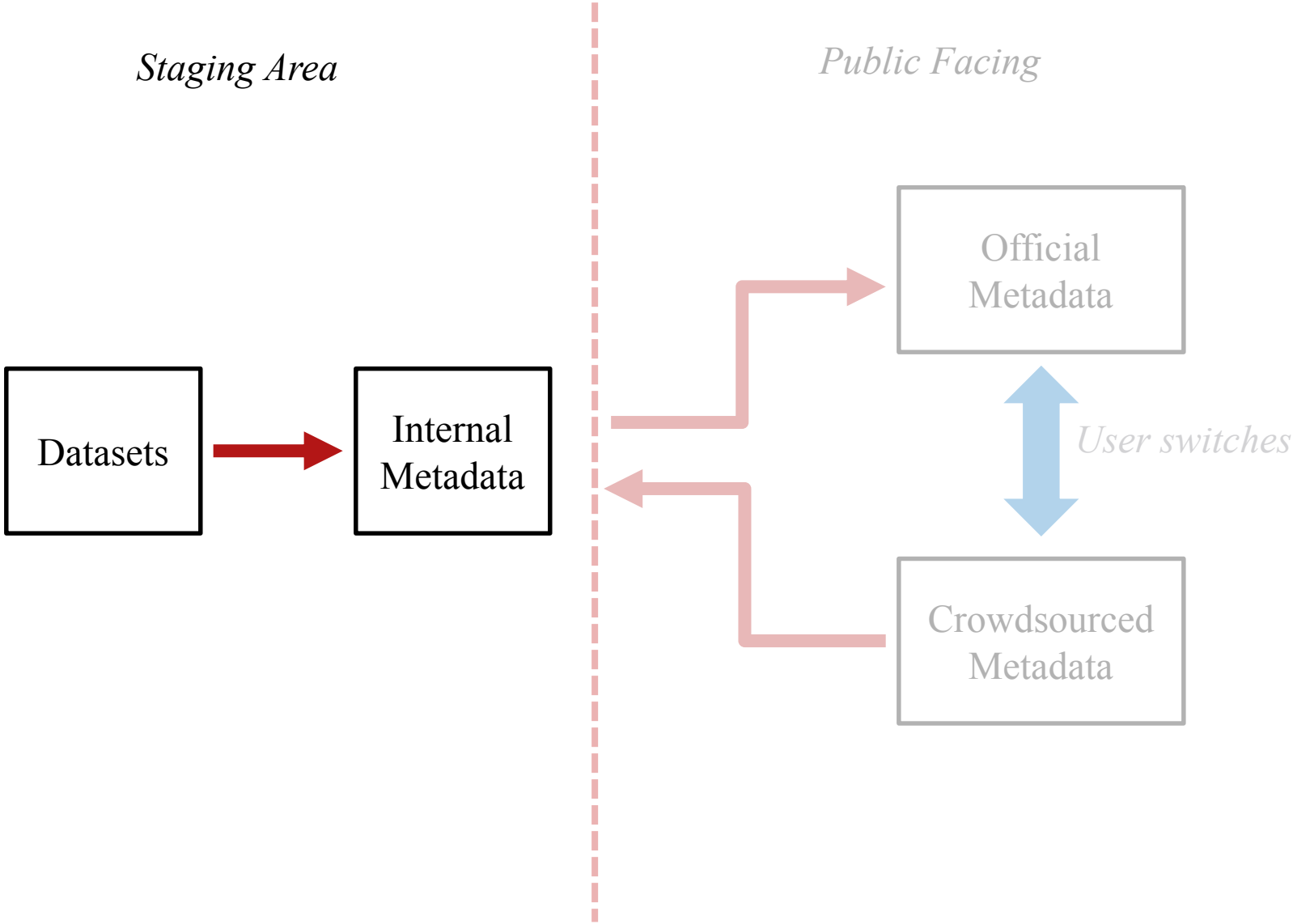
© 2012-2015, Cornell Institute for Social and Economic Research

Report a Bug | Email us | Copyright Information | NCRN GitHub

Basic Information Flow



Basic Information Flow



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

Save

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.

p

This field supports ASCII math See FAQ for details.

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

Internal Processing: Access Control

- Marking elements with different restrictions

Select what sub-elements to mark

- [ ] Select All
  - [ ] Mean
  - [ ] Median
  - [ ] Mode
  - [ ] Valid
  - [ ] Invalid
  - [x] Min
  - [x] Max
  - [ ] Standard Deviation
  - [ ] Other Summary Statistics
  - [ ] Values
  - [x] Value Frequencies
  - [x] Value Percentages
  - [x] Value Crosstabs
  - [x] Other Value Statistics
  - [ ] Label
  - [ ] Notes

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

restricted ▾ Change Levels

	Variable Name	Label	Top Access Level
[x]	afdc_MN	Indicator for receipt of AFDC or TANF benefits	released
[x]	afdcamt_MN	Amount of AFDC received	released
[ ]	birthdate	Date of Birth	released
[ ]	current_enroll_coll	Currently Enrolled in College	released
[ ]	current_enroll_hs	Currently Enrolled in HS (or less)	released

Internal Processing: Scoring

- Provide feedback to improve sparse documentation

CED2AR / SIPP Synthetic Beta v6 / Score

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%

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

SIPP Synthetic Beta v5.1

View Variables *(102 variables)*
Last update to metadata: 2014-11-13 10:38:45 (auto-generated)
Document Date: June 19th 2014

Codebook prepared by: Cornell NSF Census Research Network

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

Combining Knowledge

Viewing changes made

Codebook	Git Status	Last Local Update	Last Remote Update
ssb.v602.xml	UP_TO_DATE	September 10, 2015 at 9:25 AM {ssbv602,fs379@cornell.edu,cover}	September 10, 2015 at 9:25 AM {ssbv602,fs379@cornell.edu,cover}
ssb.v6.xml	UP_TO_DATE	November 12, 2015 at 4:16 PM {ssbv601,lorireeder@gmail.com,var,layoff_M} {ssbv601,lorireeder@gmail.com,var,fsamt_M} {ssbv601,lorireeder@gmail.com,var,hicov_M} {ssbv601,lorireeder@gmail.com,var,hiemp_M} {ssbv601,lorireeder@gmail.com,var,hispanic} {ssbv601,lorireeder@gmail.com,var,homeequity} {ssbv601,lorireeder@gmail.com,var,ind_4cat} {ssbv601,lorireeder@gmail.com,var,ind_exist} {ssbv601,lorireeder@gmail.com,var,defer_der_fica_YYYY} {ssbv51,lorireeder@gmail.com,var,defer_der_nonfica_YYYY} {ssbv6,lorireeder@gmail.com,var,defer_der_nonfica_YYYY}	November 12, 2015 at 4:16 PM {ssbv601,lorireeder@gmail.com,var,layoff_M} {ssbv601,lorireeder@gmail.com,var,fsamt_M} {ssbv601,lorireeder@gmail.com,var,hicov_M} {ssbv601,lorireeder@gmail.com,var,hiemp_M} {ssbv601,lorireeder@gmail.com,var,hispanic} {ssbv601,lorireeder@gmail.com,var,homeequity} {ssbv601,lorireeder@gmail.com,var,ind_4cat} {ssbv601,lorireeder@gmail.com,var,ind_exist} {ssbv601,lorireeder@gmail.com,var,defer_der_fica_YYYY} {ssbv51,lorireeder@gmail.com,var,defer_der_nonfica_YYYY} {ssbv6,lorireeder@gmail.com,var,defer_der_nonfica_YYYY}

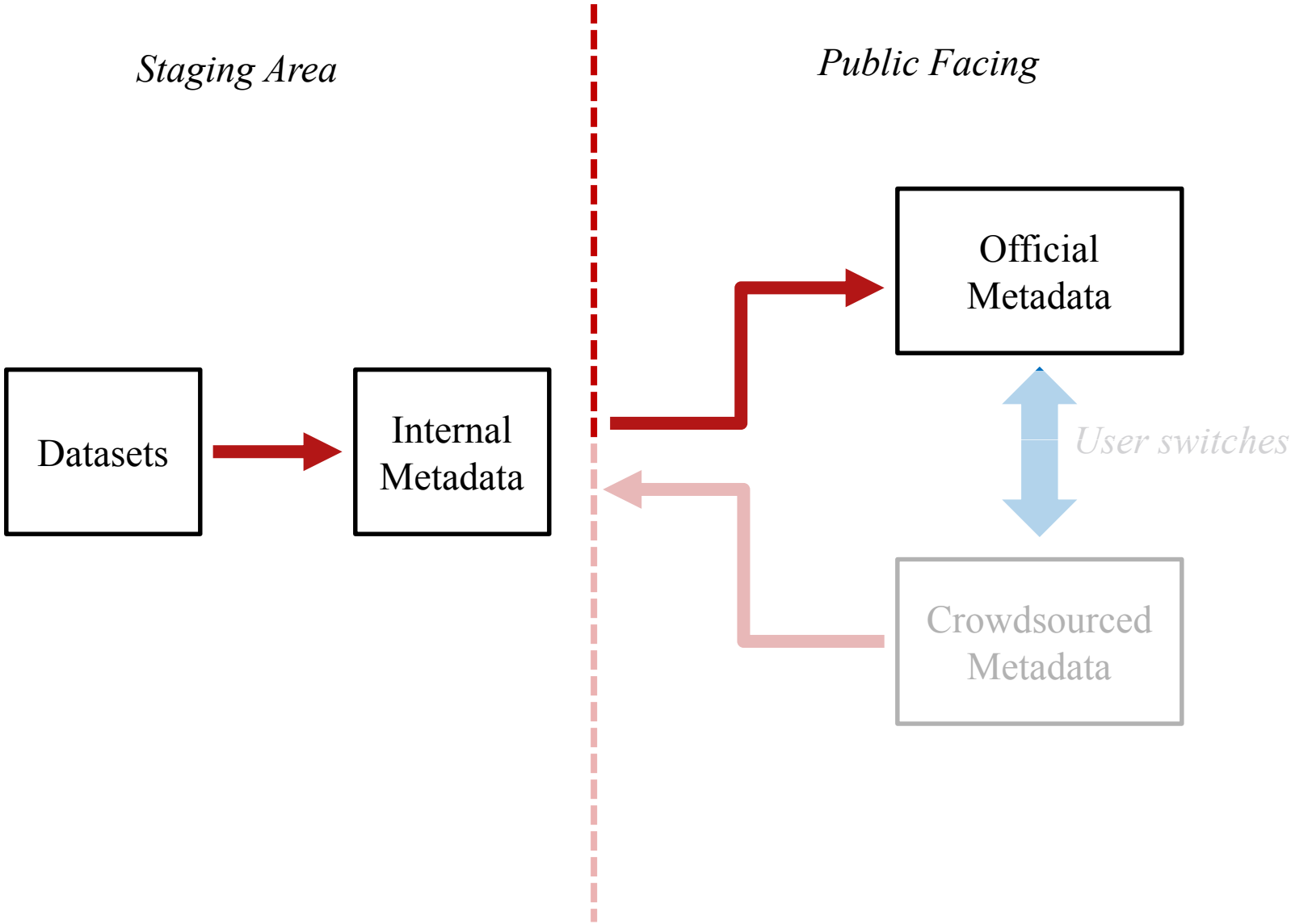
Versioning

All changes are logged externally via Git

Commits

Author	Commit	Message		Date
tomcat7	0fea515	{ssbv601,lars@vilhuber.com,cover}	ssbtesting	37 minutes ago
tomcat7	5e824de	{ssbv601,lars@vilhuber.com,cover}{ssbv601,lars@vilhuber.com,var,fl...	ssbtesting	an hour ago
tomcat7	c03c50f	Commiting codebooks retrieved directly from BaseX	cestesting	4 days ago
venkata	a61abe3	{testlbdv1,anonymous,edit}	vrk4	4 days ago
venkata	5b1e51e	{testlbdv1,anonymous,edit}	vrk4	4 days ago
tomcat7	5edbff9	{acs2009,bap63@cornell.edu,edit}	cestesting	5 days ago
tomcat7	d66d3d4	{ssbv601,lorireeder@gmail.com,var,phus_ssdi_benefit_totamt_k}{ss...	ssbtesting	5 days ago
tomcat7	1f845c1	{siabv1,warren.brown48@gmail.com,cover}{siabv1,bap63@cornell.e...	cestesting	5 days ago
tomcat7	eb77f31	{siabv1,warren.brown48@gmail.com,var,bild}{siabv1,warren.brown4...	cestesting	2015-11-17
tomcat7	b34a118	{siabv1,warren.brown48@gmail.com,var,persnr}{siabv1,warren.brow...	cestesting	2015-11-17
venkata	2cb6d7d	{lbdv2,anonymous,cover}	vrk4	2015-11-17
tomcat7	1263bcf	{siabv1,warren.brown48@gmail.com,var,bnn}{siabv1,warren.brown4...	cestesting	2015-11-17
tomcat7	aaf94f1	{siabv1,bap63@cornell.edu,edit}{blss2011,bap63@cornell.edu,edit}{...	cestesting	2015-11-17
tomcat7	0e52a6e	Commiting codebooks retrieved directly from BaseX	cestesting	2015-11-17

Basic Information Flow



Official view

CED²AR

Official Server - The Comprehensive Extensible Data Documentation and Access Repository

Search Variables Browse Variables

You are viewing the official / *crowdsourced contributions.*

CED2AR / SIPP Synthetic Beta

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.
http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html

Crowdsourced view

CED²AR

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

You are viewing crowdsourced metadata. View the official version .

SAS Stata

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

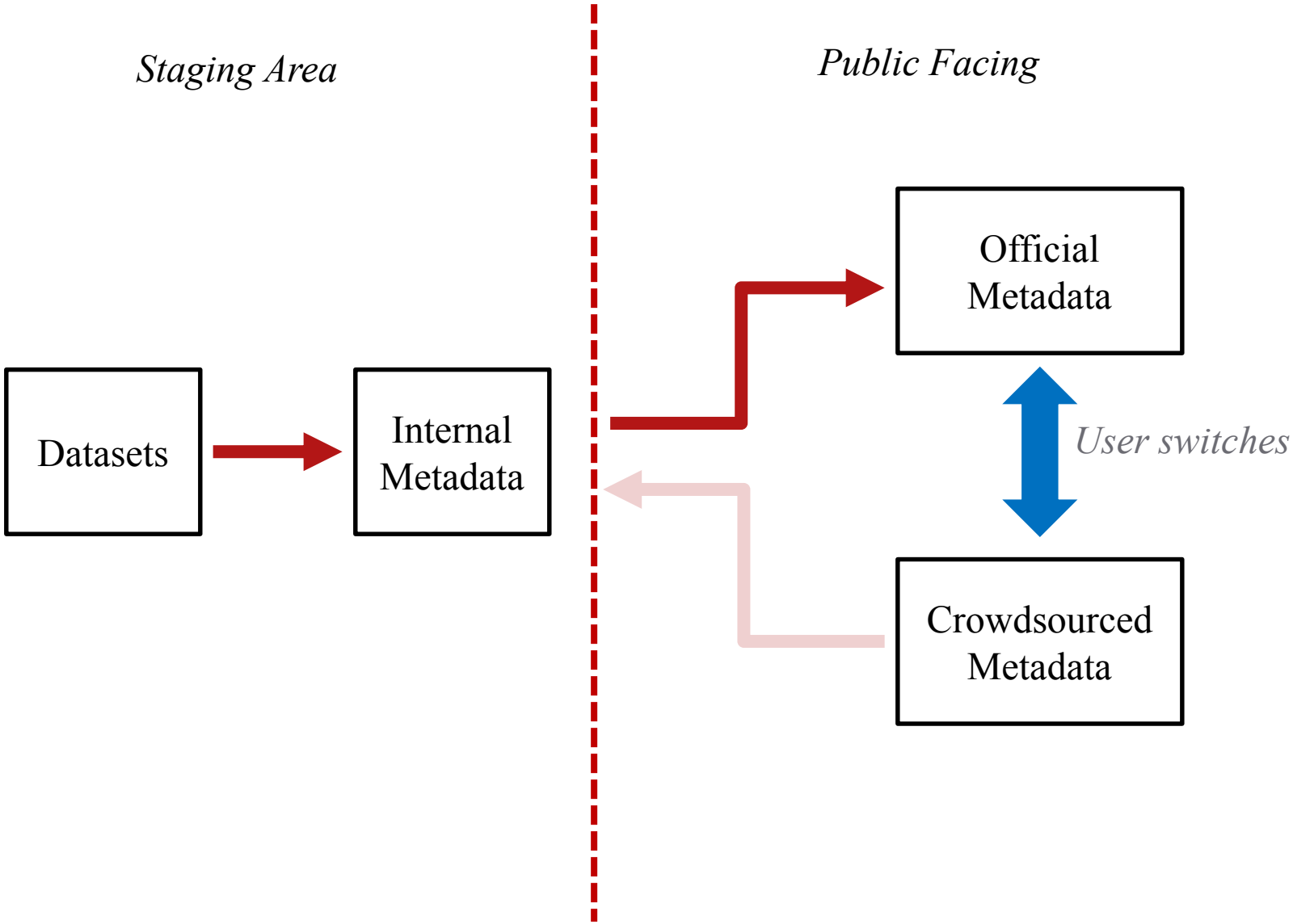
Data prepared by: United States Department of Commerce. Bureau of the Census.

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.
http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html

Basic Information Flow

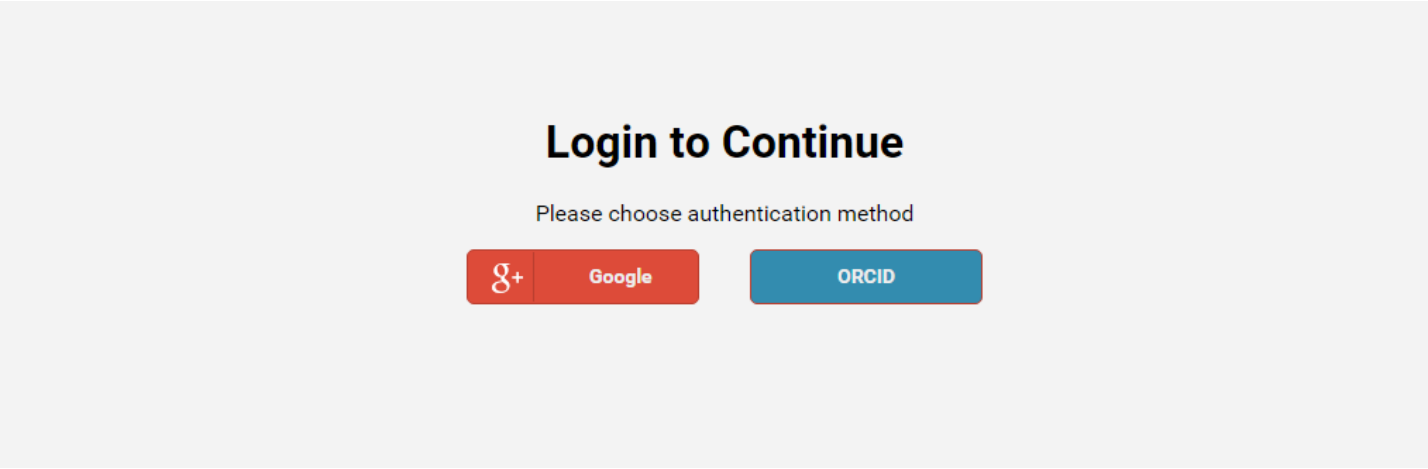


Authentication and Attribution

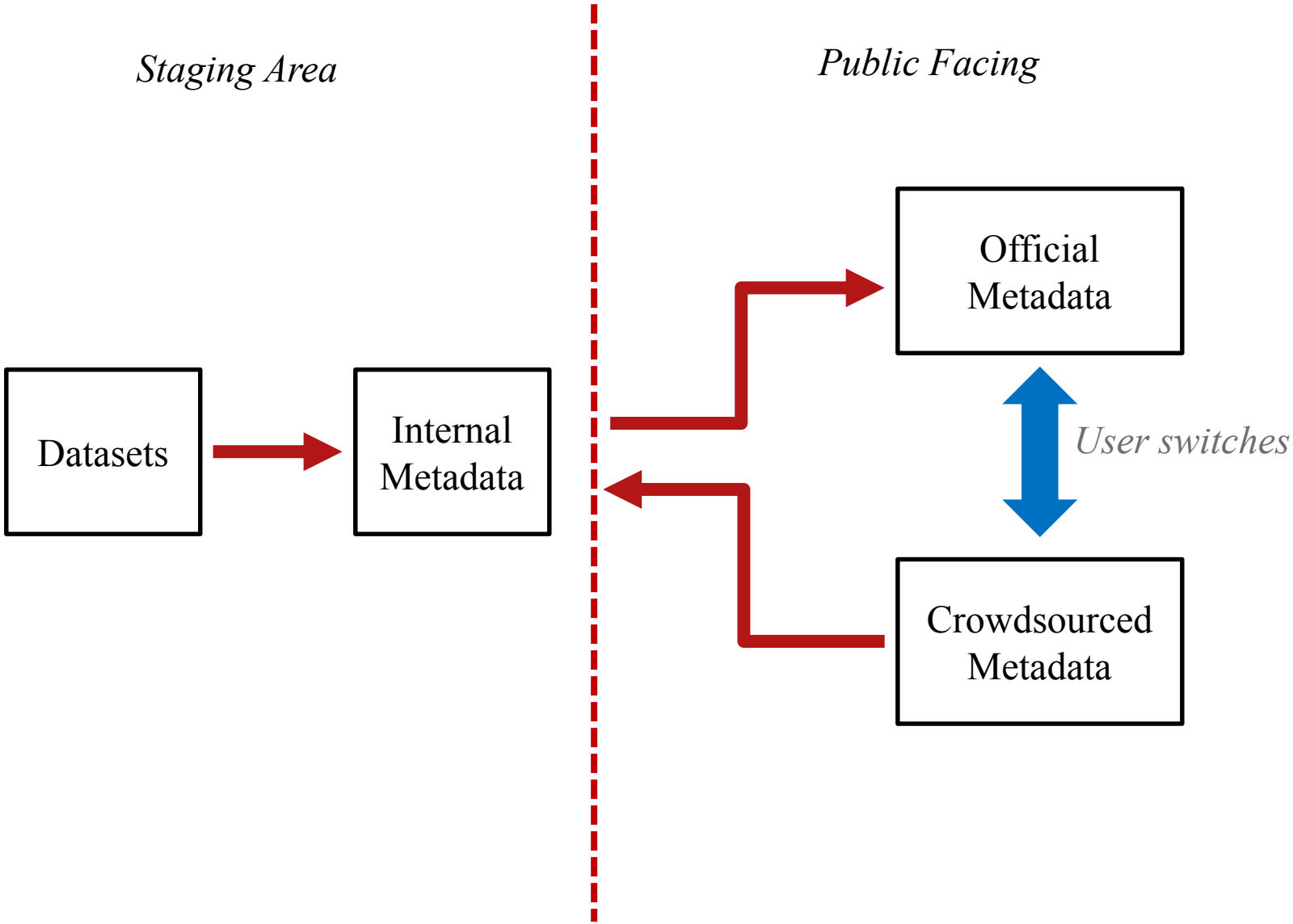
- When opening up contributions to a wide audience, how to triage between "rants" and meaningful contributions?
- Use of ORCID (academic network) for authentication
- Public attribution with link to (verified) academic ID is key for positive feedback (your effort is recognized) and prevention of negative 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
- CED^{2}AR handles identity management

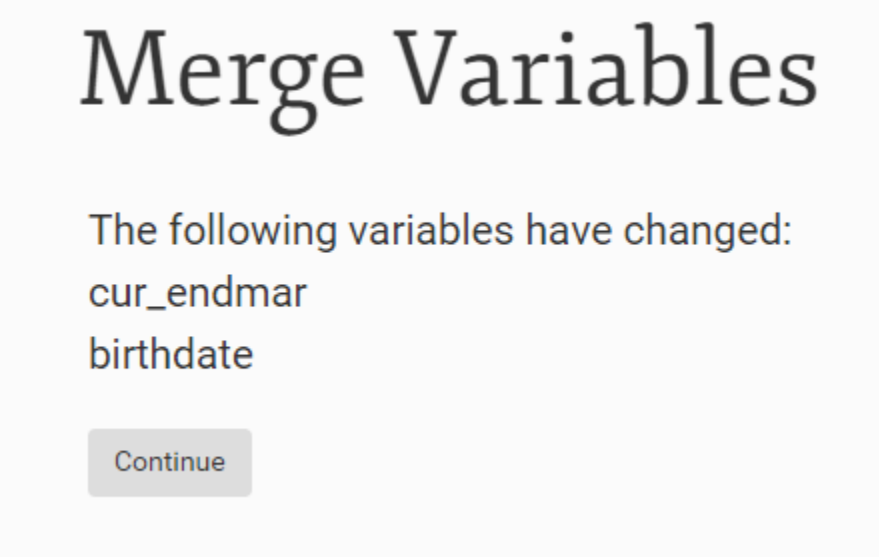


Basic Information Flow



Combining Knowledge: Merging

- Curators are given an interface to merge crowdsourced documentation with official



Combining Knowledge: Merging

current_enroll_coll

Crowdsourced Documentation

Variable Name	current_enroll_coll
Label	Currently Enrolled in College
Codebook	SIPP Synthetic Beta v6
Concept	
Concept Vocabulary	
Concept Vocabulary URI	
Type	numeric
Files	

Official Documentation

Variable Name	current_enroll_coll
Label	☐ Use crowdsourced ☐ Use original Currently Enrolled
Codebook	SIPP Synthetic Beta v6
Concept	
Concept Vocabulary	
Concept Vocabulary URI	
Type	numeric
Files	

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:

Comprehensive Extensible Data Documentation and Access Repository. Codebook for the SIPP Synthetic Beta 5.1 [Codebook file]. Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, 2013

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:

☐ Use crowdsourced ☐ Use original June 15, 2014

Citation

Please cite this codebook as:

Comprehensive Extensible Data Documentation and Access Repository. Codebook for the SIPP Synthetic Beta 5.1 [Codebook file]. Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, 2013

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 crowdsourced ☐ 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

Combining Knowledge: Citations

- Contributors can be tracked for each of their changes

CED2AR / SIPP Synthetic Beta v6.01 / Variable Versions

Modified Variables

Variable Name	Date Changed	Commit Message	User	Origin
phus_ssdi_benefit_totamt_k	November 18, 2015 at 11:15 AM	View commit	lorireeder@gmail.com	*Remote Change*
phus_ssdi_benefit_totamt_k	November 18, 2015 at 11:15 AM	View commit	lorireeder@gmail.com	*Remote Change*
pos_phus_ssdi_benefit_totamt_k	November 18, 2015 at 11:15 AM	View commit	lorireeder@gmail.com	*Remote Change*
phus_ssdi_benefit_totamt_k	November 18, 2015 at 11:15 AM	View commit	lorireeder@gmail.com	*Remote Change*
pos_phus_ssdi_benefit_totamt_k	November 18, 2015 at 11:15 AM	View commit	lorireeder@gmail.com	*Remote Change*
pos_phus_retire_benefit_totamt	November 18, 2015 at 11:15 AM	View commit	lorireeder@gmail.com	*Remote Change*
phus_ssdi_benefit_totamt_k	November 18, 2015 at 11:15 AM	View commit	lorireeder@gmail.com	*Remote Change*
pos_phus_ssdi_benefit_totamt_k	November 18, 2015 at 11:15 AM	View commit	lorireeder@gmail.com	*Remote Change*

Combining Knowledge: Citations

ORCID

1,757,580 ORCID iDs and counting. See more...

Lars Vilhuber

ORCID ID

orcid.org/0000-0001-5733-8932

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

Sort

CED^{2}AR: 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

SAS Stata

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:
Cornell University. Survey Research Institute. Cornell National Social Survey (CNSS), 2012[Computer file]. CISER version 1. Ithaca, NY: Cornell Institute for Social and Economic Research [producer and distributor], 2015

Please cite this dataset as:
Cornell University. Survey Research Institute. Cornell National Social Survey (CNSS), 2012[Computer file]. CISER version 1. Ithaca, NY: Cornell Institute for Social and Economic Research [producer and distributor], 2015

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-l@cornell.edu

Extra Slides

Technologies Used

What technologies are used to build CED^{2}AR?

- 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)