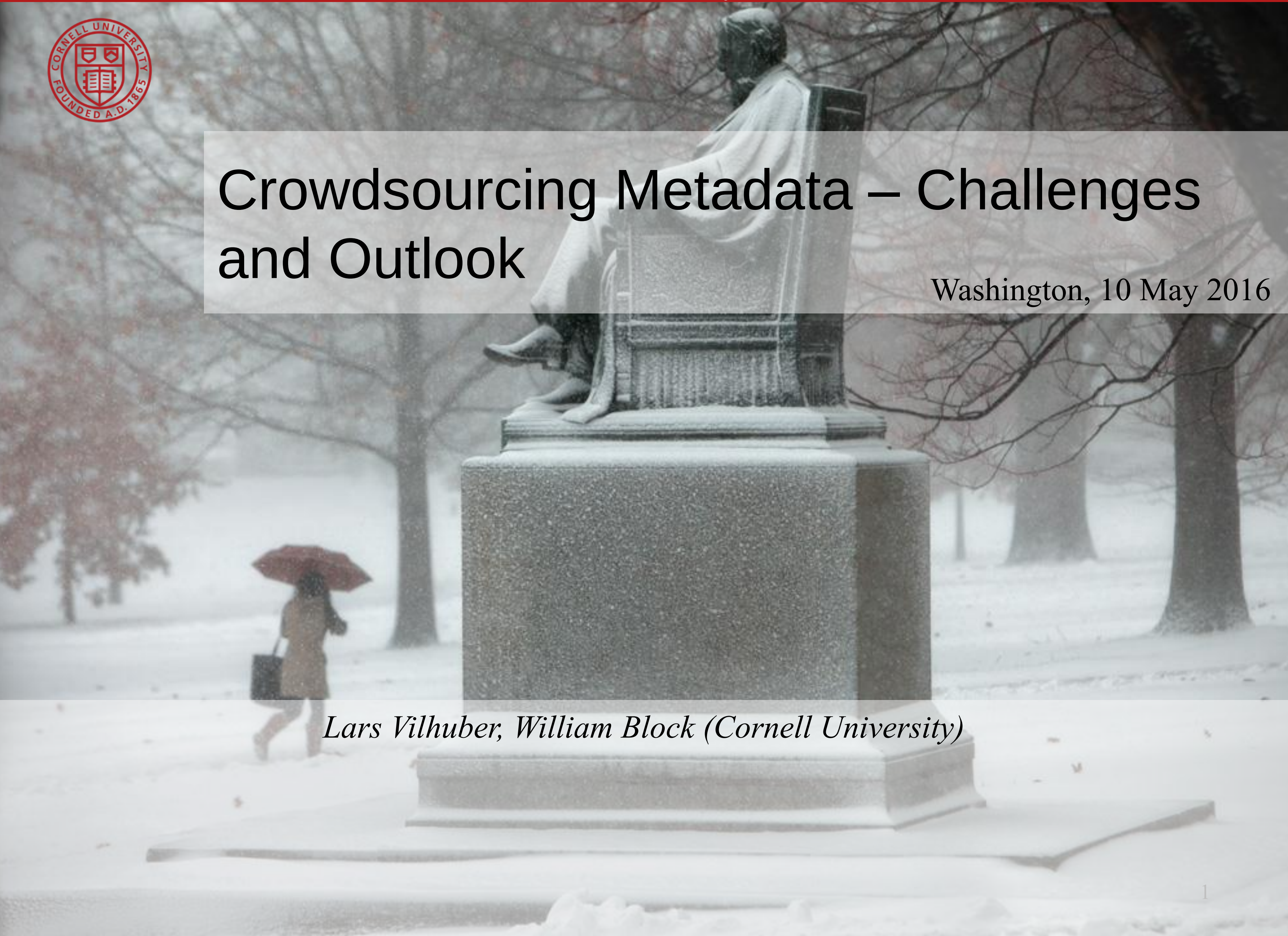
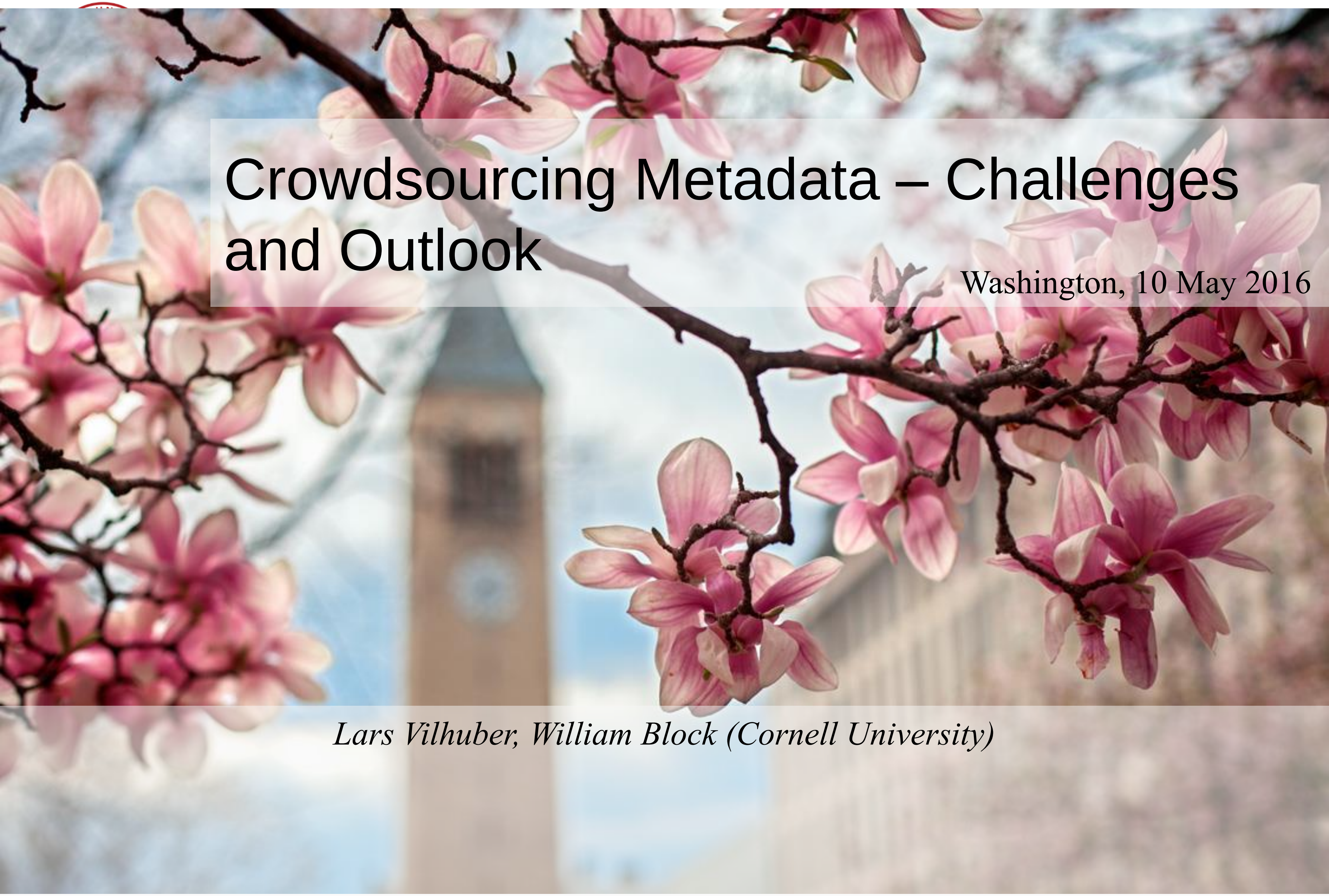
Cornell University



Crowdsourcing Metadata – Challenges and Outlook

Washington, 10 May 2016

Lars Vilhuber, William Block (Cornell University)



Crowdsourcing Metadata – Challenges and Outlook

Washington, 10 May 2016

Lars Vilhuber, William Block (Cornell University)

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Kingi, Stanchi, Vilhuber (2016, unpublished)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”

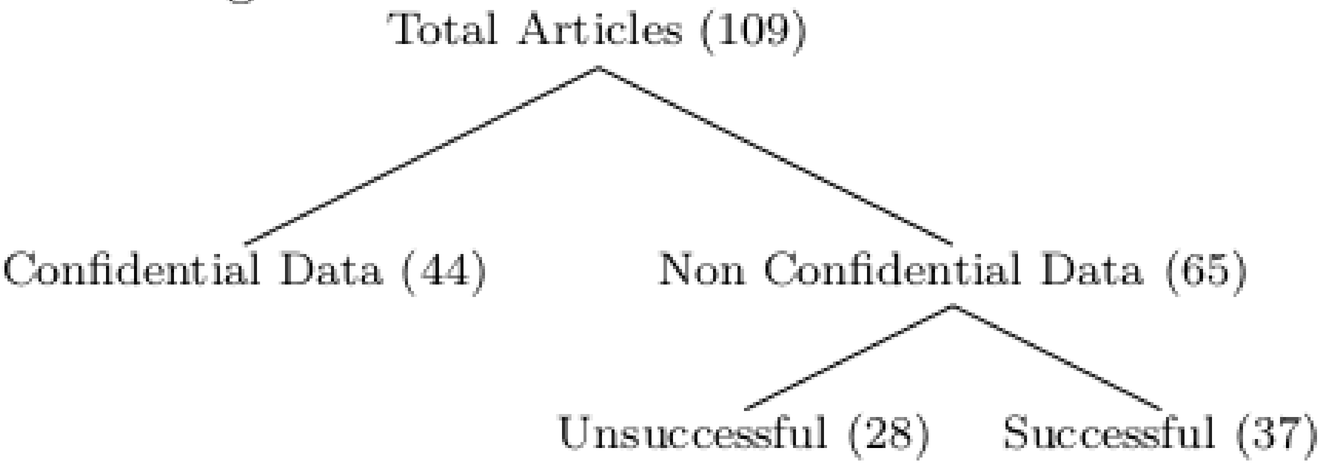
Kingi, Stanchi, Vilhuber (2016)

- 109 articles in American Economic Journal: Applied Economics
- Simpler test:

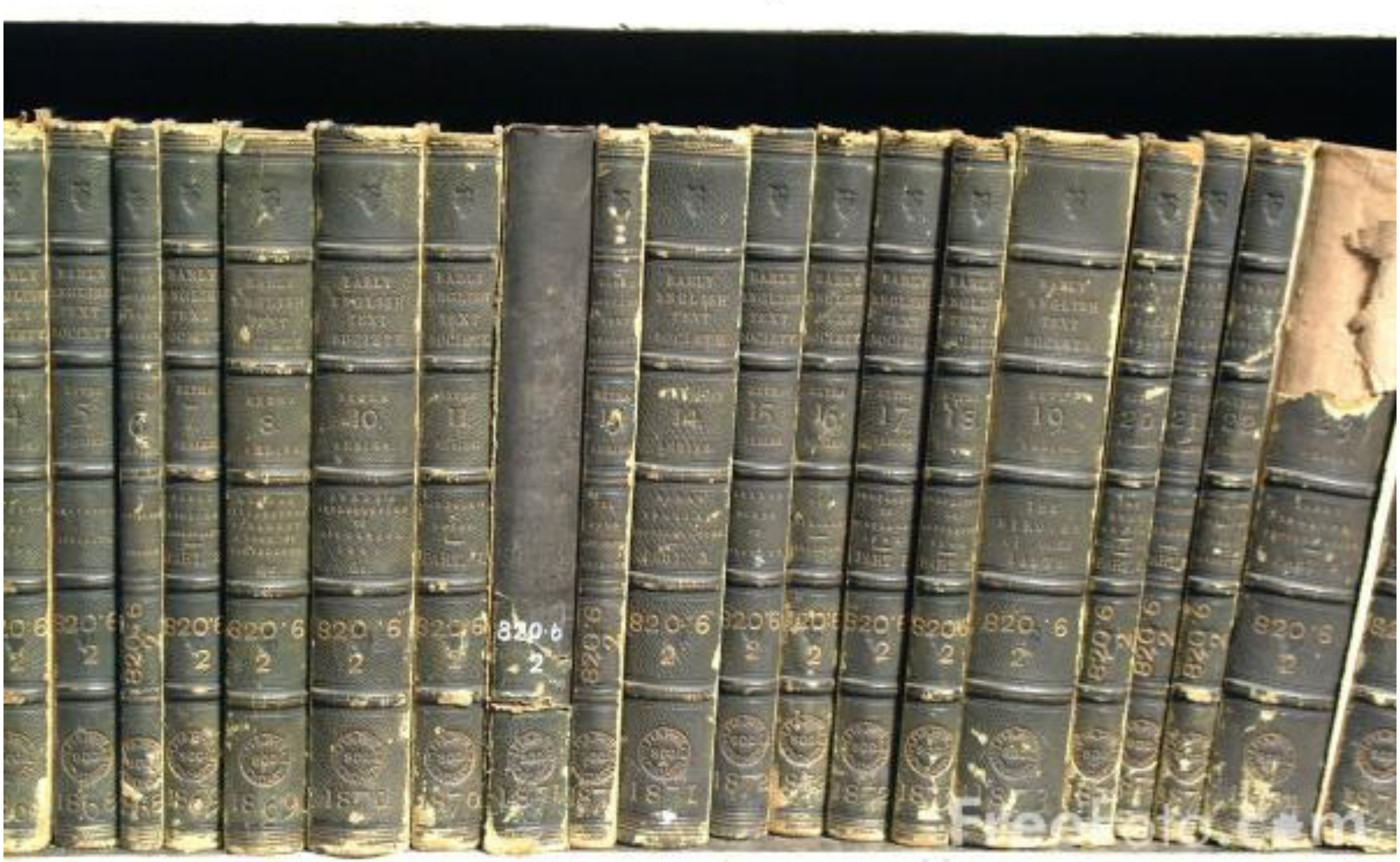
Do the provided data and programs yield the published results?

Kingi, Stanchi, Vilhuber (2016)

Figure 1: A Breakdown of the Articles



The old source of knowledge – and data!

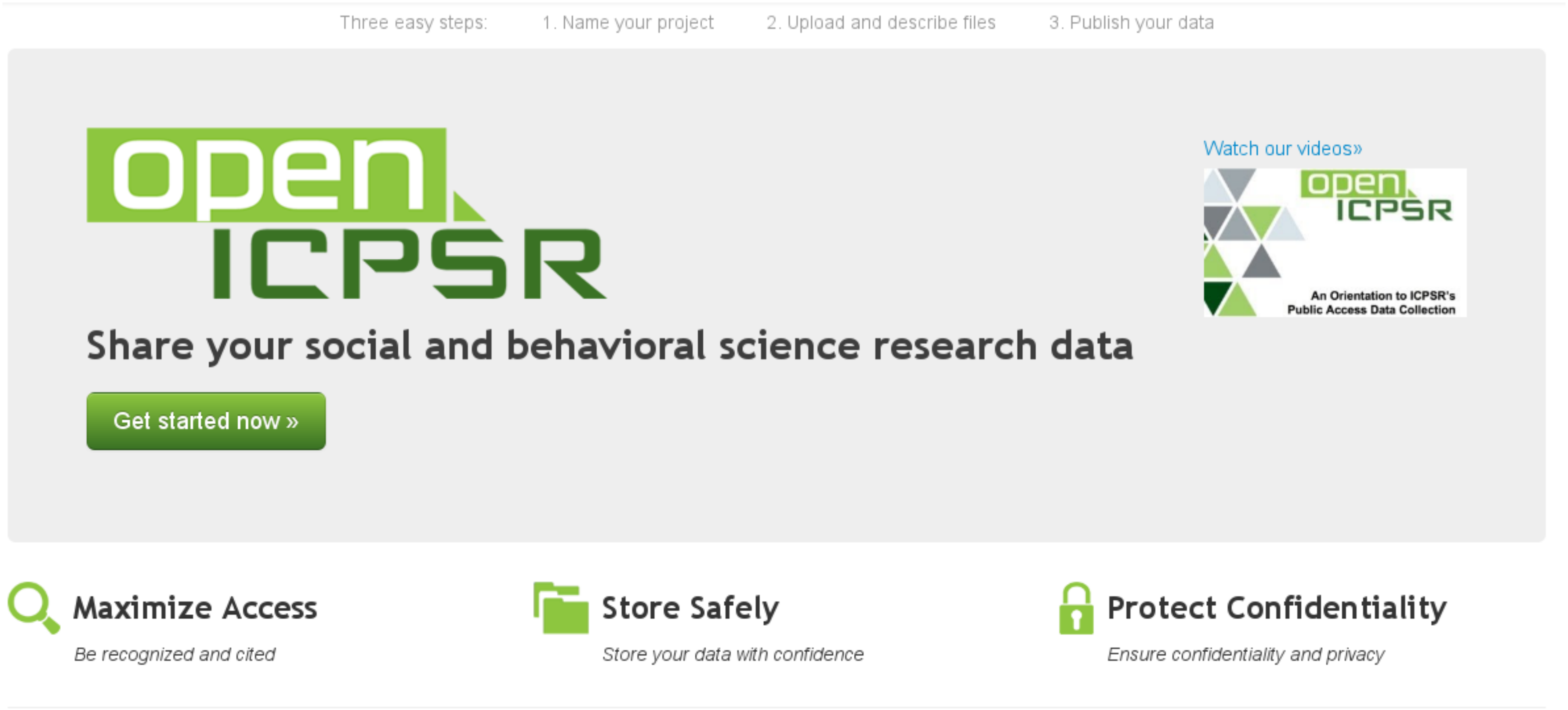


- Source: http://www.thecontrarianmedia.com/2011/08/dispatches-from-the-stacks-15/
- License unknown

https://www.flickr.com/photos/15472273@N07/6931305440/in/photolist-byuKCd-oh6Bxu-bMpoYn-aUxMUX-aUxPKF-aUxNwv-aUxNiH-aUxN2V-aUxNb6-aUxNqe-aUxP4F-aUxNLB-biAicF-biA9ar-aUxMte-biAftF-aUxLXr-aUxLKT-aUxMBc-fUMhDZ-pQm1F1-oaGC6t-6Juv4g-6JuiTM-oh78Mr-oyA1Gk-qssnzU-bMpunF-byuB5L-bMpf3n-qKP4GF-qKDrYF-dBNTRV-hEQRCt-Eag36-6Juv46-6Juv3R-6JxXpw-6Ju7Rt-6JxXpN-eJGN8U-6JuiTB-6JuiTx-6JuiU4-6JsUMD-6JsUMZ-6JuiTg-6Ju7R6-6Ju7RH-6JxXpo (Stadtbibliothek Stuttgart am Mailänder Platz) CC 2.0 (cc) BY

Options are available

- Social and behavioral sciences



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

openICPSR

Share your social and behavioral science research data

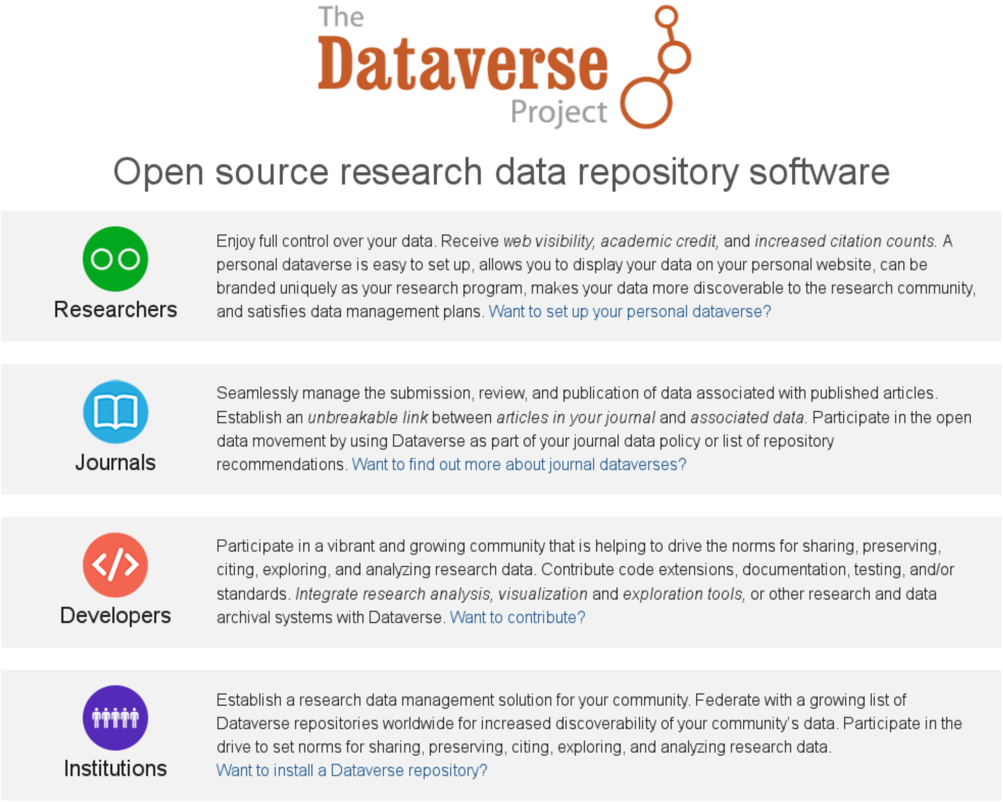
Maximize Access
Be recognized and cited

Store Safely
Store your data with confidence

Protect Confidentiality
Ensure confidentiality and privacy

Options are available

- “Research data”



The Dataverse Project

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. 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?

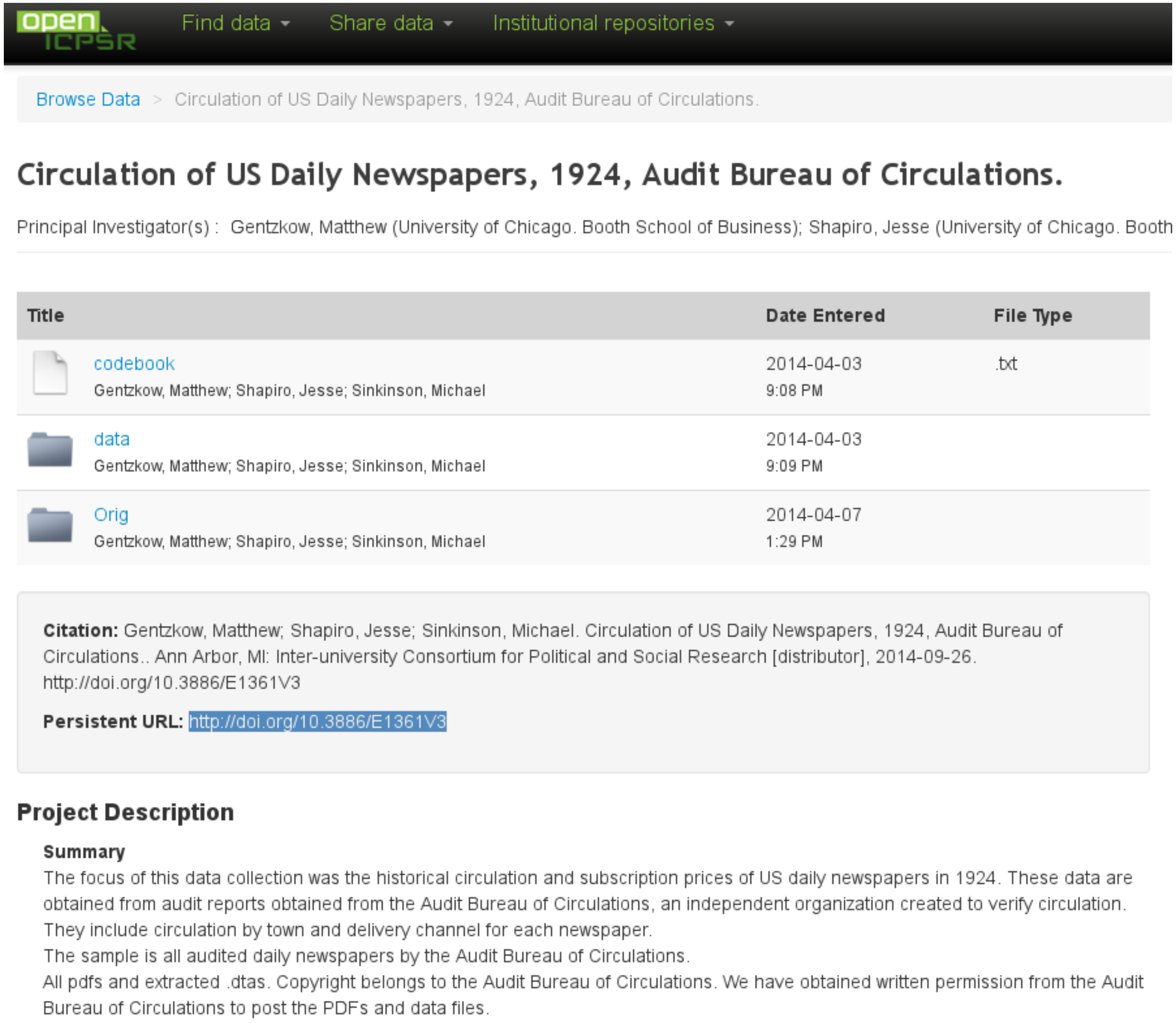
Researchers don’t use them

Training? Incentives? Ease of use?

A Good Example

Gentzkow, Shapiro, Sinkinson (2014)

- Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson. 2014. "*Competition and Ideological Diversity: Historical Evidence from US Newspapers.*"
  American Economic Review, 104(10): 3073-3114.
  DOI: 10.1257/aer.104.10.3073
- Data at http://doi.org/10.3886/E1361V3



Browse Data > Circulation of US Daily Newspapers, 1924, Audit Bureau of Circulations.

Circulation of US Daily Newspapers, 1924, Audit Bureau of Circulations.

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

Title	Date Entered	File Type
codebook Gentzkow, Matthew; Shapiro, Jesse; Sinkinson, Michael	2014-04-03 9:08 PM	.txt
data Gentzkow, Matthew; Shapiro, Jesse; Sinkinson, Michael	2014-04-03 9:09 PM	
Orig Gentzkow, Matthew; Shapiro, Jesse; Sinkinson, Michael	2014-04-07 1:29 PM	

Citation: Gentzkow, Matthew; Shapiro, Jesse; Sinkinson, Michael. Circulation of US Daily Newspapers, 1924, Audit Bureau of Circulations.. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2014-09-26. http://doi.org/10.3886/E1361V3

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

Project Description

Summary

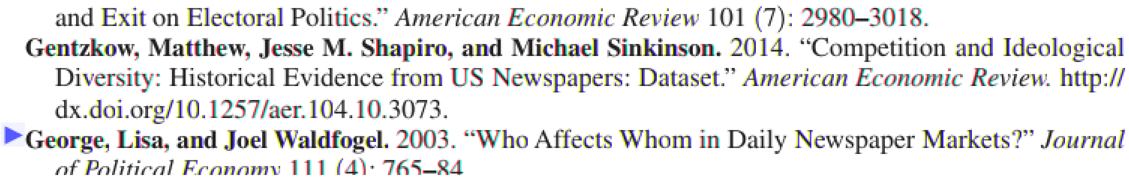
The focus of this data collection was the historical circulation and subscription prices of US daily newspapers in 1924. These data are obtained from audit reports obtained from the Audit Bureau of Circulations, an independent organization created to verify circulation. They include circulation by town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

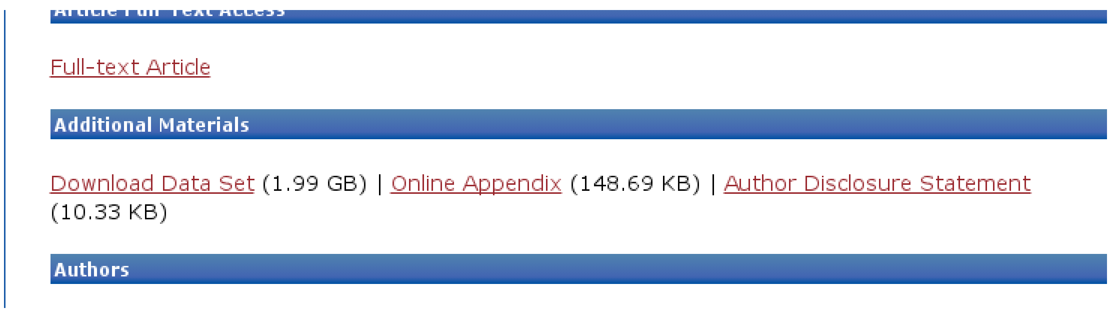
Not Perfect

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



and Exit on Electoral Politics." *American Economic Review* 101 (7): 2980–3018.
Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson. 2014. "Competition and Ideological Diversity: Historical Evidence from US Newspapers: Dataset." *American Economic Review.* http://dx.doi.org/10.1257/aer.104.10.3073.
George, Lisa, and Joel Waldfogel. 2003. "Who Affects Whom in Daily Newspaper Markets?" *Journal of Political Economy* 111 (4): 765–84

- Conversely, "online appendix" just a "blob"



Article Full Text Access

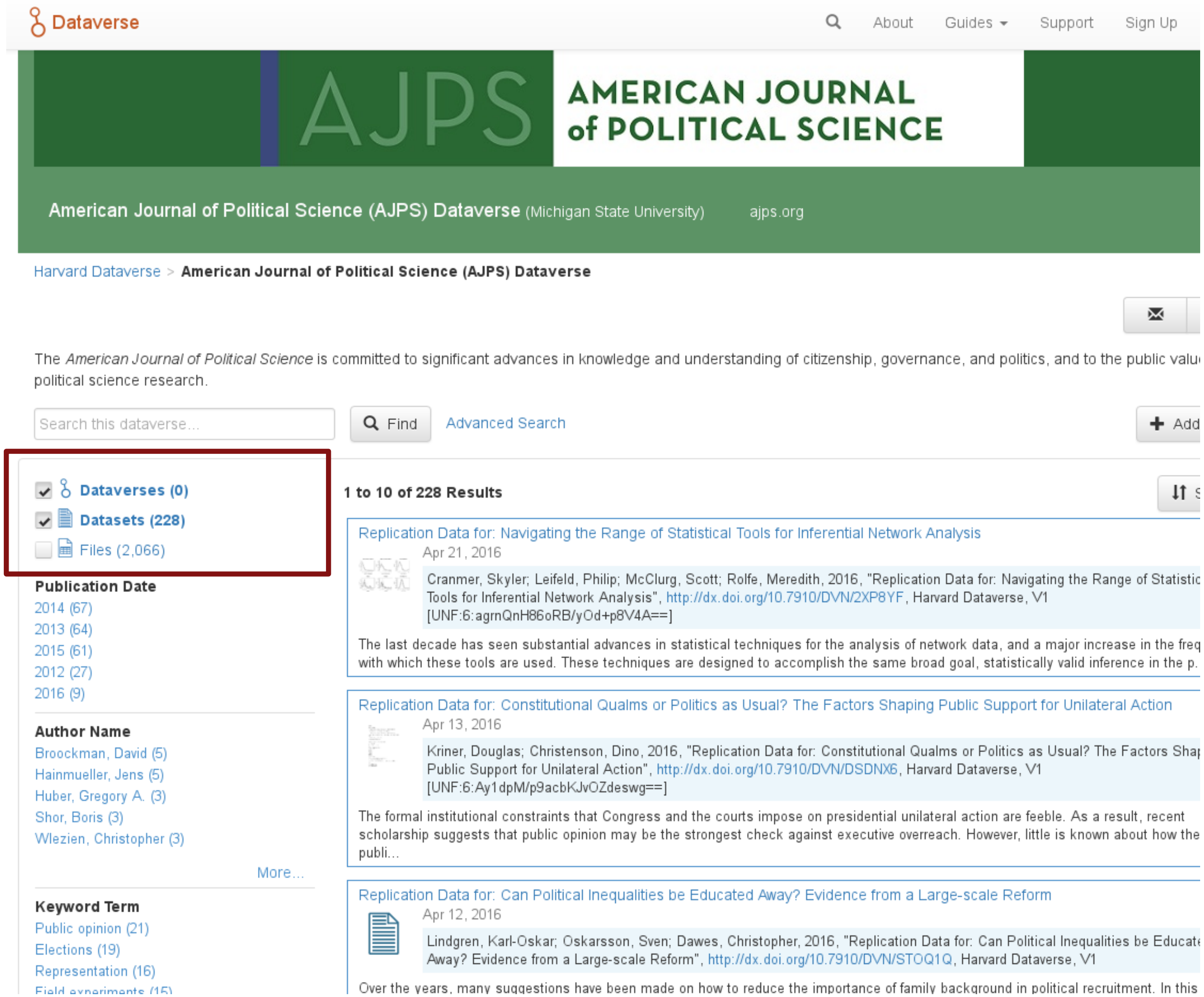
Full-text Article

Additional Materials

Download Data Set (1.99 GB) | Online Appendix (148.69 KB) | Author Disclosure Statement (10.33 KB)

Authors

Journals are starting to use them



Dataverse

About Guides Support Sign Up

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

American Journal of Political Science (AJPS) Dataverse (Michigan State University) ajps.org

Harvard Dataverse > **American Journal of Political Science (AJPS) Dataverse**

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

Search this dataverse... Find Advanced Search + Add

☑ **Dataverses (0)**
☑ **Datasets (228)**
□ Files (2,066)

Publication Date
2014 (67)
2013 (64)
2015 (61)
2012 (27)
2016 (9)

Author Name
Broockman, David (5)
Hainmueller, Jens (5)
Huber, Gregory A. (3)
Shor, Boris (3)
Wlezien, Christopher (3)
More...

Keyword Term
Public opinion (21)
Elections (19)
Representation (16)
Field experiments (15)

1 to 10 of 228 Results

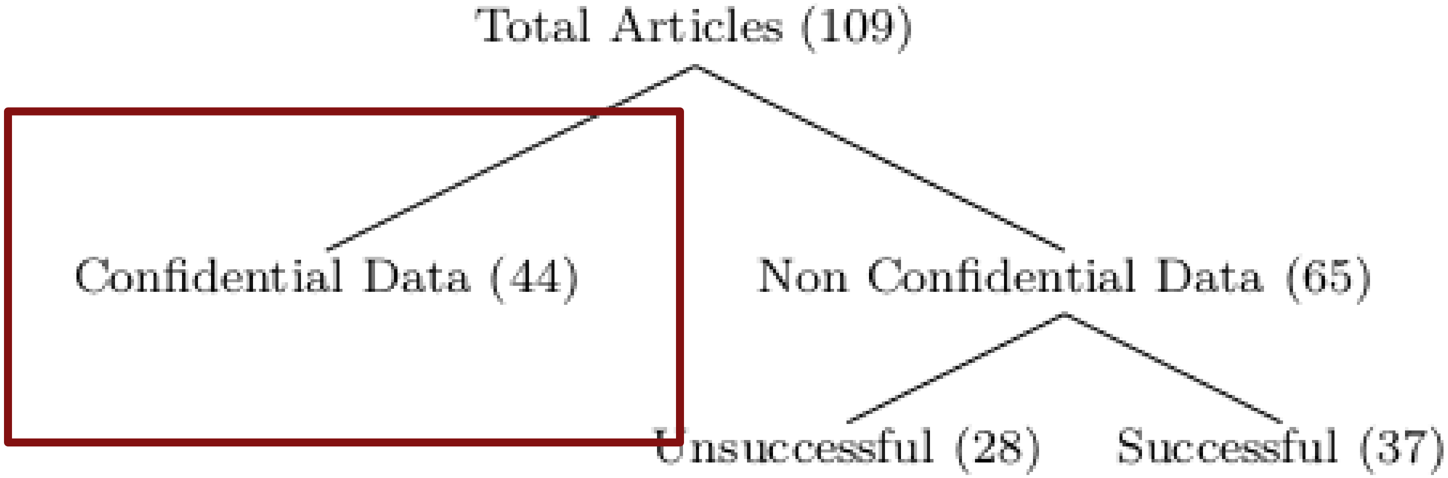
Replication Data for: Navigating the Range of Statistical Tools for Inferential Network Analysis
Apr 21, 2016
Cranmer, Skyler; Leifeld, Philip; McClurg, Scott; Rolfe, Meredith, 2016, "Replication Data for: Navigating the Range of Statistic
Tools for Inferential Network Analysis", http://dx.doi.org/10.7910/DVN/2XP8YF, Harvard Dataverse, V1
[UNF:6:agrnQnH86oRB/yOd+p8V4A==]
The last decade has seen substantial advances in statistical techniques for the analysis of network data, and a major increase in the freq
with which these tools are used. These techniques are designed to accomplish the same broad goal, statistically valid inference in the p.

Replication Data for: Constitutional Qualms or Politics as Usual? The Factors Shaping Public Support for Unilateral Action
Apr 13, 2016
Kriner, Douglas; Christenson, Dino, 2016, "Replication Data for: Constitutional Qualms or Politics as Usual? The Factors Shap
Public Support for Unilateral Action", http://dx.doi.org/10.7910/DVN/DSDNX6, Harvard Dataverse, V1
[UNF:6:Ay1dpM/p9acbKJvOZdeswg==]
The formal institutional constraints that Congress and the courts impose on presidential unilateral action are feeble. As a result, recent
scholarship suggests that public opinion may be the strongest check against executive overreach. However, little is known about how the
publi...

Replication Data for: Can Political Inequalities be Educated Away? Evidence from a Large-scale Reform
Apr 12, 2016
Lindgren, Karl-Oskar; Oskarsson, Sven; Dawes, Christopher, 2016, "Replication Data for: Can Political Inequalities be Educate
Away? Evidence from a Large-scale Reform", http://dx.doi.org/10.7910/DVN/STOQ1Q, 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

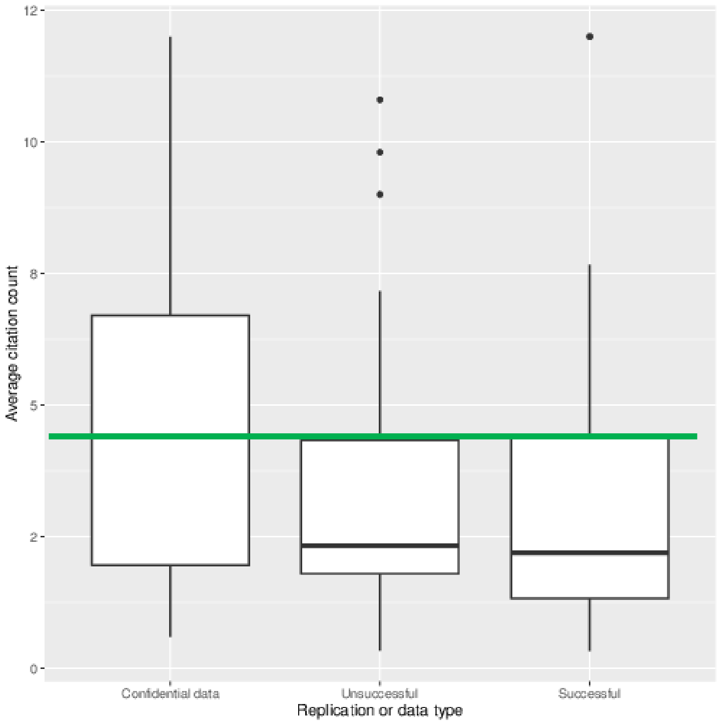
But the biggest problem…

Figure 1: A Breakdown of the Articles



Back to confidential data

- Articles using confidential data are (weakly) more cited than others

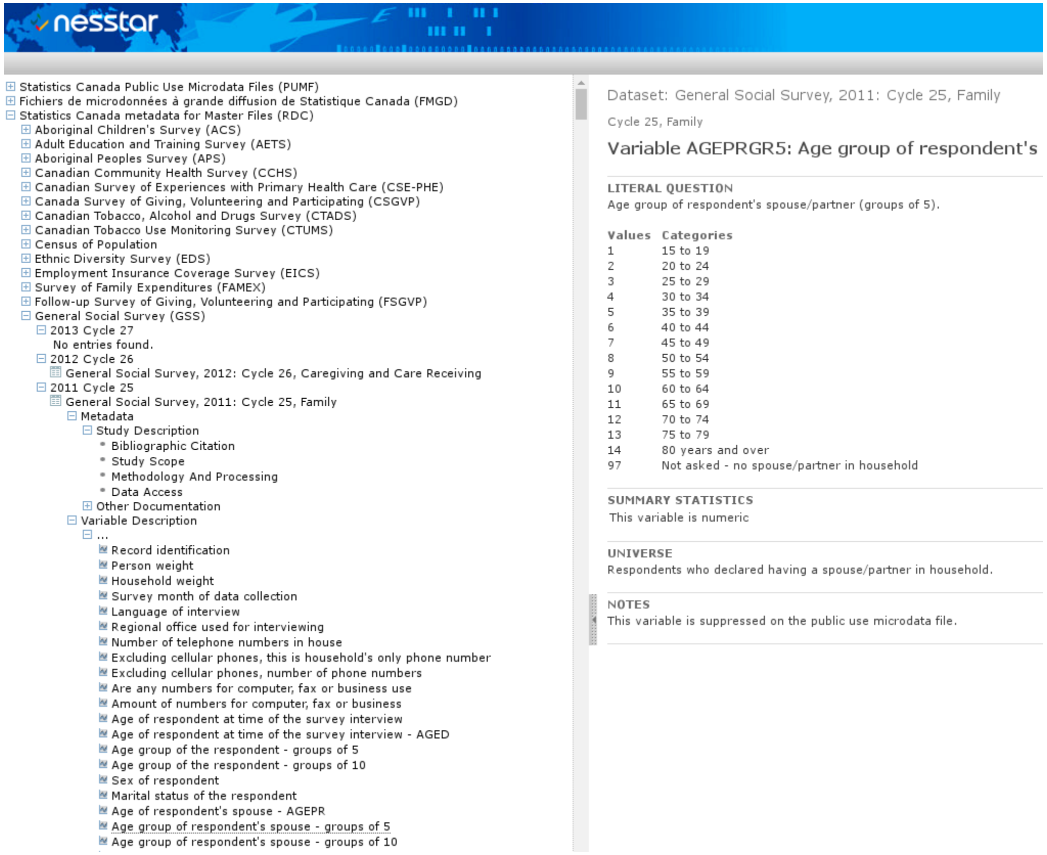


But: for confidential data…

- Data is not available
- Metadata is not available
- Programs? So-so…

Should We Just Trust These Guys?

Some are quite commendable



- ⊞ Statistics Canada Public Use Microdata Files (PUMF)
- ⊞ Fichiers de microdonnées à grande diffusion de Statistique Canada (FMGD)
- ⊟ Statistics Canada metadata for Master Files (RDC)
  - ⊞ Aboriginal Children's Survey (ACS)
  - ⊞ Adult Education and Training Survey (AETS)
  - ⊞ Aboriginal Peoples Survey (APS)
  - ⊞ Canadian Community Health Survey (CCHS)
  - ⊞ Canadian Survey of Experiences with Primary Health Care (CSE-PHE)
  - ⊞ Canada Survey of Giving, Volunteering and Participating (CSGVP)
  - ⊞ Canadian Tobacco, Alcohol and Drugs Survey (CTADS)
  - ⊞ Canadian Tobacco Use Monitoring Survey (CTUMS)
  - ⊞ Census of Population
  - ⊞ Ethnic Diversity Survey (EDS)
  - ⊞ Employment Insurance Coverage Survey (EICS)
  - ⊞ Survey of Family Expenditures (FAMEX)
  - ⊞ Follow-up Survey of Giving, Volunteering and Participating (FSGVP)
  - ⊟ General Social Survey (GSS)
    - ⊟ 2013 Cycle 27
      - No entries found.
    - ⊟ 2012 Cycle 26
      - General Social Survey, 2012: Cycle 26, Caregiving and Care Receiving
    - ⊟ 2011 Cycle 25
      - General Social Survey, 2011: Cycle 25, Family
        - ⊟ Metadata
          - ⊟ Study Description
            - Bibliographic Citation
            - Study Scope
            - Methodology And Processing
            - Data Access
          - ⊞ Other Documentation
        - ⊟ Variable Description
          - ⊟ ...
            - Record identification
            - Person weight
            - Household weight
            - Survey month of data collection
            - Language of interview
            - Regional office used for interviewing
            - Number of telephone numbers in house
            - Excluding cellular phones, this is household's only phone number
            - Excluding cellular phones, number of phone numbers
            - Are any numbers for computer, fax or business use
            - Amount of numbers for computer, fax or business
            - Age of respondent at time of the survey interview
            - Age of respondent at time of the survey interview - AGED
            - Age group of the respondent - groups of 5
            - Age group of the respondent - groups of 10
            - Sex of respondent
            - Marital status of the respondent
            - Age of respondent's spouse - AGEPR
            - Age group of respondent's spouse - groups of 5
            - Age group of respondent's spouse - groups of 10

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

Cycle 25, Family

Variable AGEPRGR5: Age group of respondent's

LITERAL QUESTION

Age group of respondent's spouse/partner (groups of 5).

Values	Categories
1	15 to 19
2	20 to 24
3	25 to 29
4	30 to 34
5	35 to 39
6	40 to 44
7	45 to 49
8	50 to 54
9	55 to 59
10	60 to 64
11	65 to 69
12	70 to 74
13	75 to 79
14	80 years and over
97	Not asked - no spouse/partner in household

SUMMARY STATISTICS

This variable is numeric

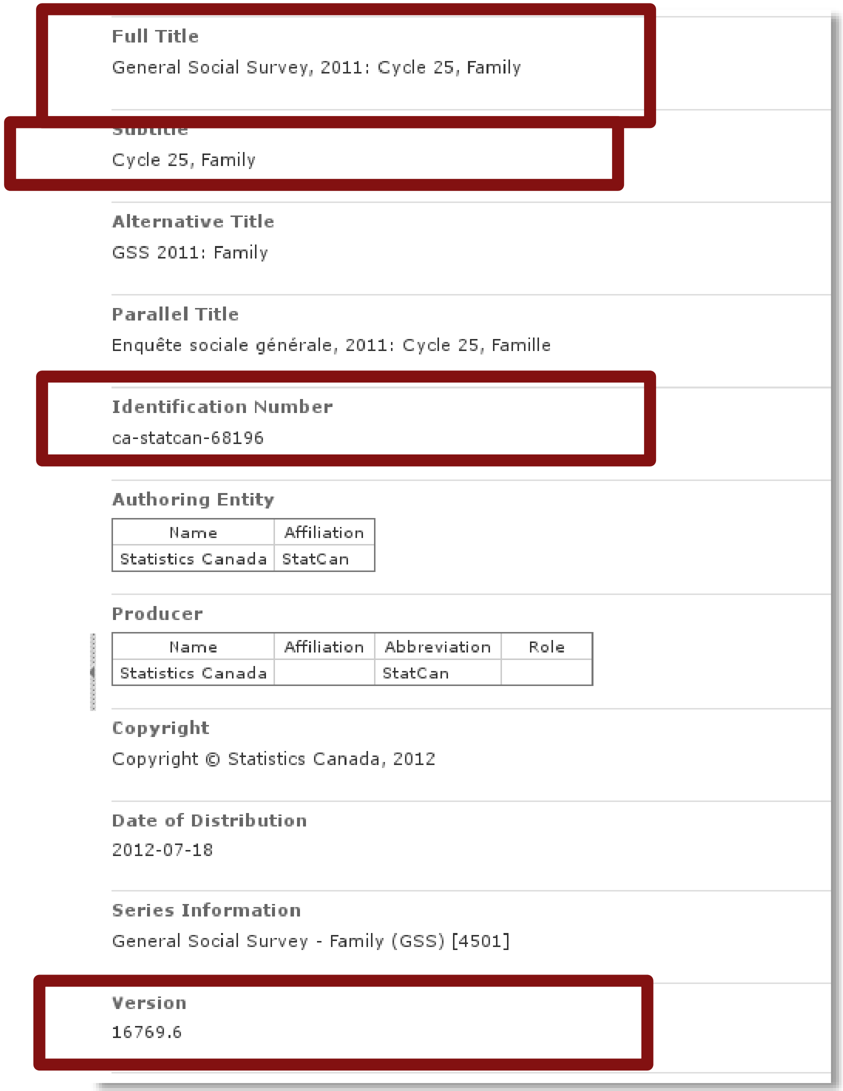
UNIVERSE

Respondents who declared having a spouse/partner in household.

NOTES

This variable is suppressed on the public use microdata file.

Even detailed information



Full Title

General Social Survey, 2011: Cycle 25, Family

Subtitle

Cycle 25, Family

Alternative Title

GSS 2011: Family

Parallel Title

Enquête sociale générale, 2011: Cycle 25, Famille

Identification Number

ca-statcan-68196

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]

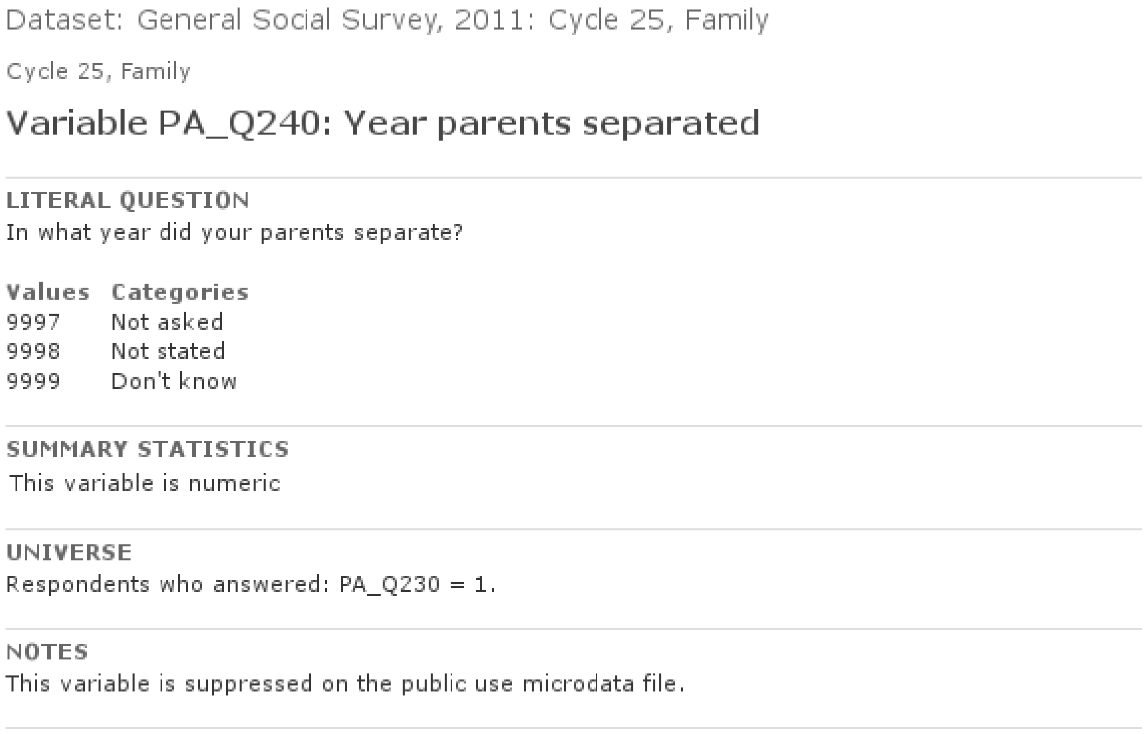
Version

16769.6

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

LITERAL QUESTION

In what year did your parents separate?

Values	Categories
9997	Not asked
9998	Not stated
9999	Don't know

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!

text
BREAKING NEWS
Free the data users GROWS LARGER
Major rally & march today in New York
nark imposes tax on foods high in saturated fats
S&P ▲ 20
CN

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?

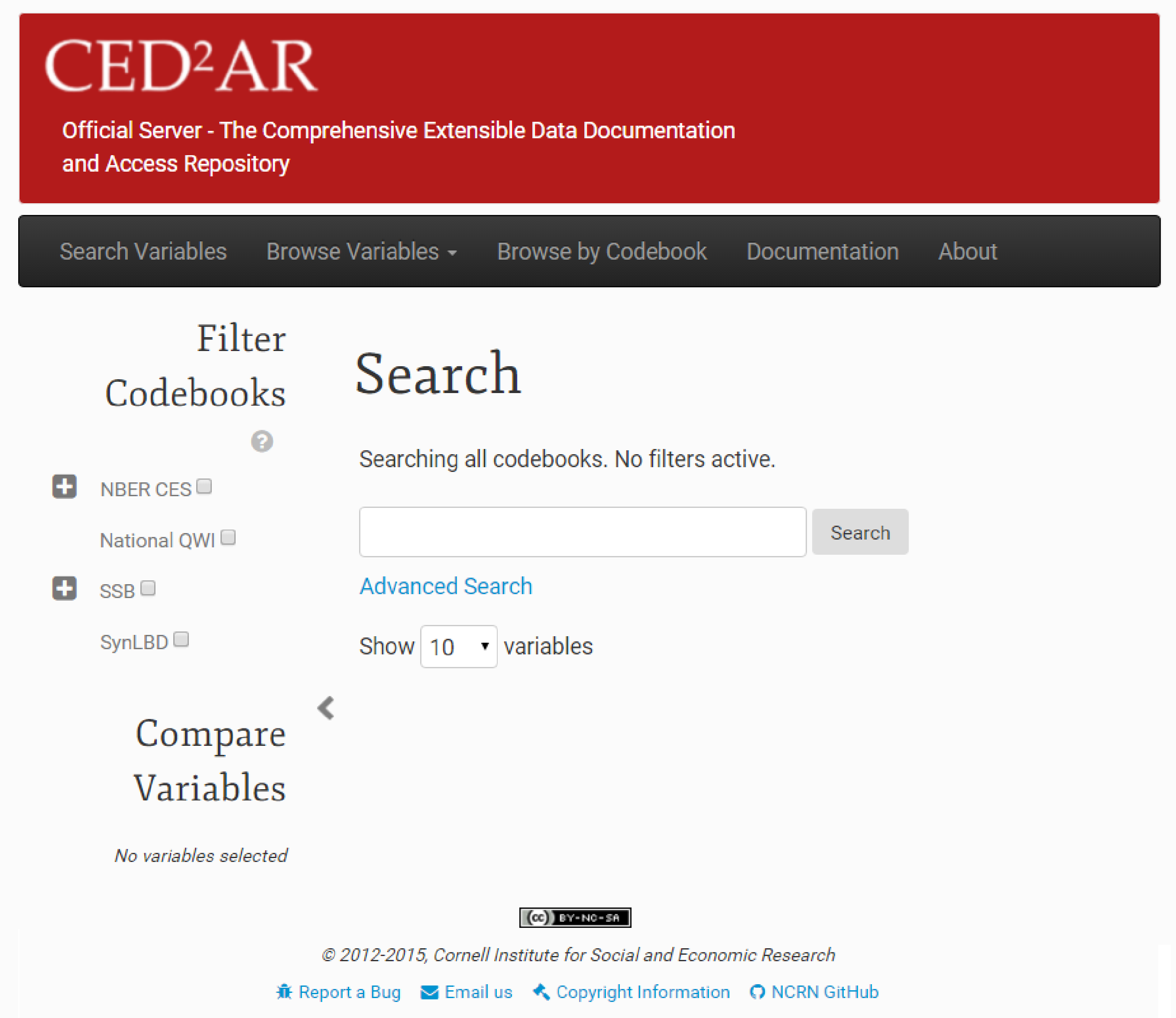
CED^{2}AR

The Comprehensive Extensible Data Documentation and Access Repository

What is 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?



CED²AR

Official Server - The Comprehensive Extensible Data Documentation and Access Repository

Filter Codebooks

- NBER CES □
- National QWI □
- SSB □
- SynLBD □

Search

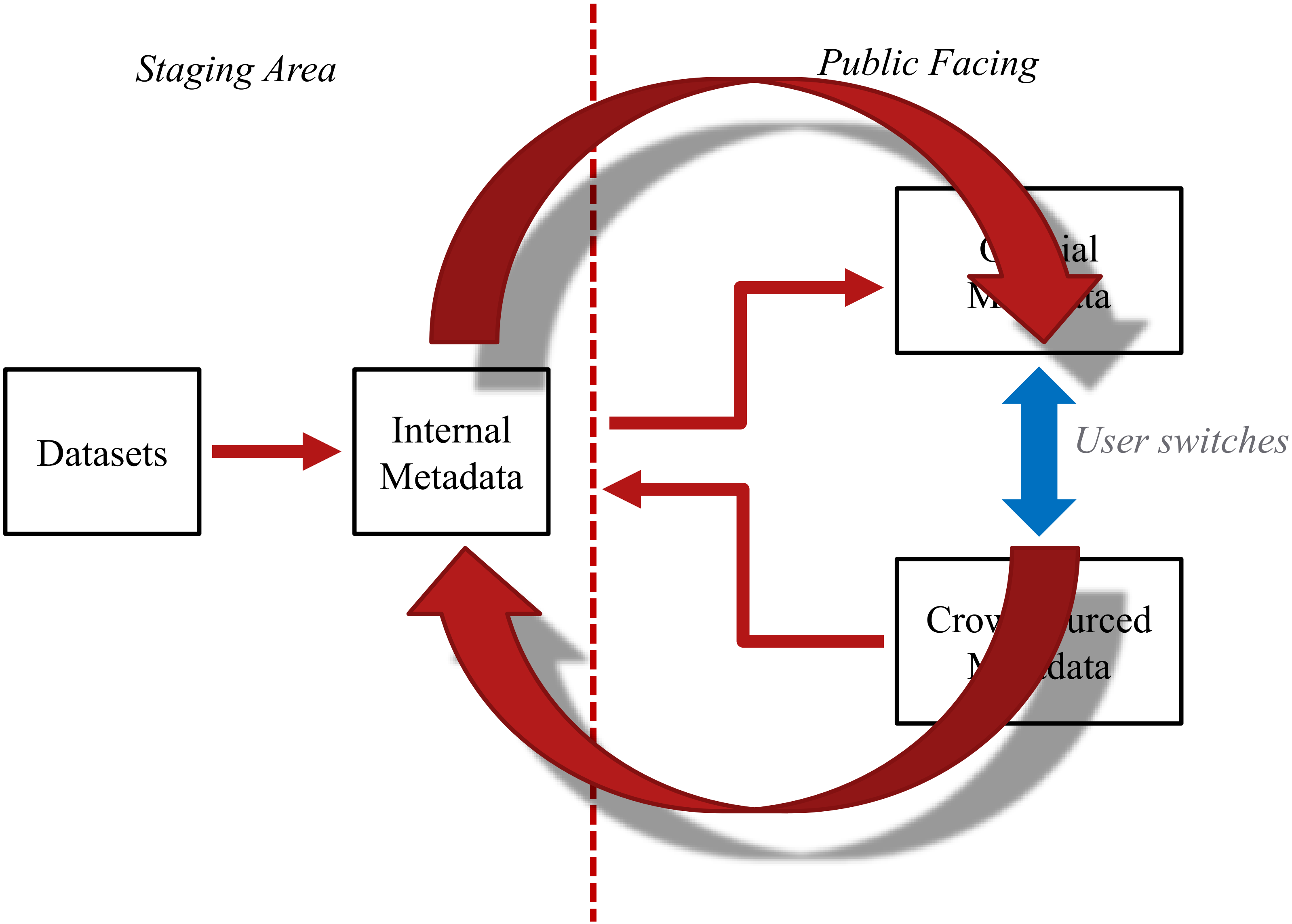
Searching all codebooks. No filters active.

Show 10 variables

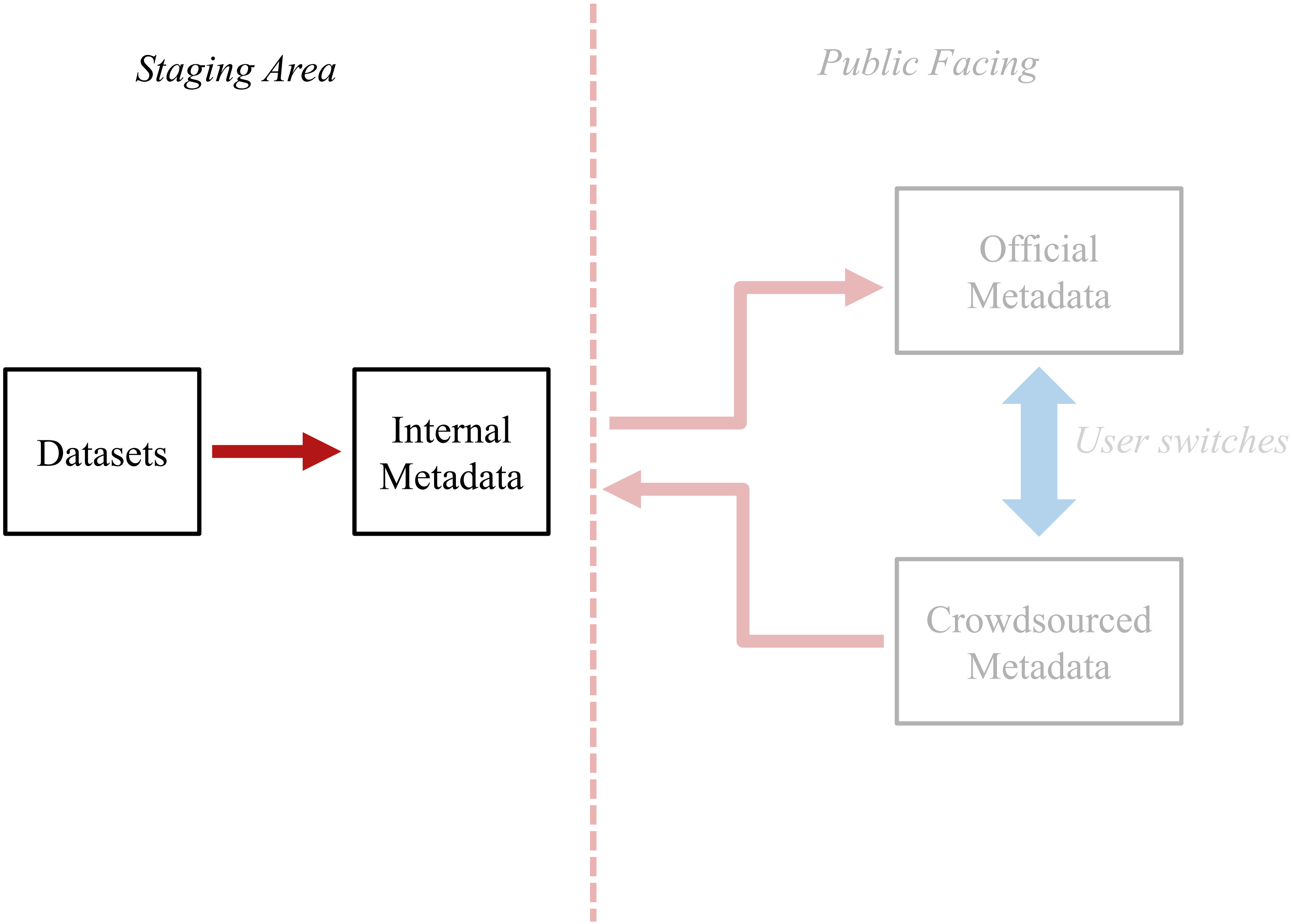
Compare Variables

No variables selected

Basic Information Flow



Basic Information Flow



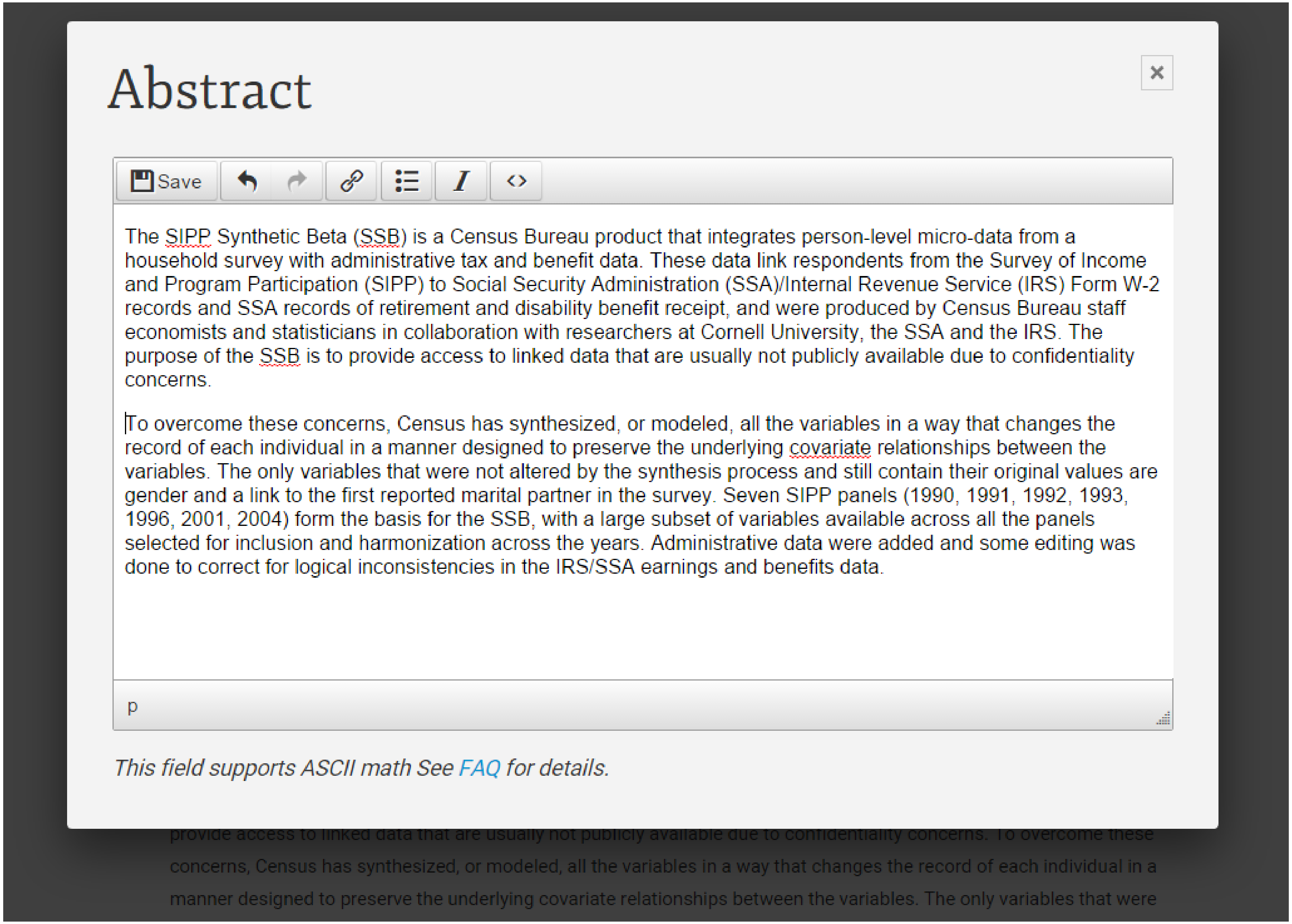
Internal Processing

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
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.

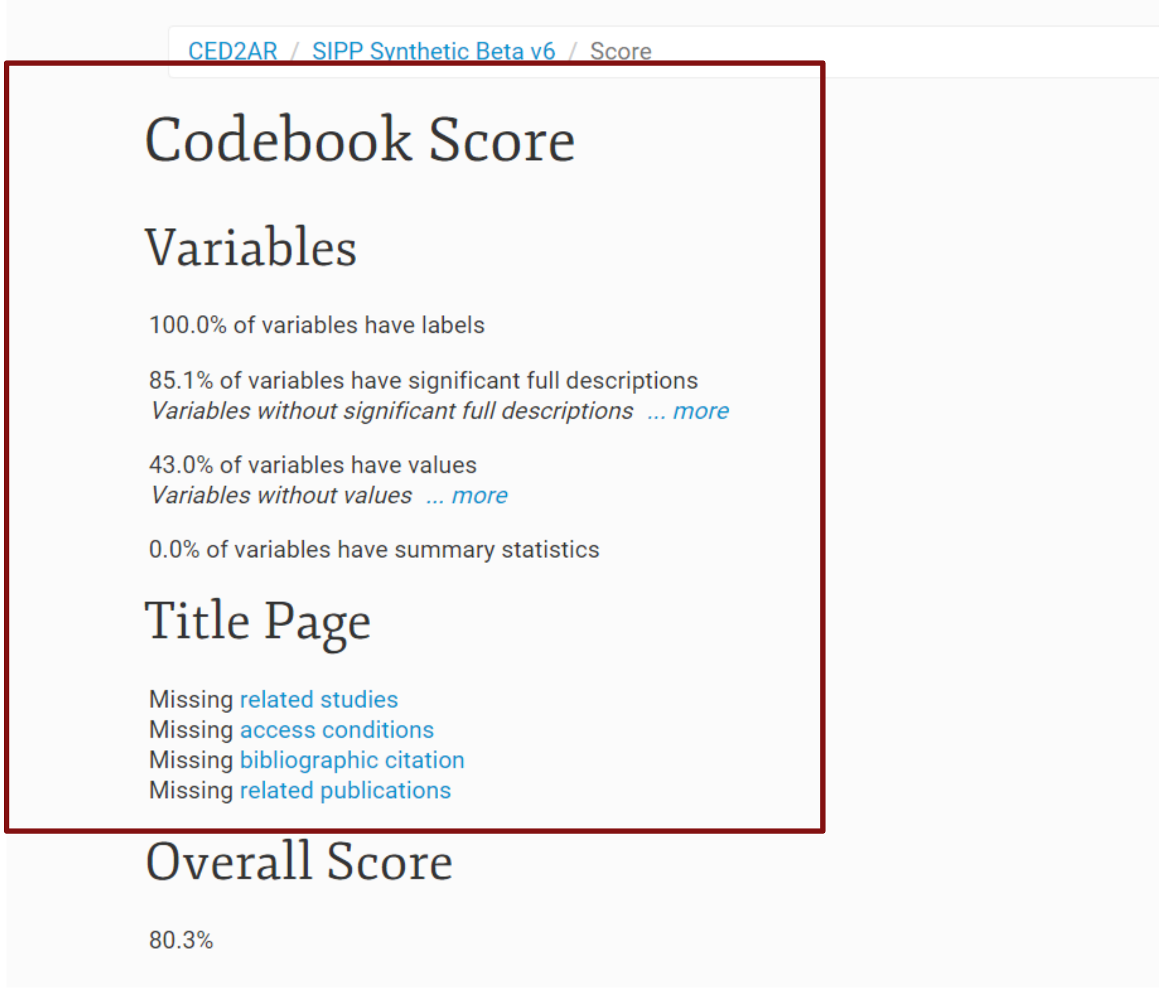
To overcome these concerns, Census has synthesized, or modeled, all the variables in a way that changes the record of each individual in a manner designed to preserve the underlying covariate relationships between the variables. The only variables that were not altered by the synthesis process and still 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.

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: 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

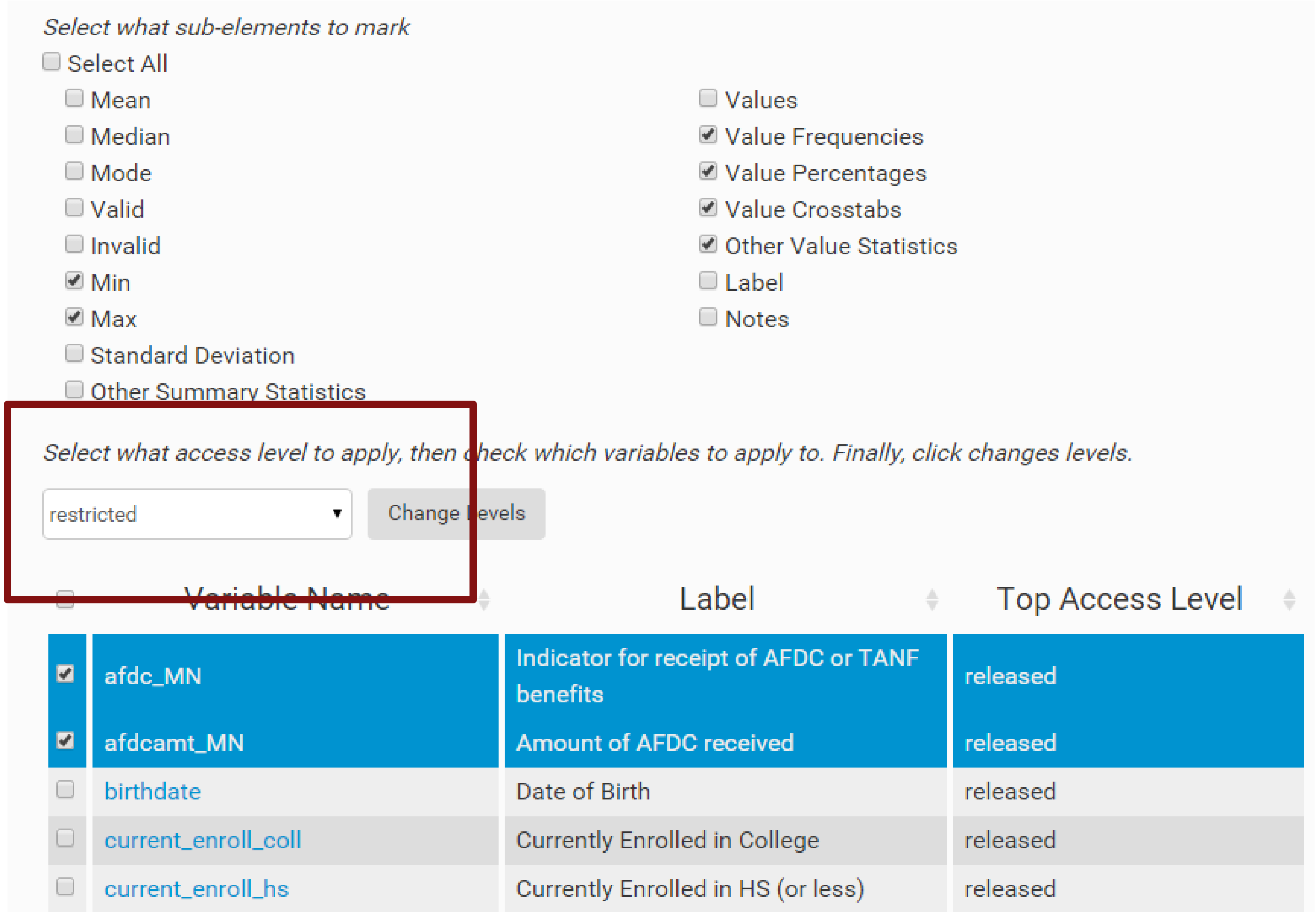
80.3%

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

□ Values
☑ Value Frequencies
☑ Value Percentages
☑ Value Crosstabs
☑ Other Value Statistics
□ Label
□ Notes

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

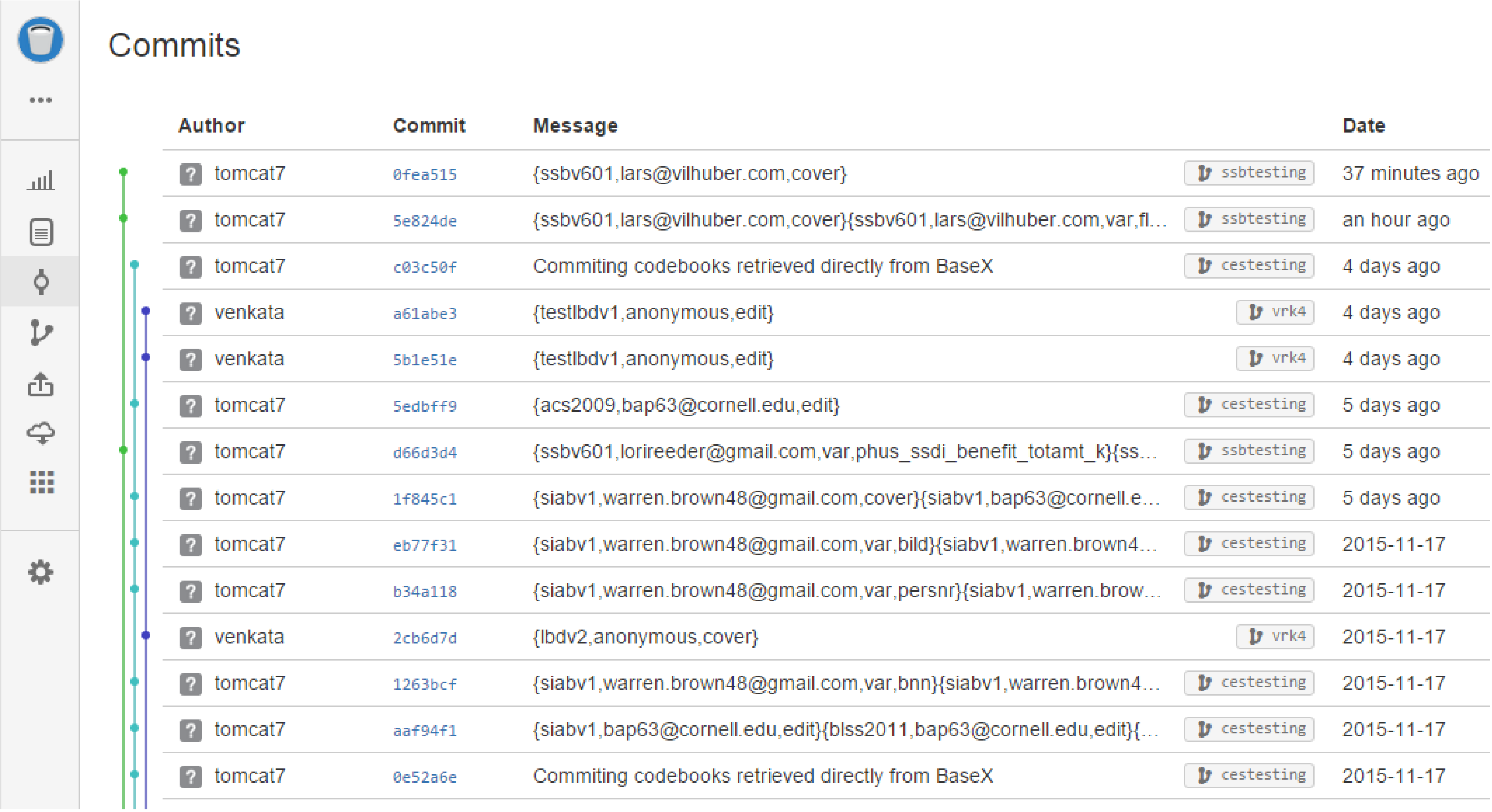
□	Variable Name	Label	Top Access Level
☑	afdc_MN	Indicator for receipt of AFDC or TANF benefits	released
☑	afdcamt_MN	Amount of AFDC received	released
□	birthdate	Date of Birth	released
□	current_enroll_coll	Currently Enrolled in College	released
□	current_enroll_hs	Currently Enrolled in HS (or less)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

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

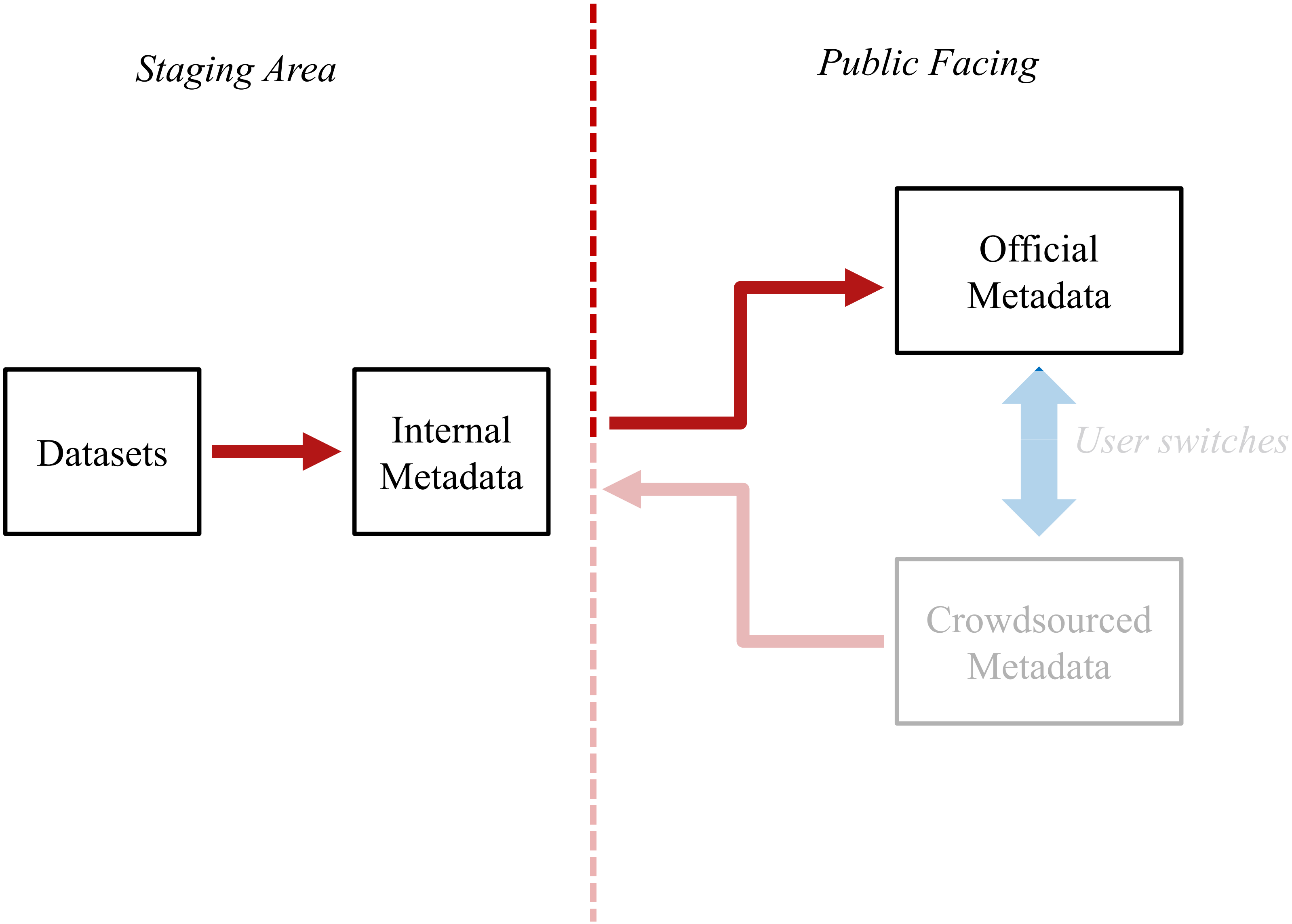
All changes are logged externally via Git



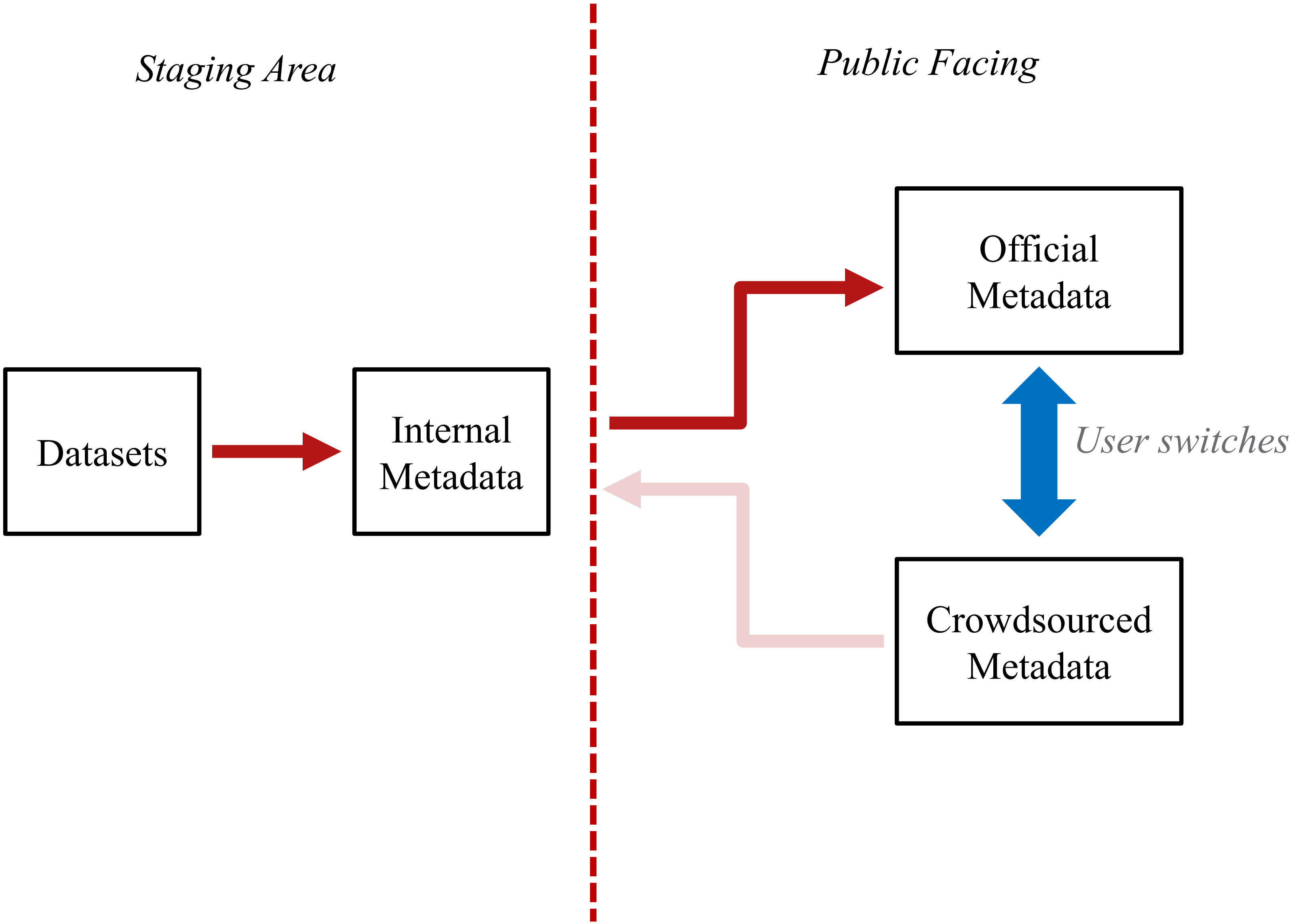
Commits

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



Basic Information Flow



Staging Area

Public Facing

Datasets

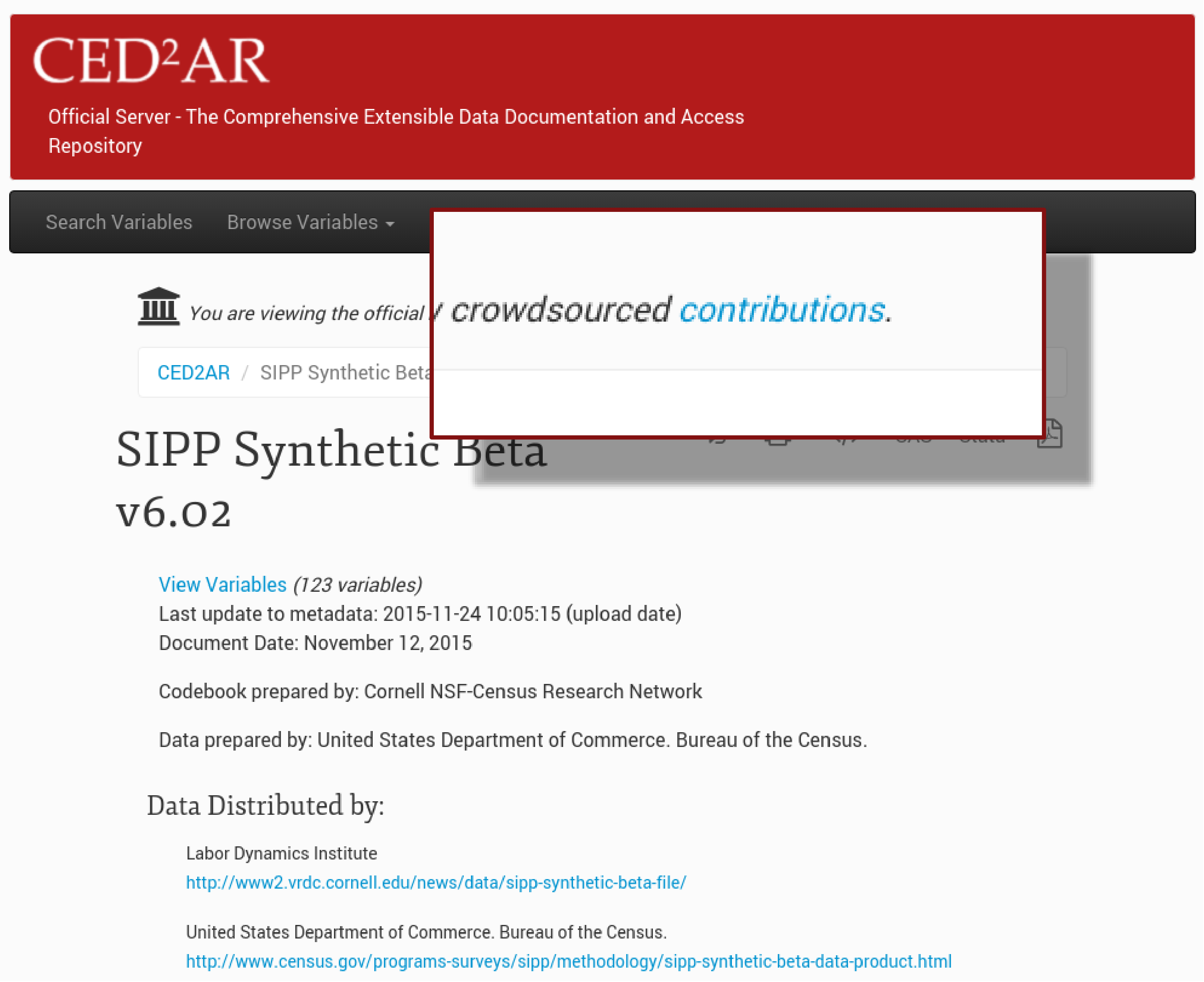
Internal Metadata

Official Metadata

User switches

Crowdsourced Metadata

Official view



CED²AR

Official Server - The Comprehensive Extensible Data Documentation and Access Repository

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

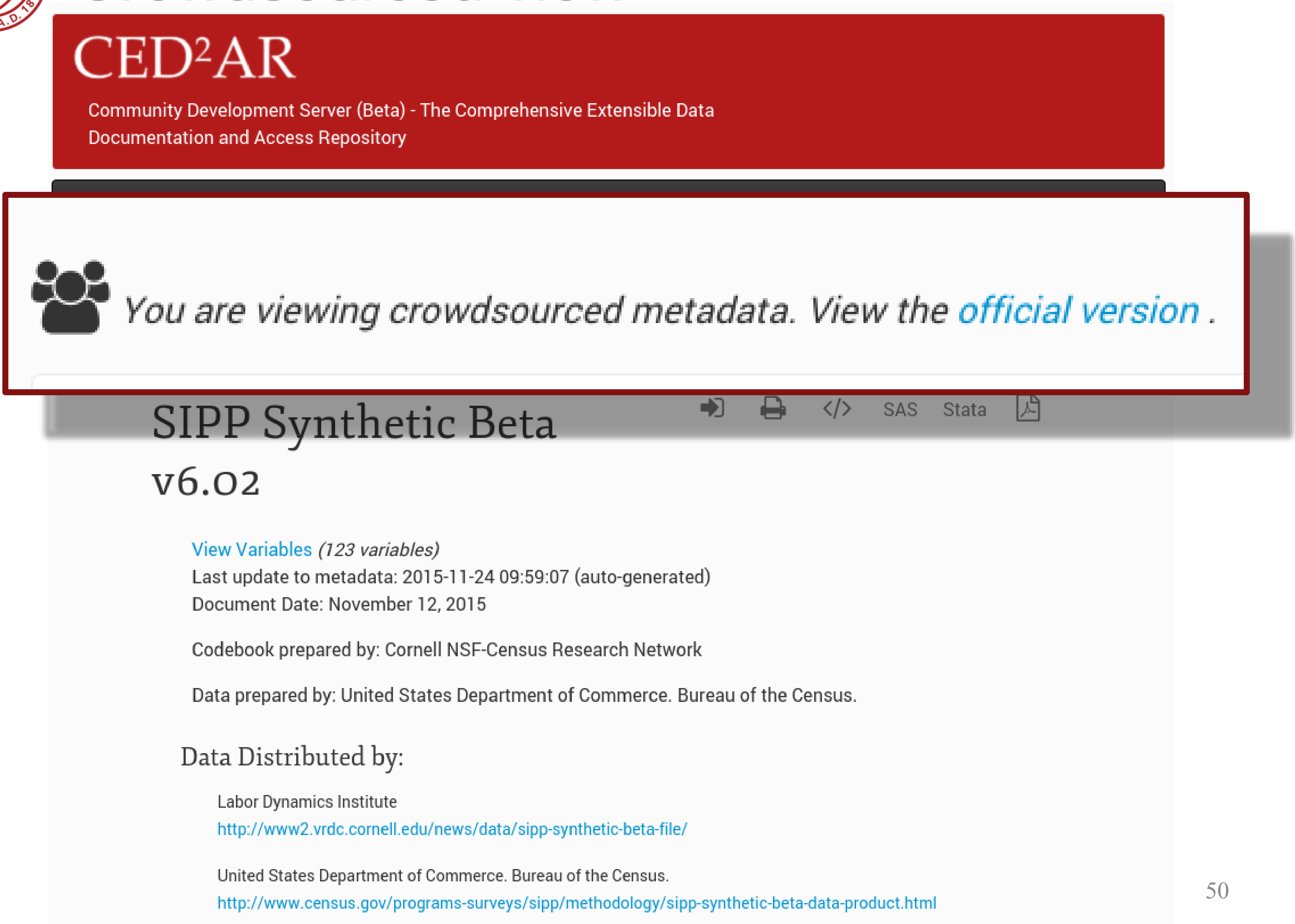
Data prepared by: United States Department of Commerce. Bureau of the Census.

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 .

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.

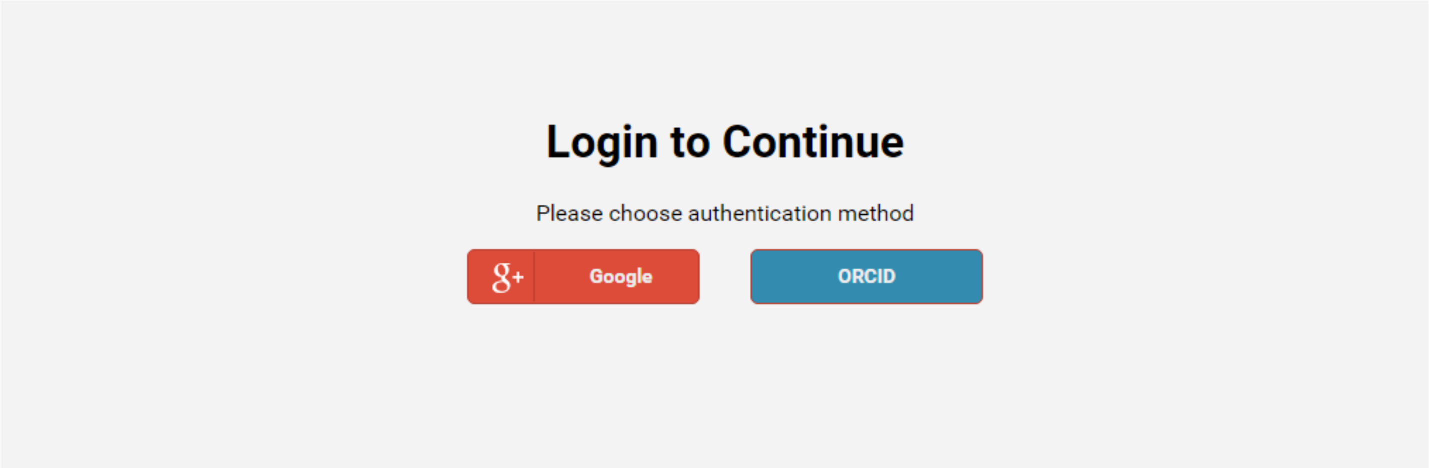
http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html

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

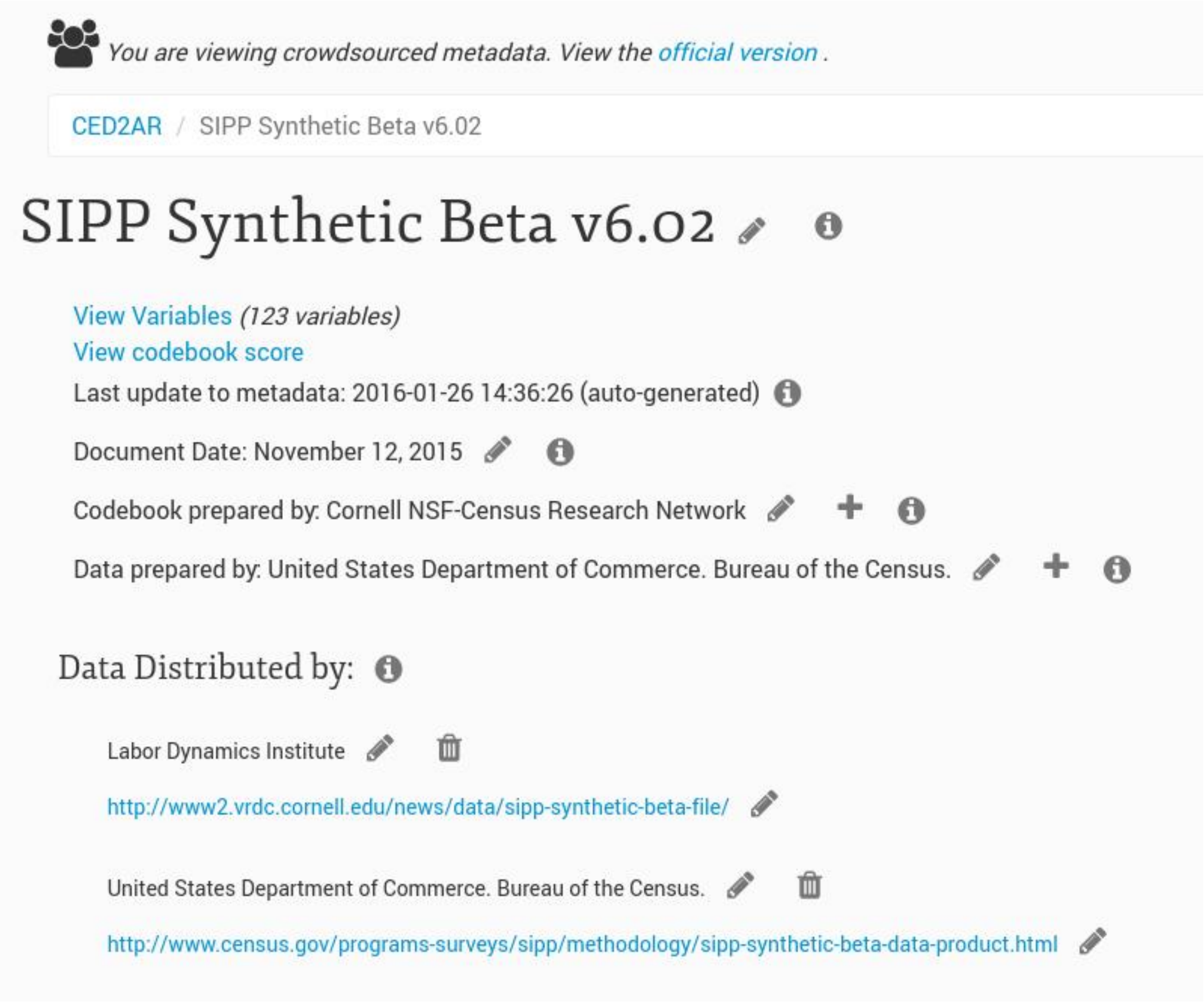
- Supports OpenID and OAuth2
  - Currently using Google and ORCID with OAuth2
  - Developing connectors to work with additional providers
- CED^{2}AR handles identity management



Login to Continue

Please choose authentication method

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 14:36:26 (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.

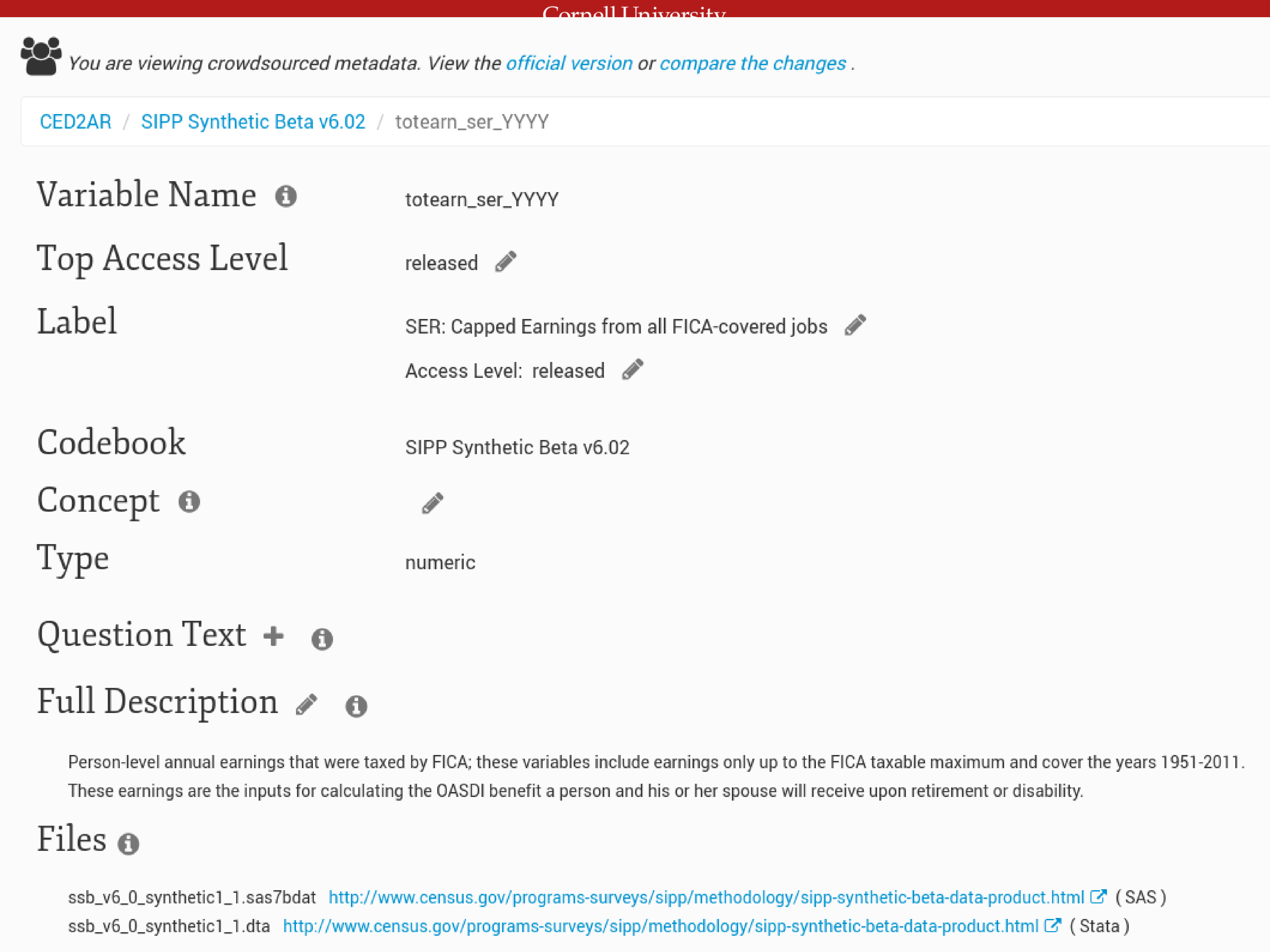
Data Distributed by:

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



You are viewing crowdsourced metadata. View the official version or compare the changes .

CED2AR / SIPP Synthetic Beta v6.02 / totearn_ser_YYYY

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

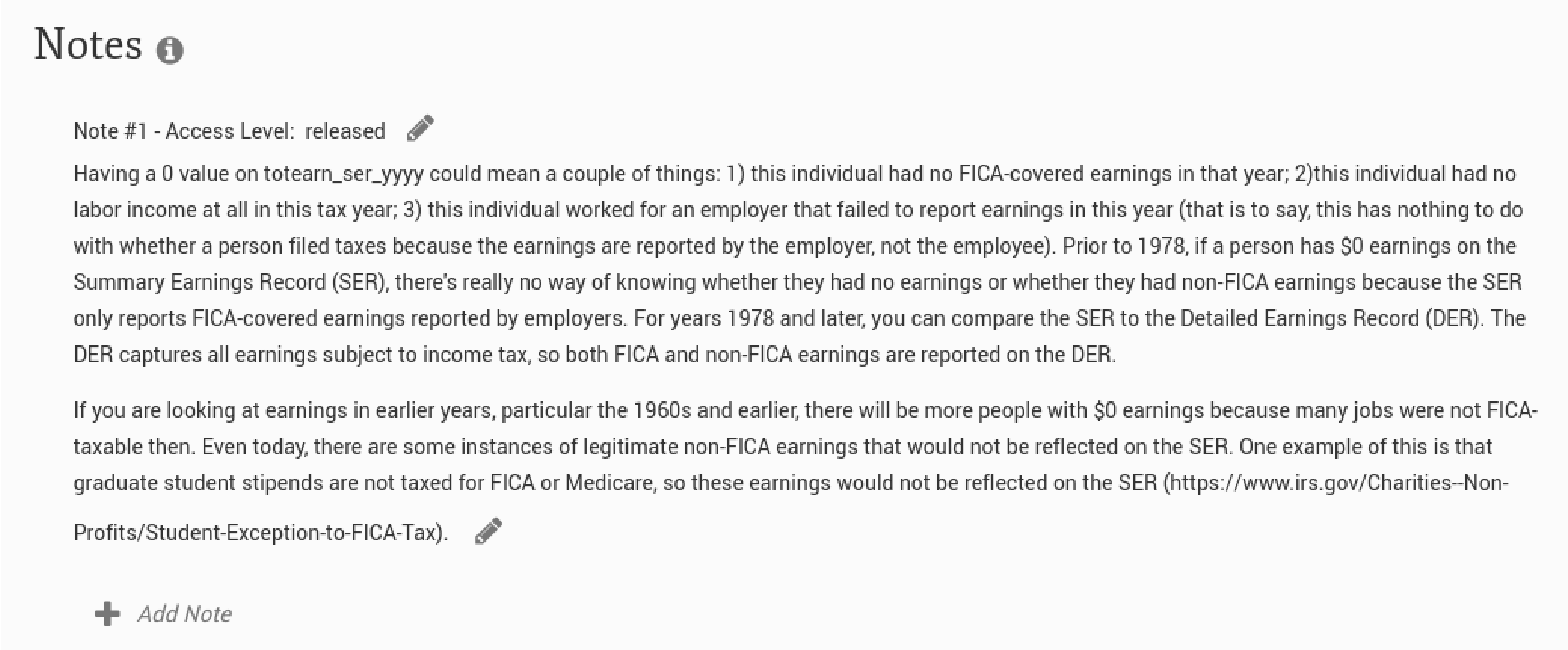
Question Text +

Full Description

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.

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)

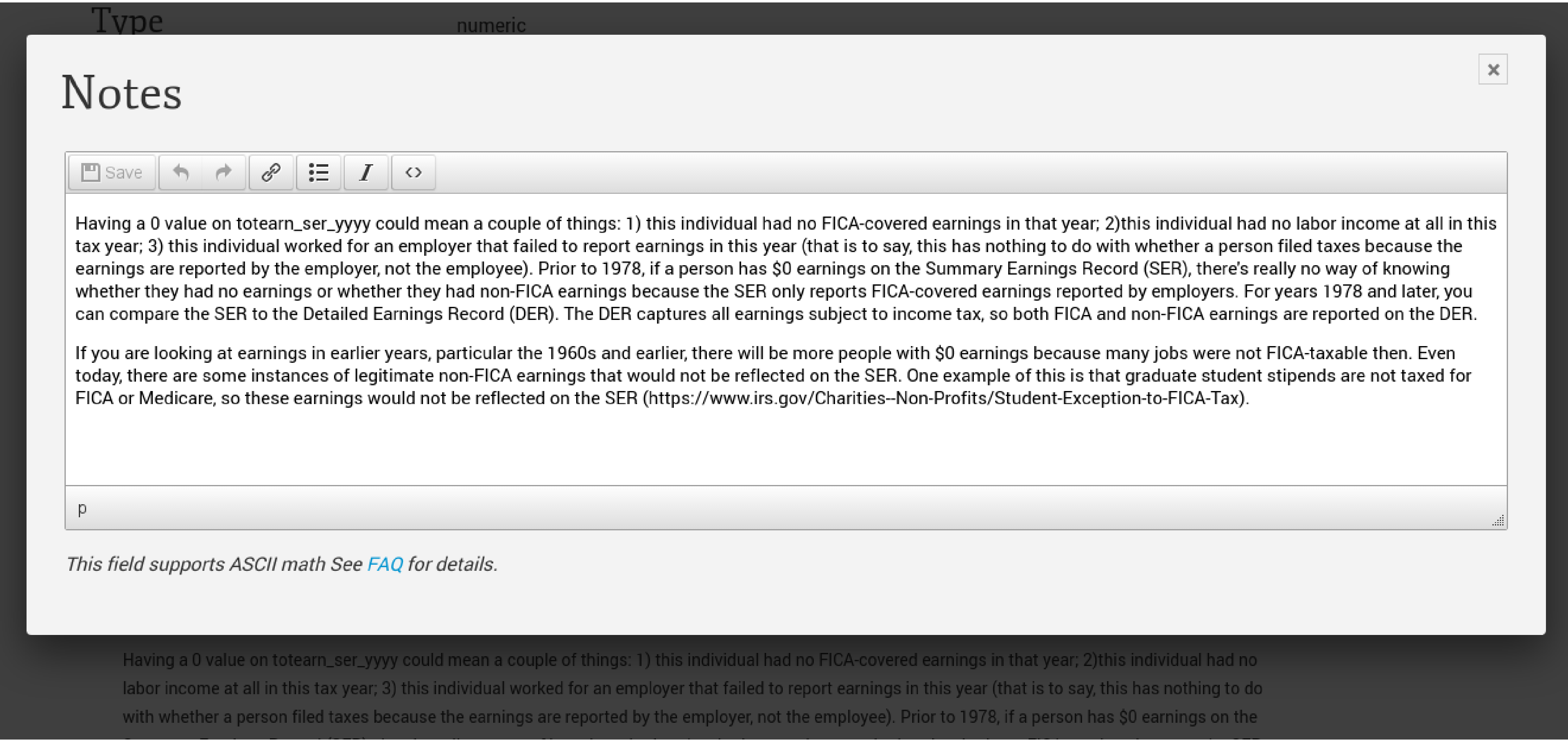


Notes

Note #1 - Access Level: released

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).



Type numeric

Notes

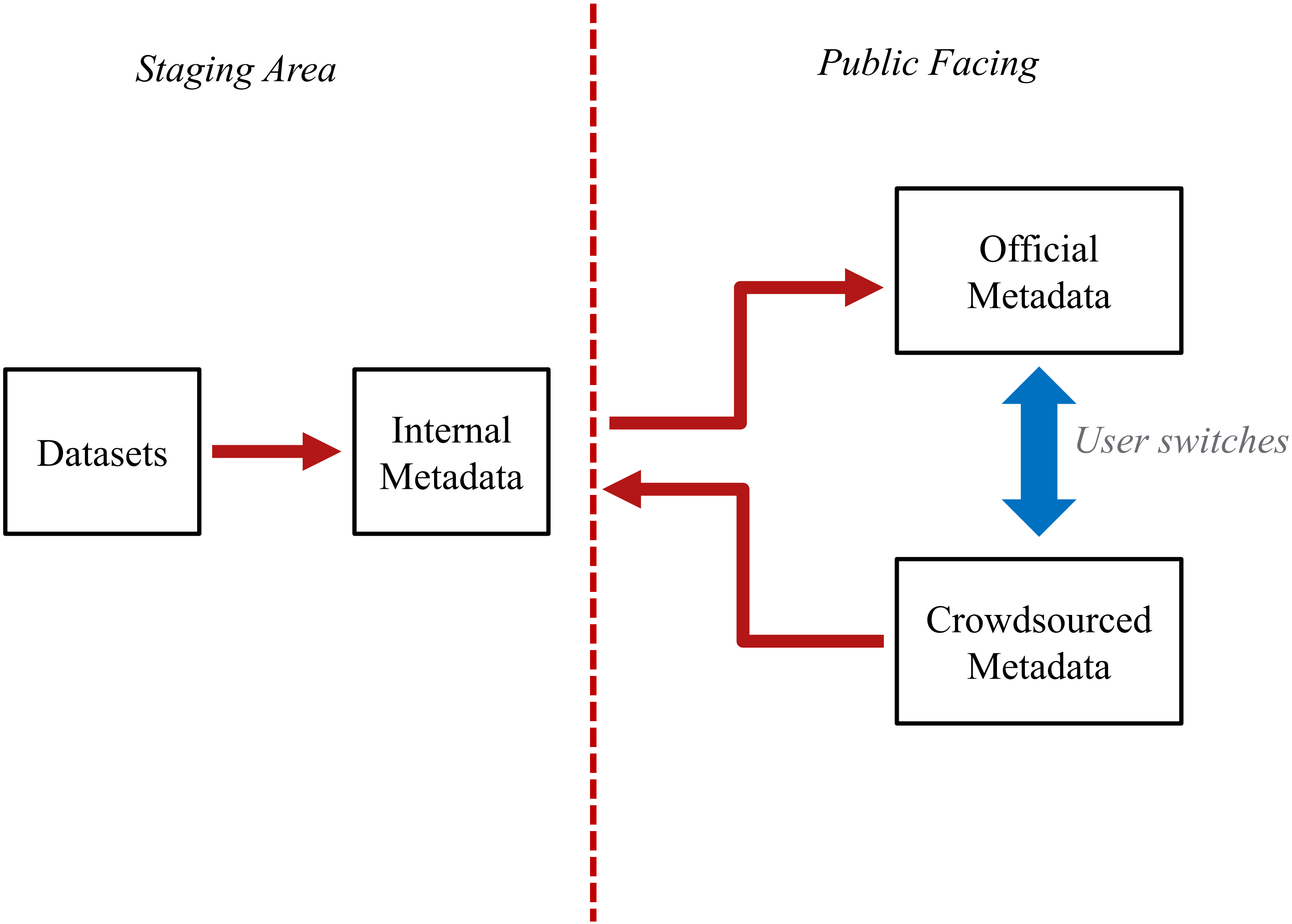
Having a 0 value on totearn_ser_yyyy could mean a couple of things: 1) this individual had no FICA-covered earnings in that year; 2)this individual had no labor income at all in this tax year; 3) this individual worked for an employer that failed to report earnings in this year (that is to say, this 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).

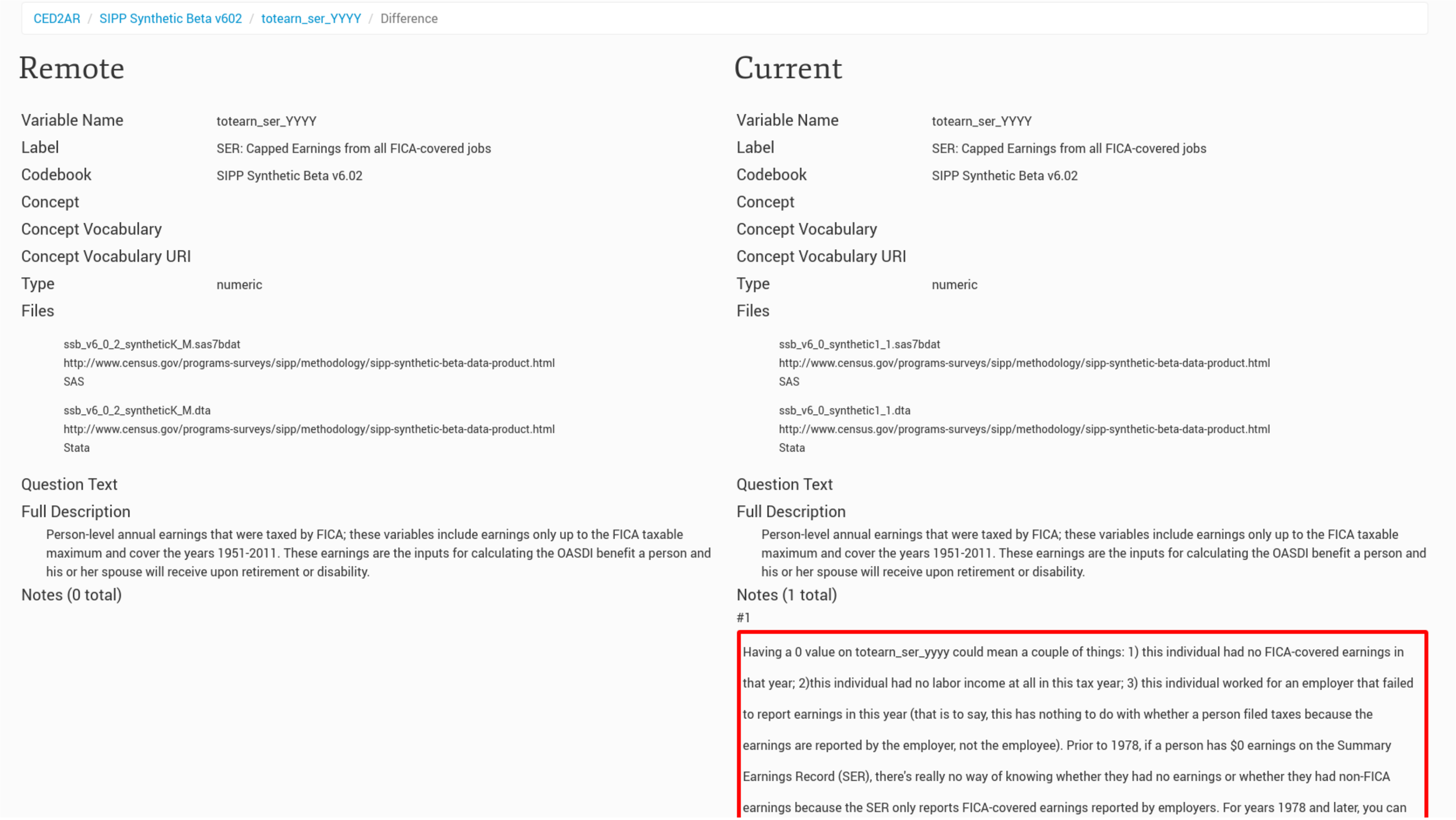
This field supports ASCII math See FAQ for details.

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

Basic Information Flow



Everybody can see changes



CED2AR / SIPP Synthetic Beta v602 / totearn_ser_YYYY / Difference

Remote

Variable Name	totearn_ser_YYYY
Label	SER: Capped Earnings from all FICA-covered jobs
Codebook	SIPP Synthetic Beta v6.02
Concept	
Concept Vocabulary	
Concept Vocabulary URI	
Type	numeric
Files	

ssb_v6_0_2_syntheticK_M.sas7bdat
http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html
SAS

ssb_v6_0_2_syntheticK_M.dta
http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html
Stata

Question Text

Full Description

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

Variable Name	totearn_ser_YYYY
Label	SER: Capped Earnings from all FICA-covered jobs
Codebook	SIPP Synthetic Beta v6.02
Concept	
Concept Vocabulary	
Concept Vocabulary URI	
Type	numeric
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

Question Text

Full Description

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.

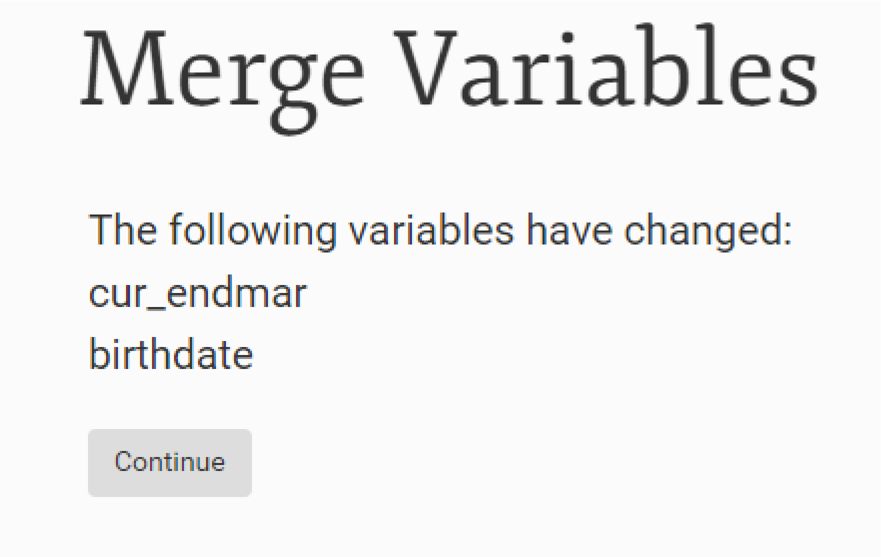
Notes (1 total)

#1

Having a 0 value on 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



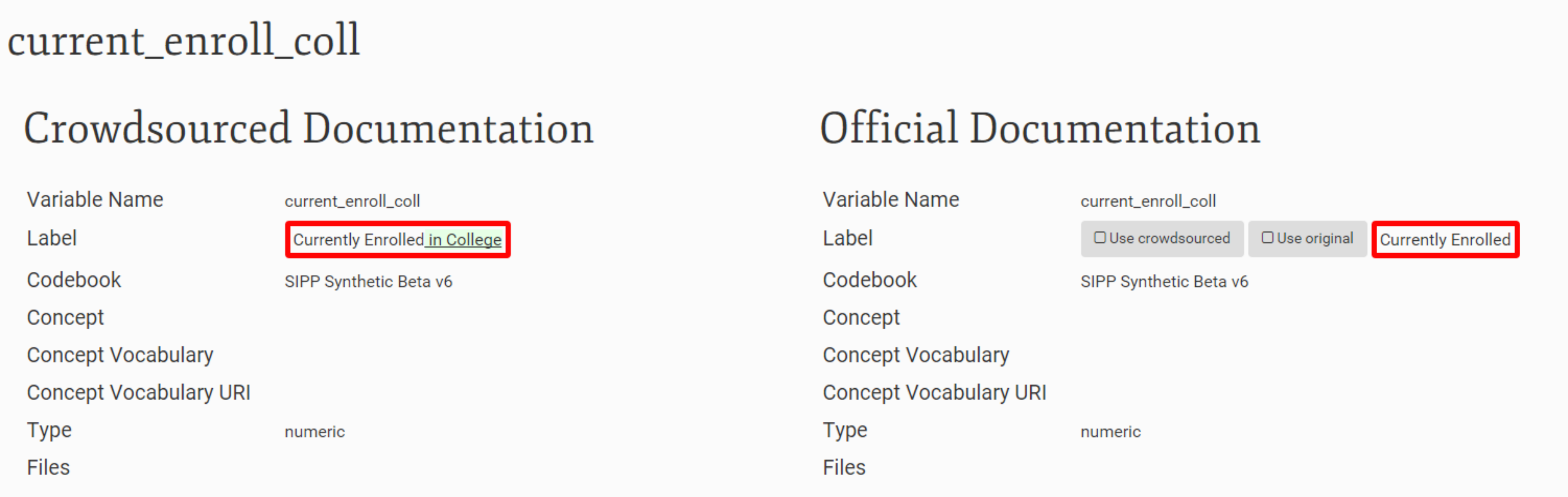
Merge Variables

The following variables have changed:

cur_endmar

birthdate

Combining Knowledge: Merging



current_enroll_coll

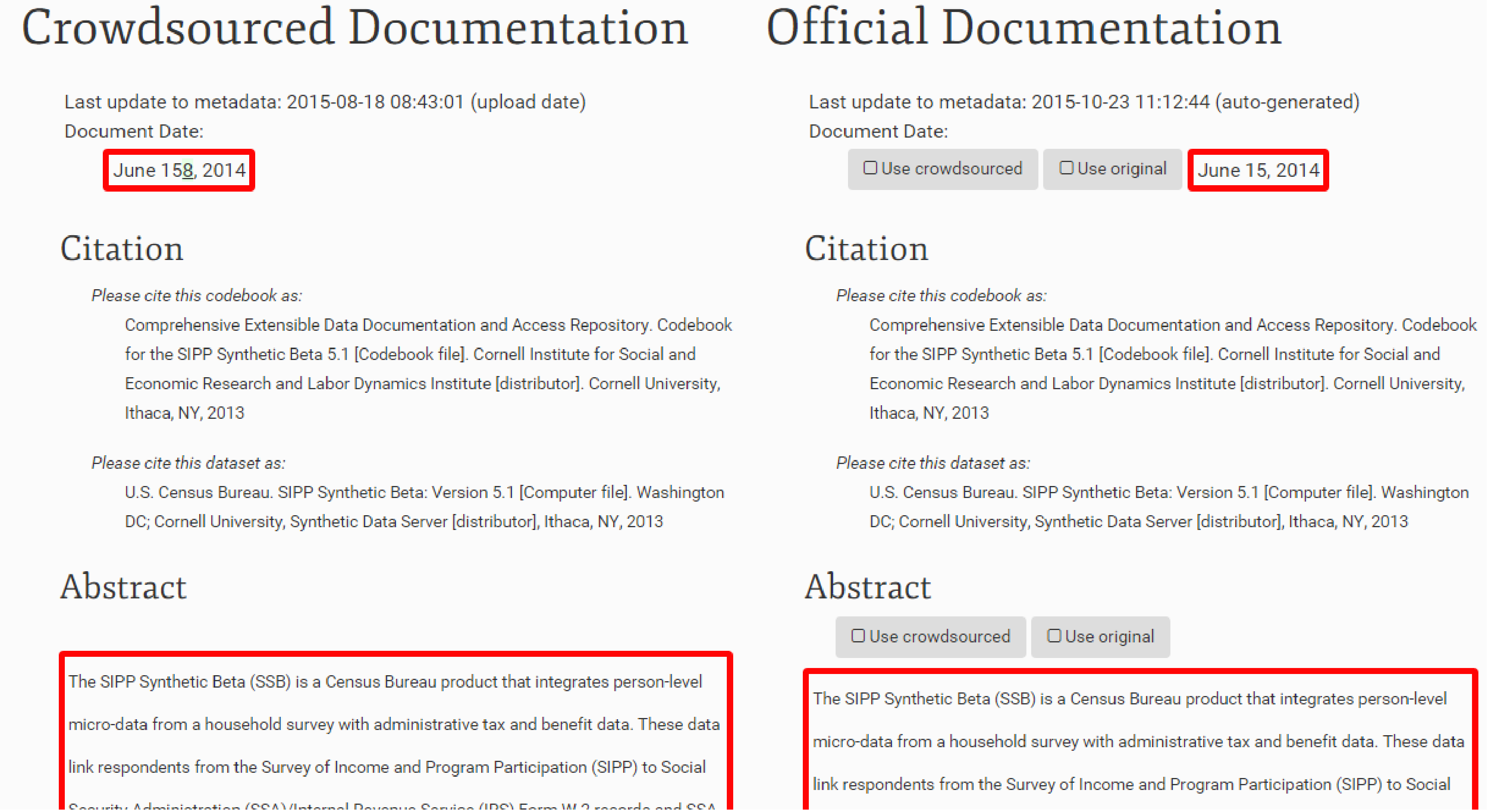
Crowdsourced Documentation

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

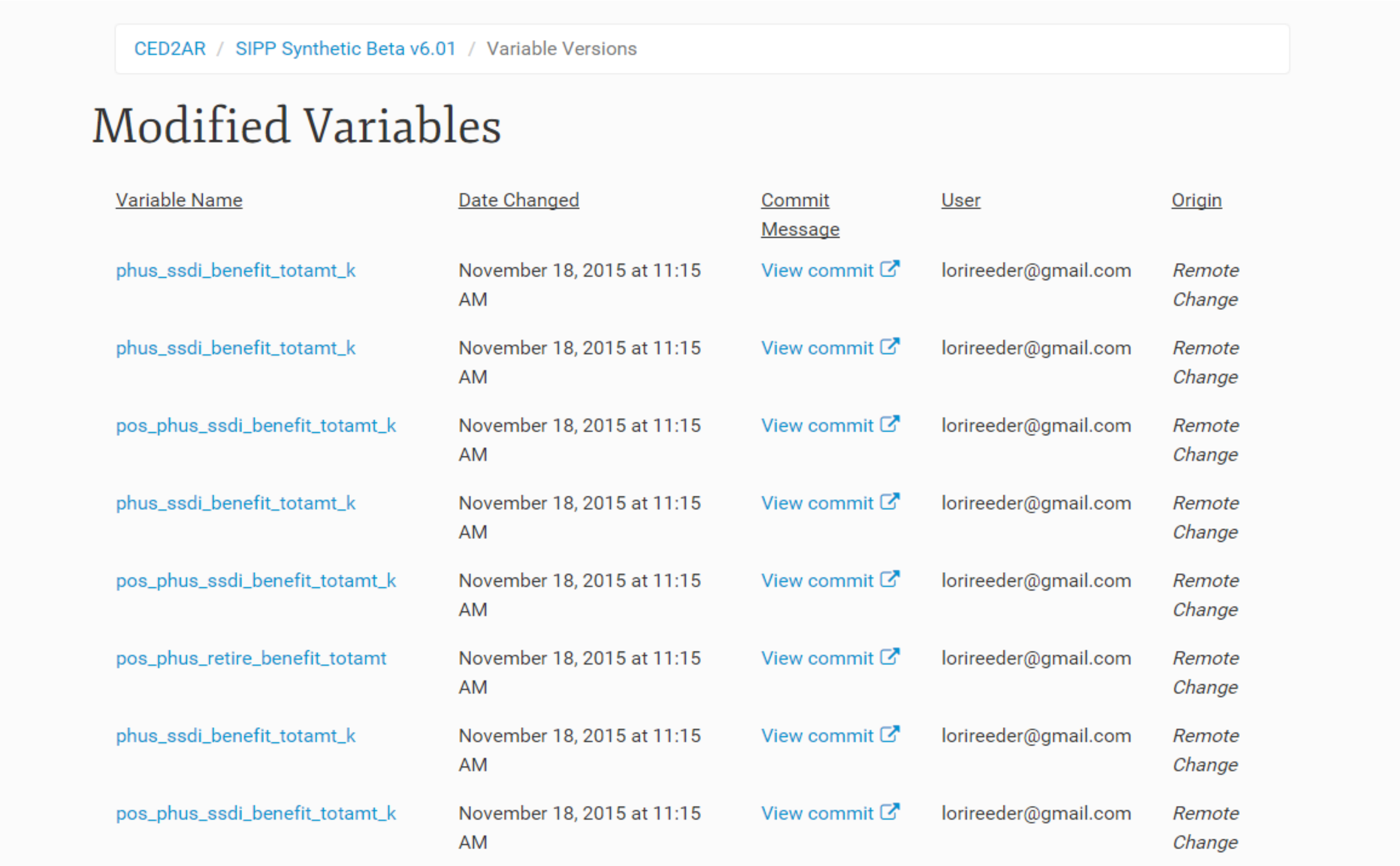
Abstract

☐ 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

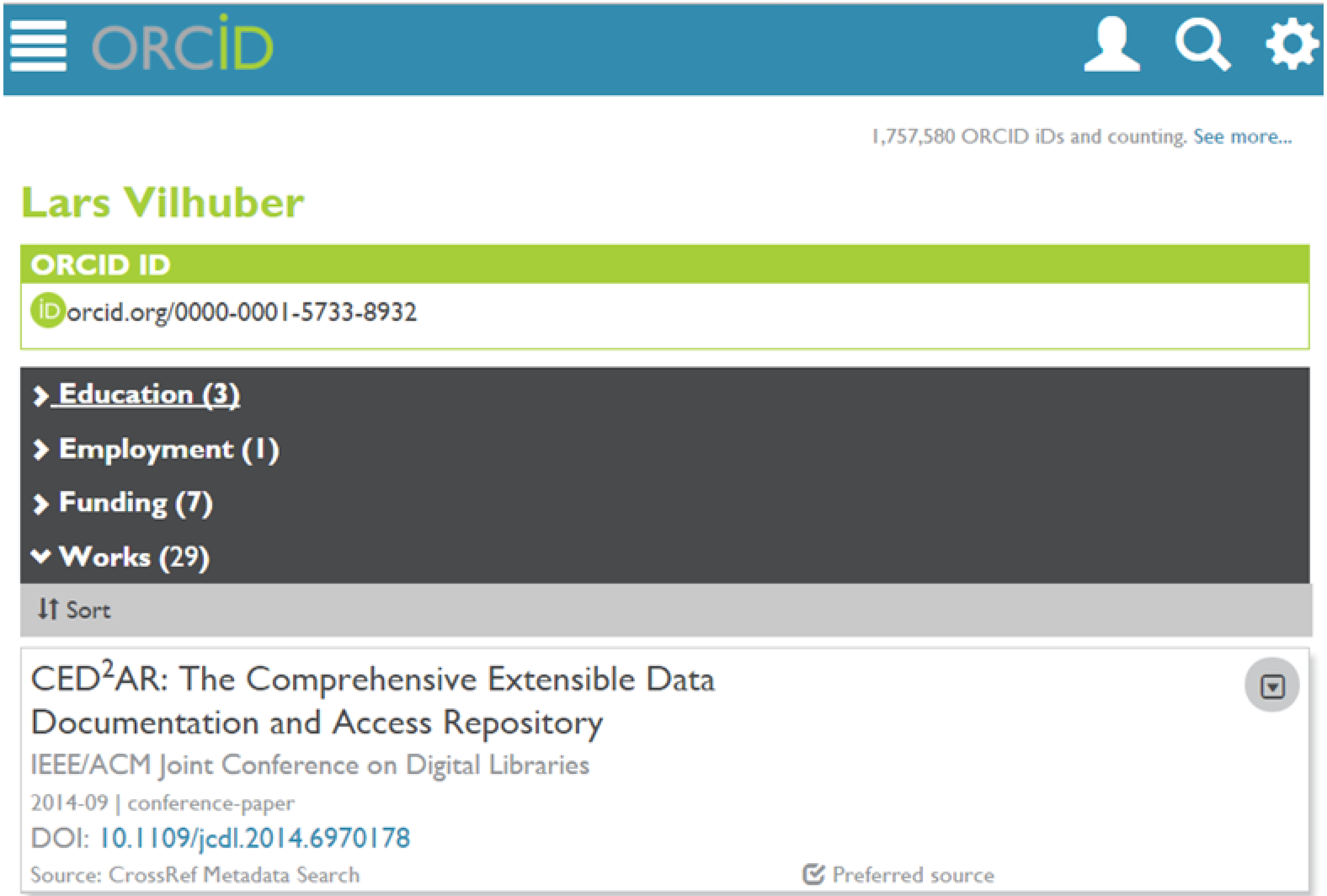


CED2AR / SIPP Synthetic Beta v6.01 / Variable Versions

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



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

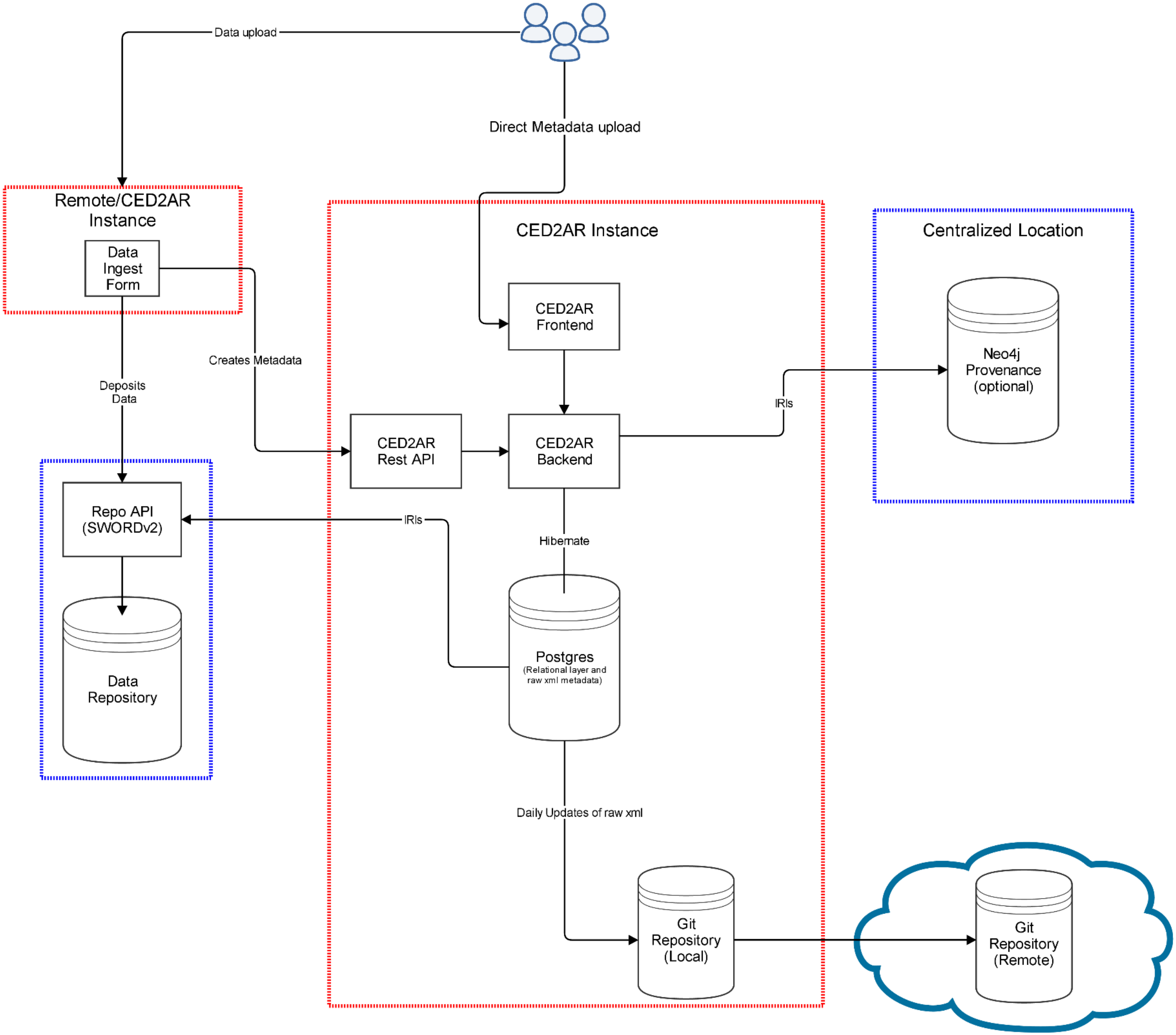
Next steps

Making CED^{2}AR 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 CED^{2}AR scalable in V3

- Current implementation of CED^{2}AR 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 CED^{2}AR is tied to a limited schema – adding schema is hard


Data upload
Direct Metadata upload
Remote/CED2AR Instance
Data Ingest Form
Creates Metadata
Deposits Data
CED2AR Instance
CED2AR Frontend
CED2AR Rest API
CED2AR Backend
IRIs
Centralized Location
Neo4j Provenance (optional)
Repo API (SWORDv2)
Data Repository
IRIs
Hibernate
Postgres
(Relational layer and raw xml metadata)
Daily Updates of raw xml
Git Repository (Local)
Git Repository (Remote)

Thank you!

Questions?

ced2ar-devs-l@cornell.edu