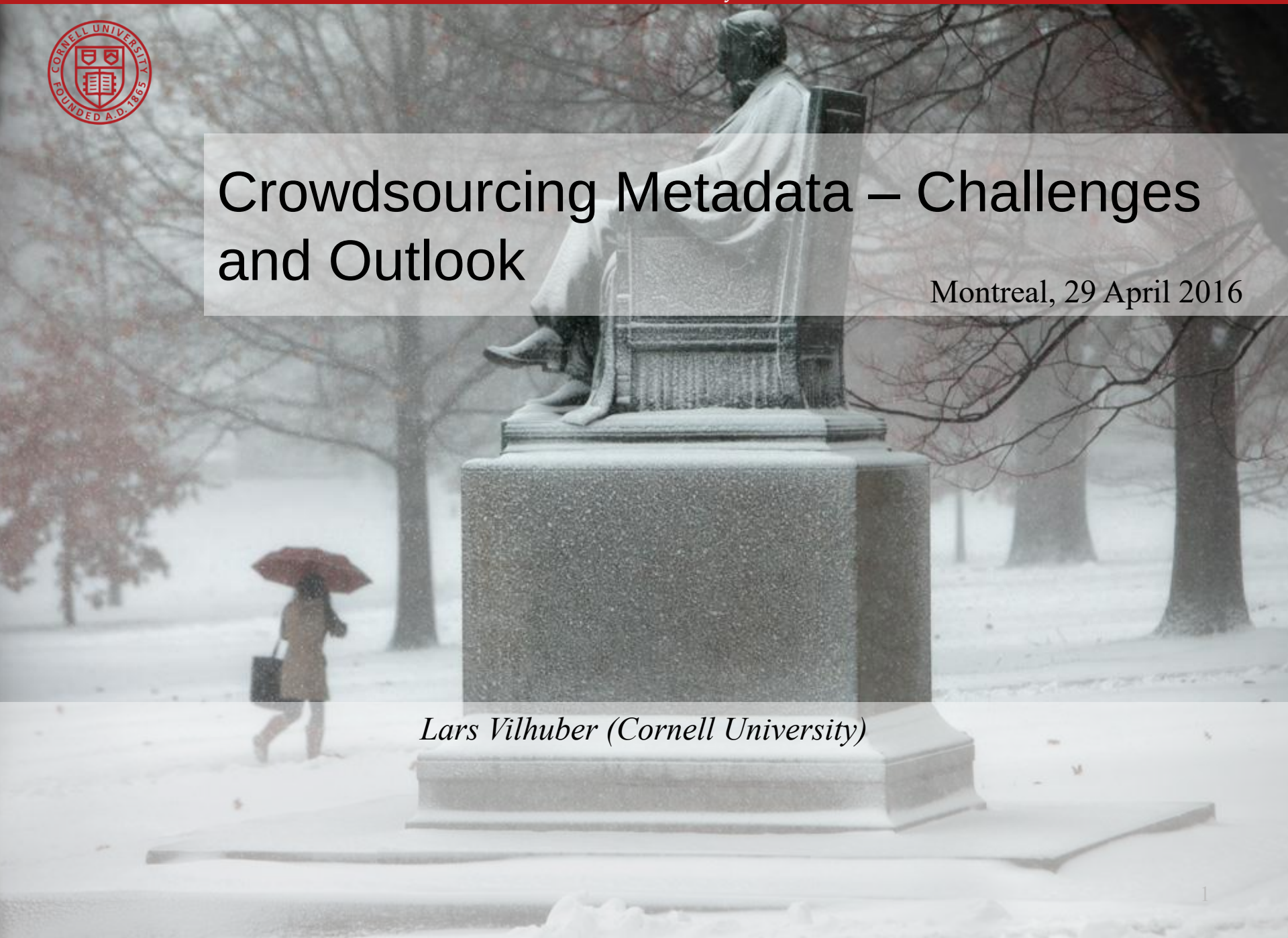
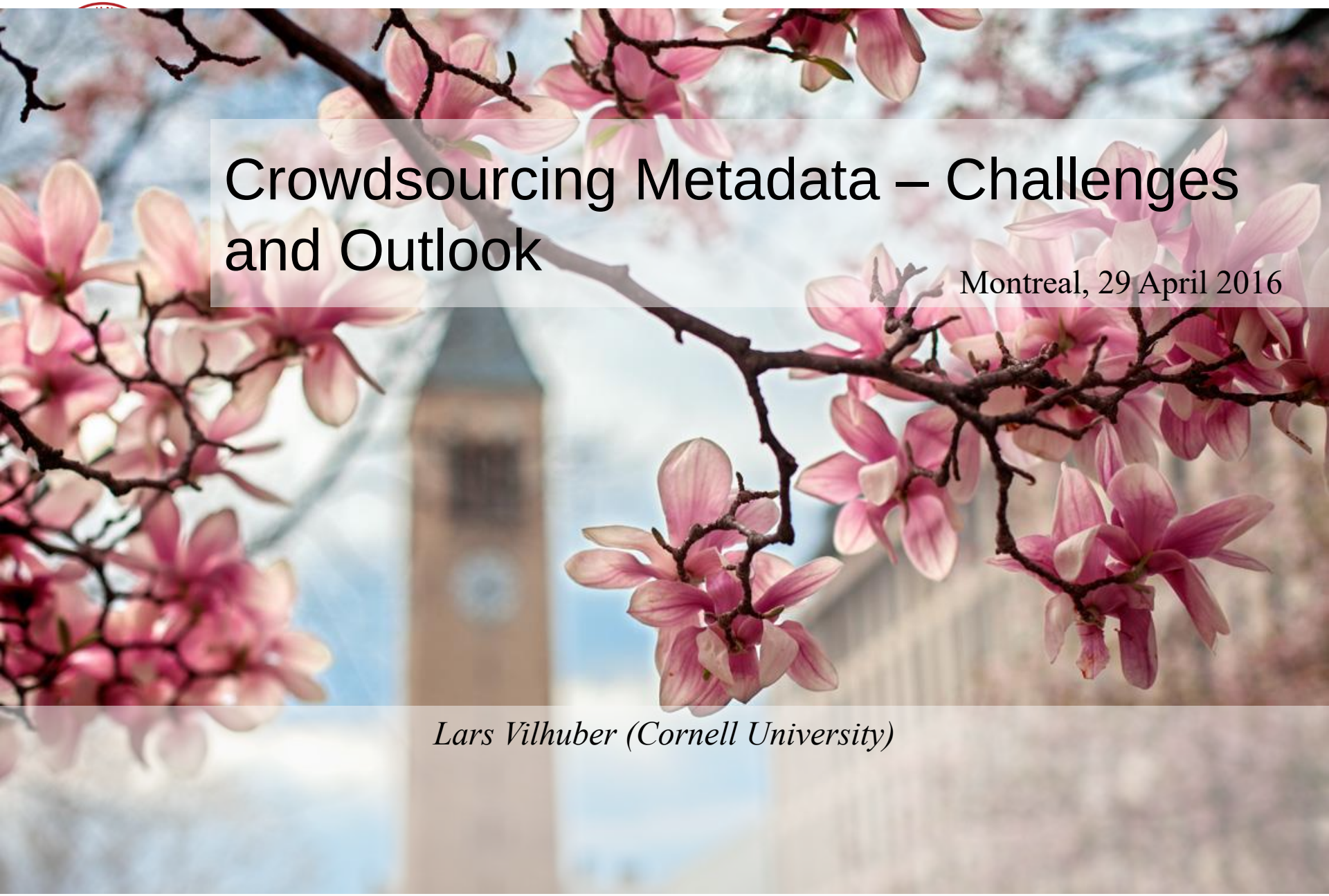


Crowdsourcing Metadata – Challenges and Outlook

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:

Open Science Collaboration (2015) "Estimating the reproducibility of psychological science," *Science* 28 Aug 2015: Vol. 349, Issue 6251, DOI: 10.1126/science.aac4716:

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”

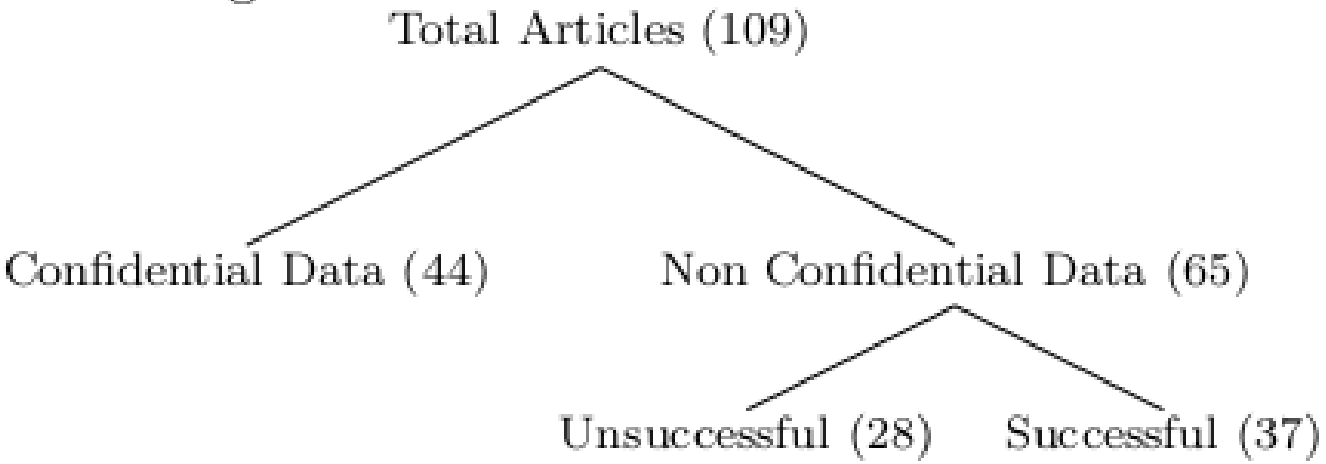
Kingi, Stanchi, Vilhuber (2016)

- 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



Kingi, Stanchi, Vilhuber (2016)

- New question:

Are the provided data and programs accessible?

Table 3: Type of Confidential Data

	Admin local	Admin National	Admin Regional	Private Commercial	Private Other	Total
2010	2	8	0	4	3	17
2011	2	9	4	1	0	16
2013	2	2	1	4	2	11
Total	6	19	5	9	5	44

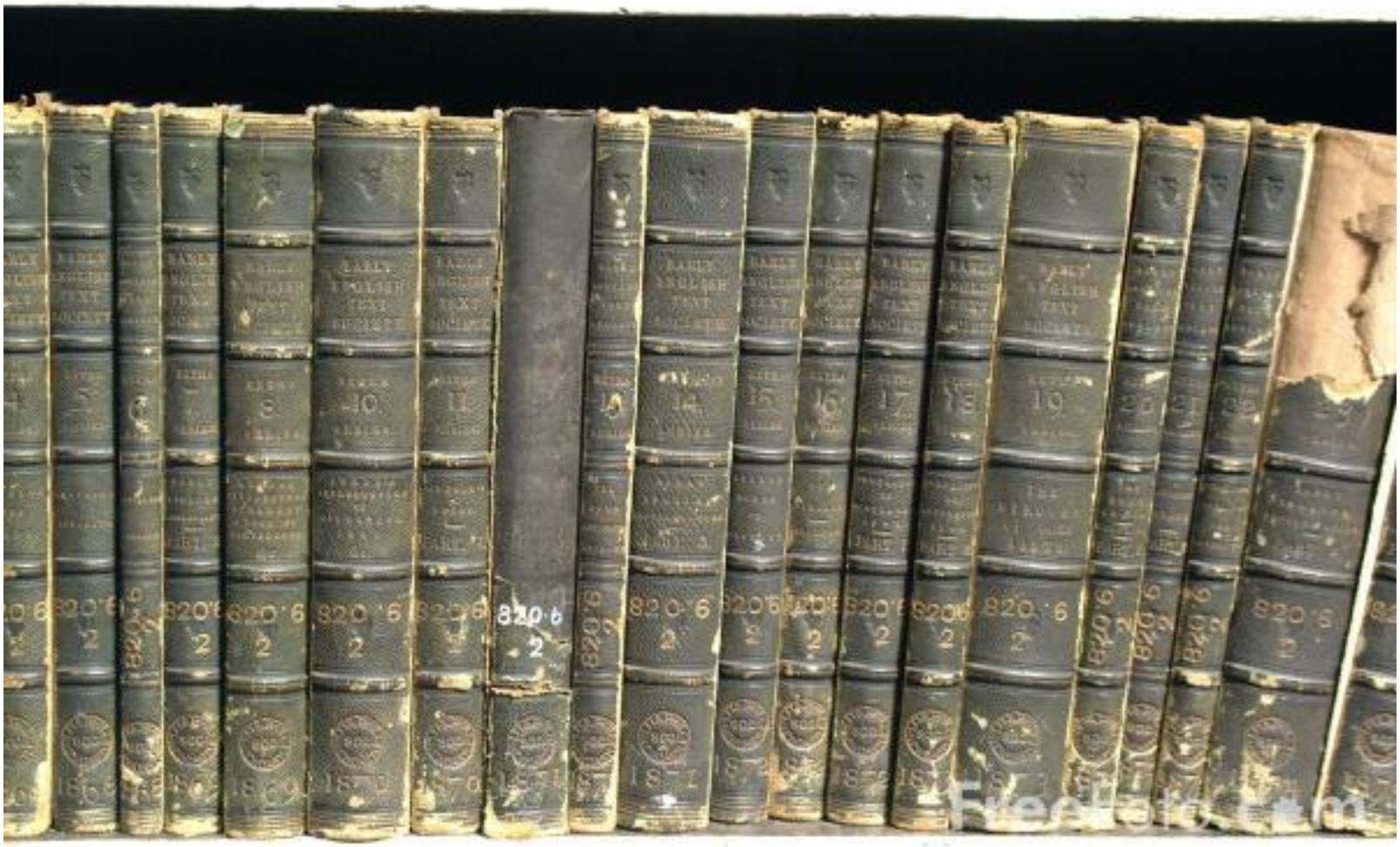
Table 4: Type of Access to Confidential Data

	Formal	Informal Commitment	Informal No Commitment	No Info	Total
2010	2	3	9	3	17
2011	3	0	10	3	16
2013	1	2	8	0	11
Total	6	5	27	6	44

What's part of the problem?

A: Lack of proper deposit

The old source of knowledge – and data!

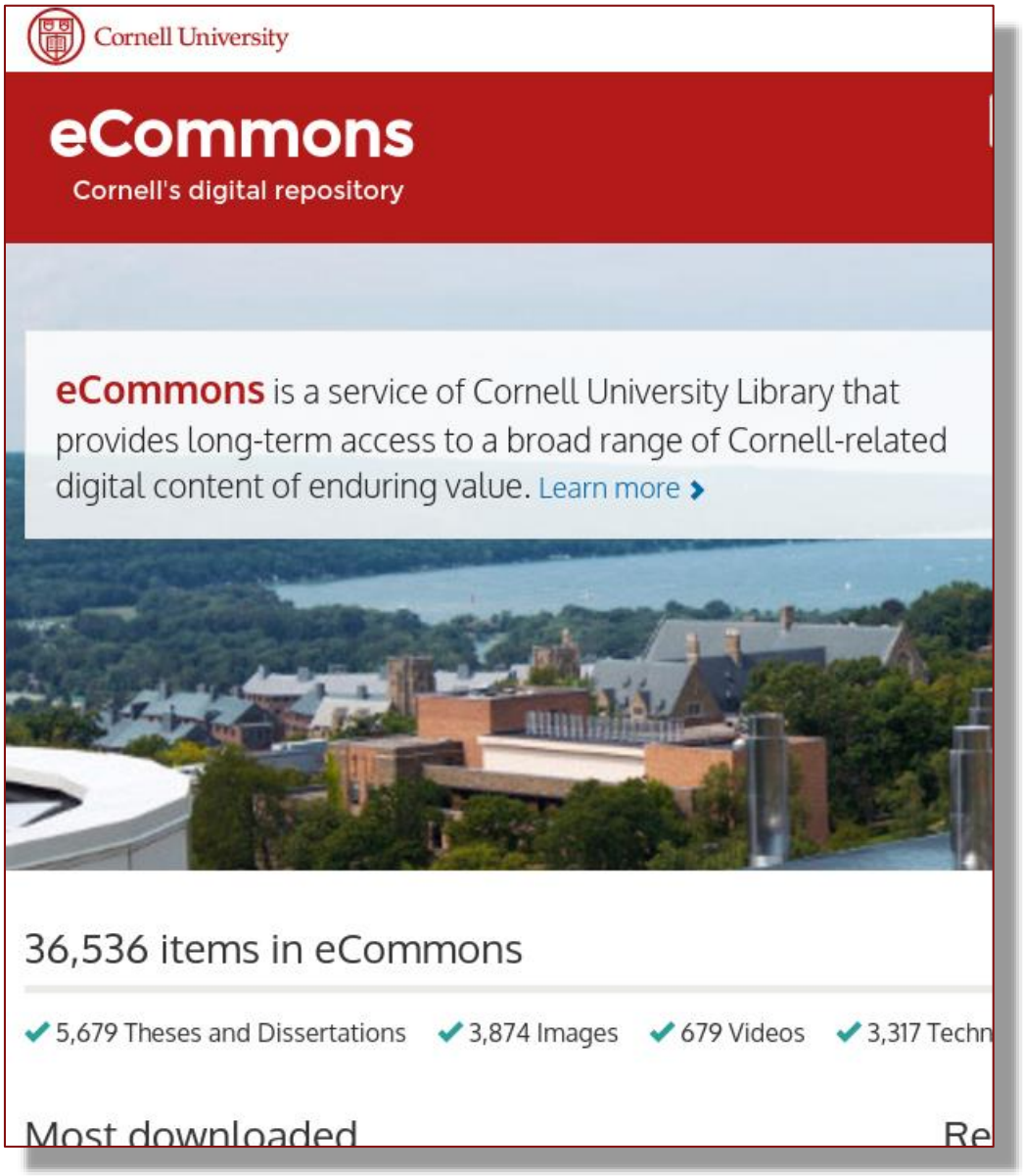


- Source: http://www.thecontrarianmedia.com/2011/08/dispatches-from-the-stacks-15/
- License unknown

It’s mental…

Where do your university's Ph.D. theses go?

Or more likely



Where do your researchers' data go?

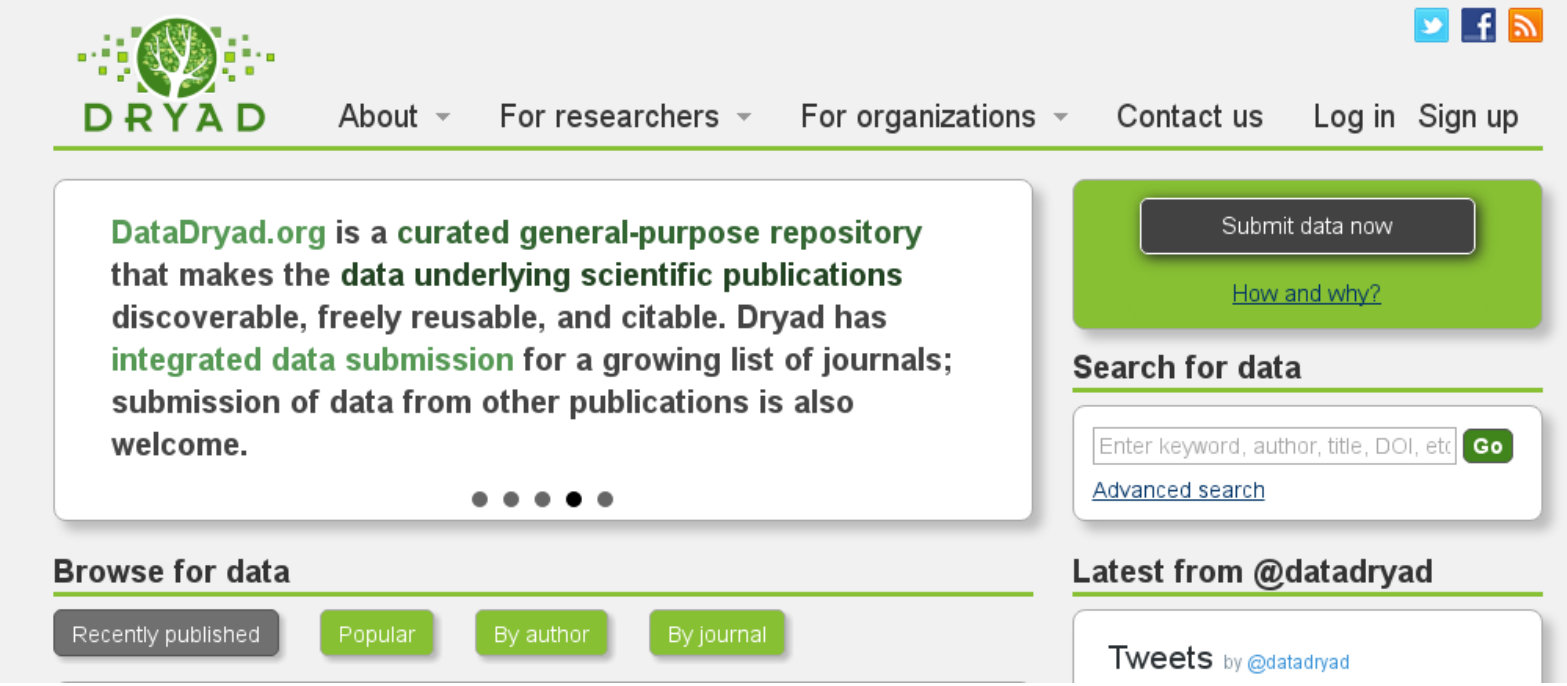
- Their computer?
- Dropbox?

Where can researchers go?

Options are available

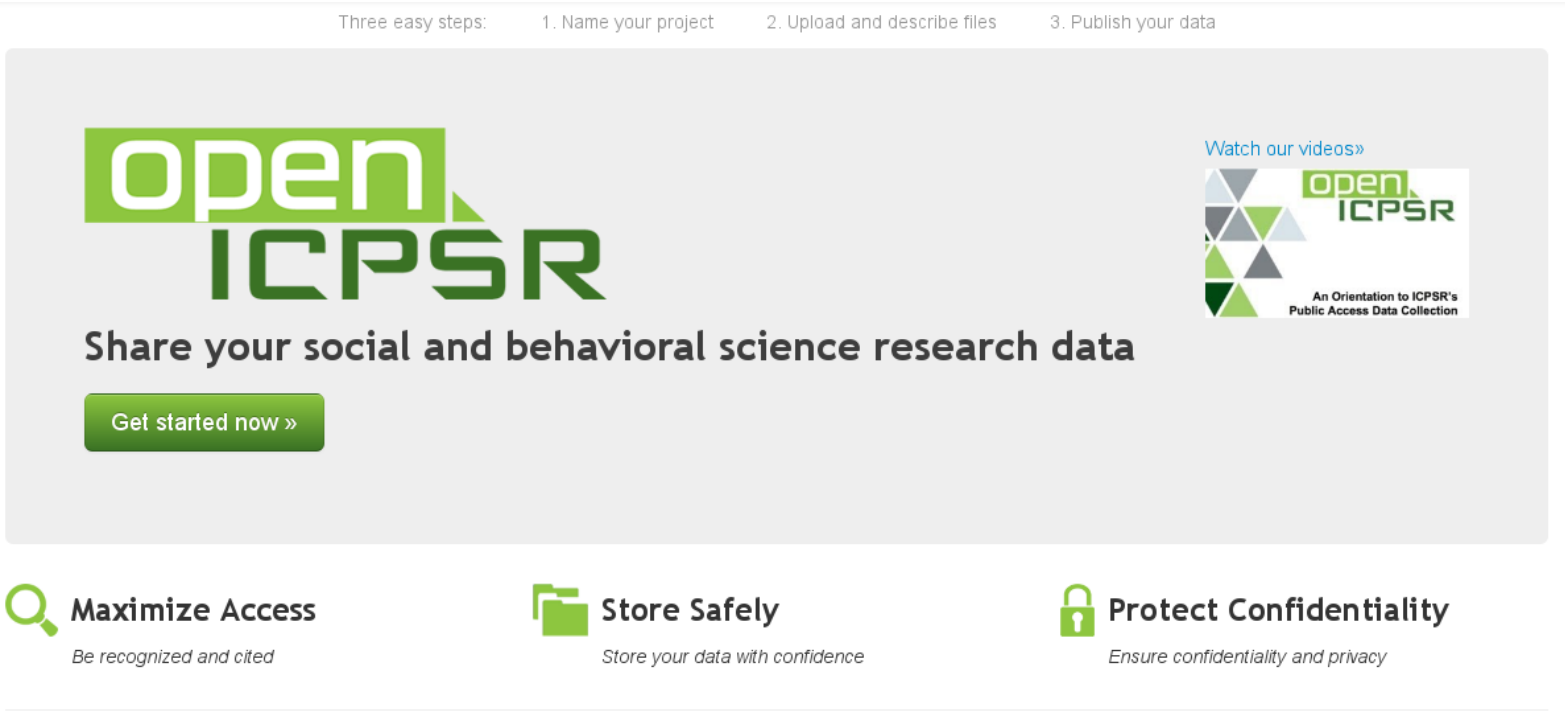
Options are available

- Earth and natural sciences



Options are available

- Social and behavioral sciences



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?

Three main problems

Three main problems

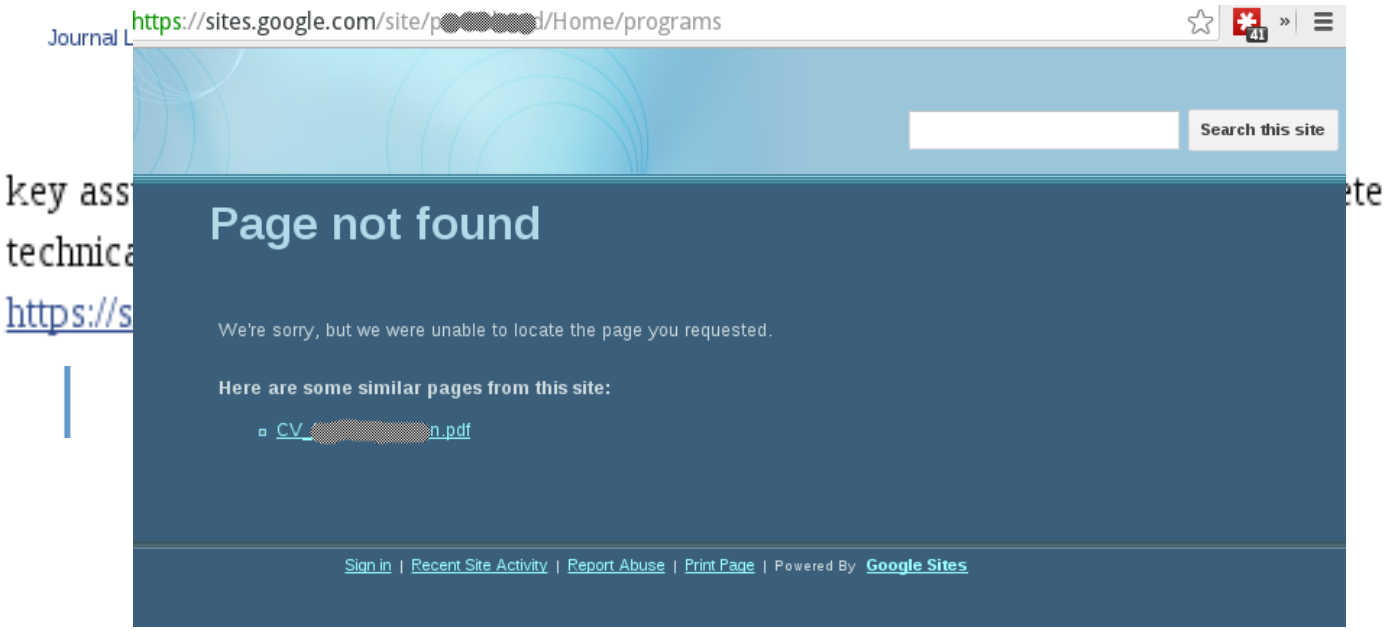
- Researchers don't use them
  - When researchers use them, usage is suboptimal (but better than nothing!)
- Journals don't point to them
- Don't work for many researchers at all

Researchers don't use them

Training? Incentives? Ease of use?

Some Case Studies

Self-archiving



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."

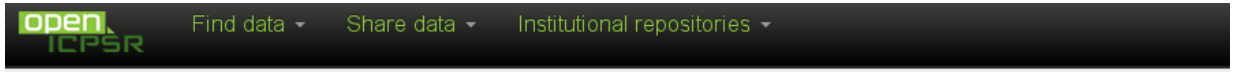
Problems When There Is a Will

- 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)

- 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"

Full-text Article

Additional Materials

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

Authors

A Self-serving Example

Abowd and Vilhuber (2012)

- Article: http://www.aeaweb.org/articles.php?doi=10.1257/aer.102.3.589
- Appendix
  - Description at http://www.aeaweb.org/aer/data/may2012/2012_2790_app.pdf (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

Press for the NBER; 2009. pp. 149–230.

5. Abowd JM, Vilhuber L. The sensitivity of economic statistics to coding errors in personal identifiers. Journal of Business and Economic Statistics. 2005;23(2):133–152.
6. Abowd JM, Zellner A. Estimating Gross Labor Force Flows. Journal of Business and Economic Statistics. 1985;3:254–283.

Abowd and Vilhuber (2011)

J Econom. Author manuscript; available in PMC 2012 Mar 1.
Published in final edited form as:
J Econom. 2011 Mar 1; 161(1): 82–99.
doi: 10.1016/j.jeconom.2010.09.008

PMCID: PMC3079891
NIHMSID: NIHMS246950

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

John M. Abowd and Lars Vilhuber

Author information ▶ Copyright and License information ▶

Abstract

Go to:

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

Abowd and Vilhuber (2011)

Dataverse About

Lars Vilhuber Dataverse (Cornell University)

Harvard Dataverse > Lars Vilhuber Dataverse >

Replication data for: National estimates of gross employment and job flows from the Quarterly Workforce Indicators with der

Metrics 4 Downloads

Replication data for: National estimates of gross employment and job flows from the Q
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/27923, 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 ev
Bureau. Unlike any other local labor market series produced in the U.S. or the rest of the v
flows for workers (accession and separations), jobs (creations and destructions) and earn
and sex), economic industry (NAICS industry groups), and detailed geography (county, Cc
Workforce Investment Area, as well as experimental, unreleased block-level estimates). Jo
the existing 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

Keyword	Employment Dynamics
Topic Classification	Economics
Related Publication	John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Wor with demographic and industry detail," Journal of Econometrics, vol. 161, iss. 1, pp. 82-99, 2011. doi: 10.1016/j.jeconom.2010.09.008 http://www2.vrdc.cornell.edu/news/data/qwi-national-data/ John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Wor with demographic and industry detail," Journal of Econometrics, vol. 161, iss. 1, pp. 82-99, 2011. doi: 10.1016/j.jeconom.2010.09.008 http://www2.vrdc.cornell.edu/news/data/qwi-national-data/ John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Wor with demographic and industry detail (with color graphs)," Center for Economic Studies, U.S. Census Bureau, Wo 11, 2010. http://ideas.repec.org/p/cen/wpaper/10-11.html
Producer	Labor Dynamics Institute (Cornell University) (LDI) http://www2.vrdc.cornell.edu/news/data/qwi-national-data/

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

THE QUARTERLY JOURNAL OF ECONOMICS

ABOUT THIS JOURNAL | CONTACT THIS JOURNAL | SUBSCRIPTIONS | CURRENT ISSUE | ARCHIVE | SEARCH

Oxford Journals › Social Sciences › Quarterly Journal of Economics

Click here to sign up for email table of contents alerts from *The Quarterly Journal of Economics*

READ THIS JOURNAL

View Current Issue (Volume 131 Issue 1 February 2016)
Advance Access
Browse the archive
Forthcoming Articles
Browse by JEL code

The Quarterly Journal of Economics is the oldest professional journal of economics in the English language. Edited at Harvard University's Department of Economics, it covers all aspects of the field.

QJE is invaluable to professional and academic economists and students around the world.

In an effort to promote consistent standards and requirements among general-interest journals in the field of economics the *Quarterly Journal of Economics* has adopted a new data availability policy. To read the full policy, click here. To navigate to the *QJE*'s Dataverse page, please click here.

The **2015 Sveriges Riksbank Prize in Economic Sciences in memory of Alfred Nobel** was awarded on the 12th of October to Angus Deaton. Click here to read a collection of Professor Deaton's articles from across OUP's economics journals.

***The Quarterly Journal of Economics* is now the top-ranked journal in Economics** according to the 2015 Journal Citation Reports® (Thomson Reuters, 2015). Click here to read a collection of highly cited articles.

***QJE* in the News - find out the latest media coverage.**

Read *QJE* Editors' Choice Articles.

MOST READ FROM THE PAST MONTH

THE JOURNAL

› About the journal
› Free Highly Cited Articles
› Free Editors' Choice Articles
› QJE Dataverse
› Rights & permissions
› We are mobile – find out more
› Journals Career Network

Click here to contact the Editorial Office.

Editorial Office:
Trina Ott, Assistant Editor
1805 Cambridge Street
Cambridge, MA 02138
617-496-3293
qje_admin@editorialexpress.com

Published on behalf of
› President and Fellows of Harvard University

Impact Factor: 6.654
5-Yr impact factor: 9.794

Journals are starting to use them

Dataverse

About Guides Support Sign Up Log In

THE QUARTERLY JOURNAL OF ECONOMICS

The Quarterly Journal of Economics Dataverse (Harvard University) Journal website

Harvard Dataverse > **The Quarterly Journal of Economics Dataverse**

The Quarterly Journal of Economics is the oldest professional journal of economics in the English language. Edited at Harvard University's Department of Economics, it covers all aspects of the field. In an effort to promote consistent standards and requirements among general-interest journals in the field of economics the Quarterly Journal of Economics has adopted a new data availability policy. To read the full policy, click here.

Search this dataverse... Find Advanced Search Add Data

Dataverses (0)
Datasets (0)
Files (0)

This dataverse currently has no dataverses, datasets, or files. Please log in to see if you are able to add to it.

Data Science at The Institute for Quantitative Social Science | Dataverse Project on | Code available at

Copyright © 2016, The President & Fellows of Harvard College | Privacy Policy

Powered by The Dataverse Project v. 4.3 build 22-1e451ff

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

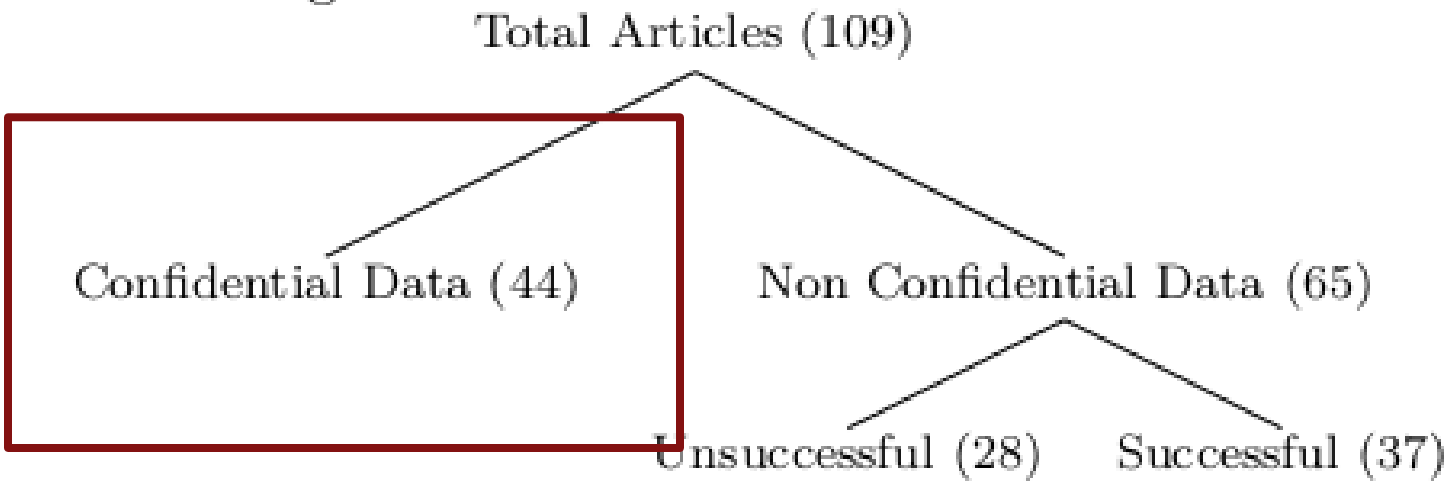
Replication Data for: Navigating the Range of Statistical Tools for Inferential Network Analysis
Apr 21, 2016
Cranmer, Skyler; Leifeld, Philip; McClurg, Scott; Rolfe, Meredith, 2016, "Replication Data for: Navigating the Range of Statistic
Tools for Inferential Network Analysis", http://dx.doi.org/10.7910/DVN/2XP8YF, 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 Educat
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

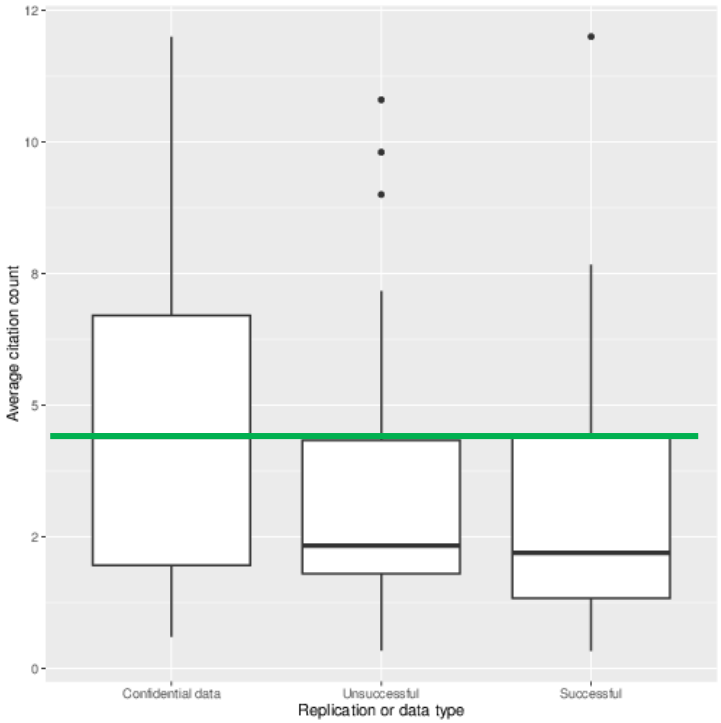
But the biggest problem…

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…

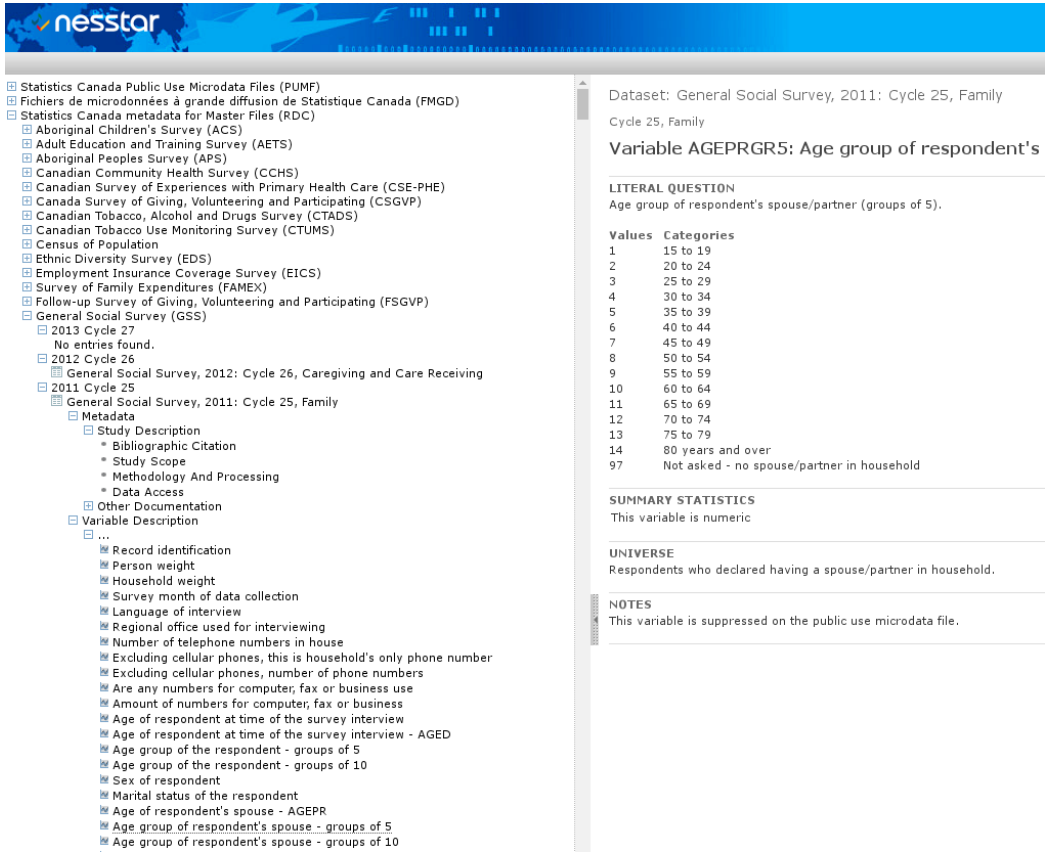
- Data is not available
- Metadata is not available
- Programs? So-so…

Should We Just Trust These Guys?

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

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.

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!

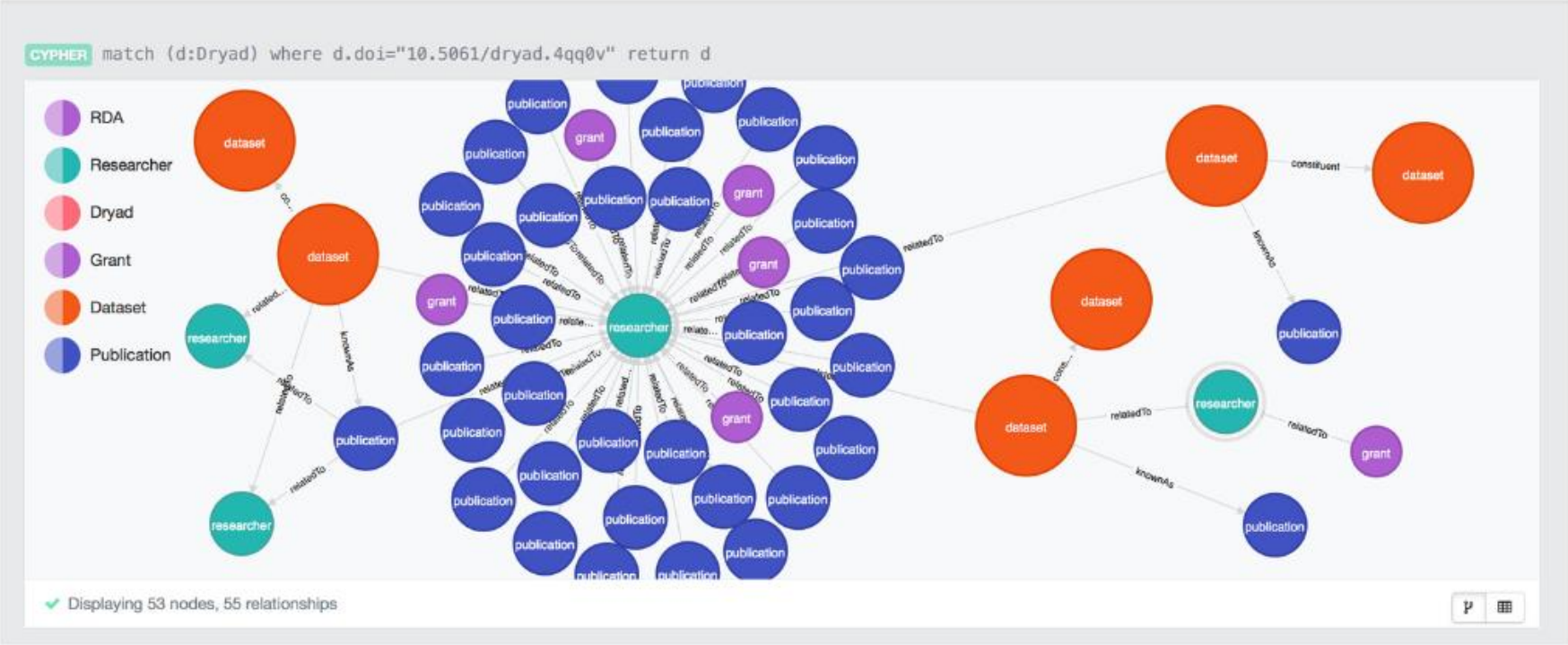
Liberate the data users!

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

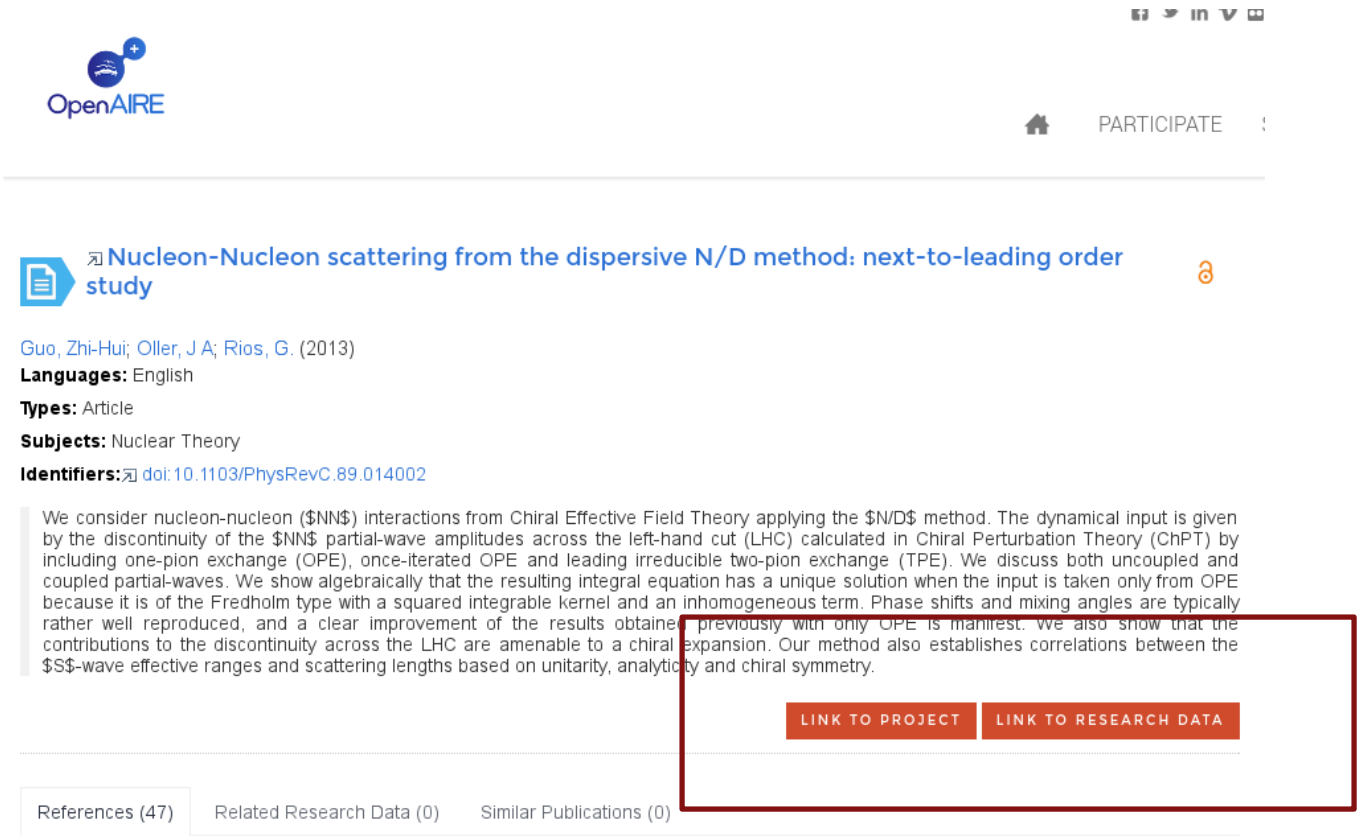
RD Switchboard is under development



But relies on existing metadata

OpenAIRE attempts to do so

- … allowing users to establish the link



Our contribution

Leverage researcher knowledge

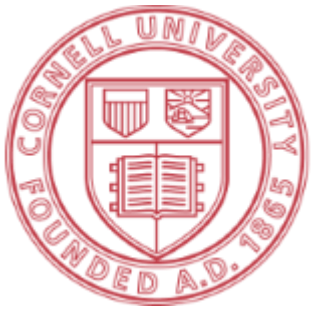
Our Approach

- 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?

CED²AR

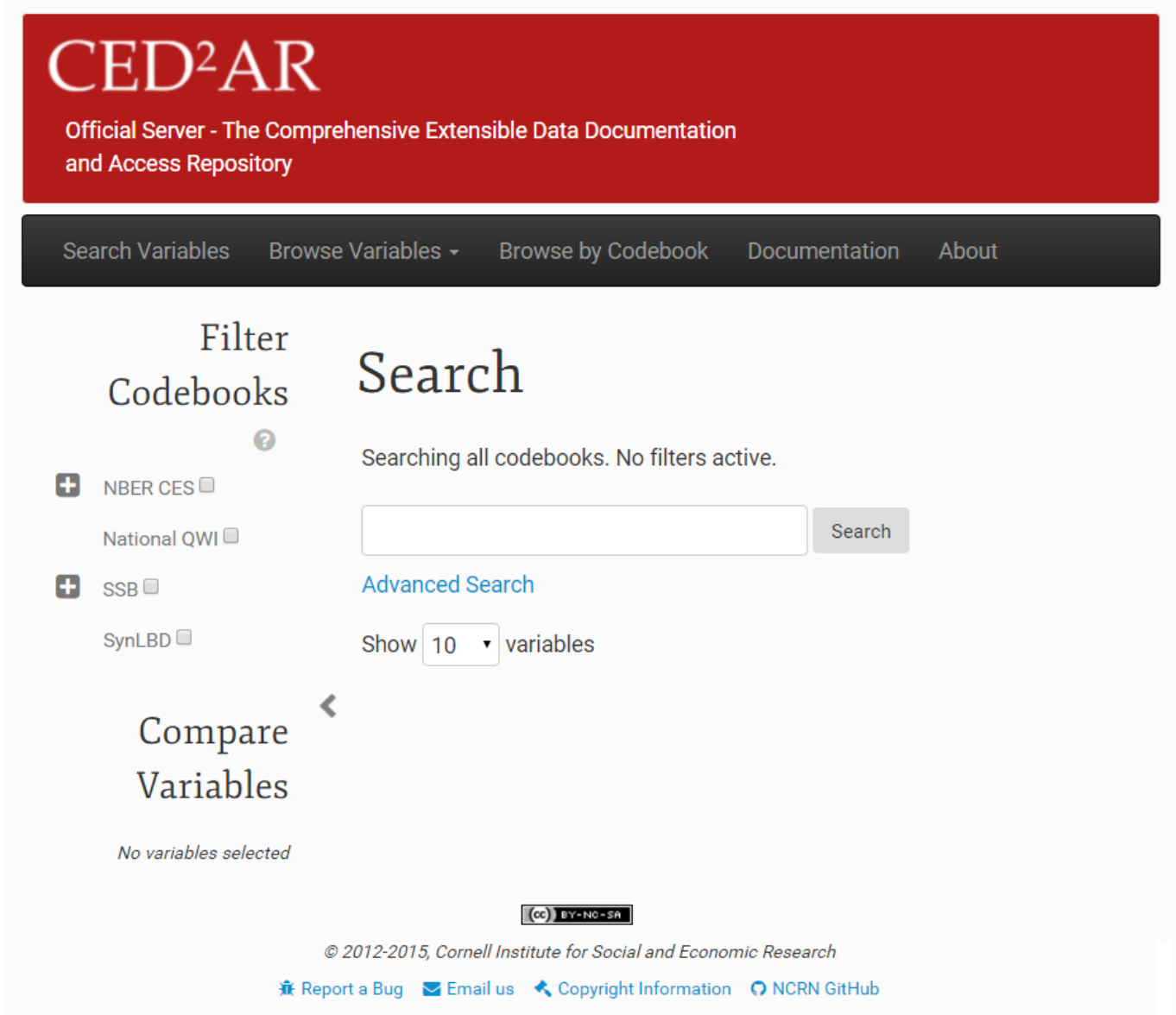
The Comprehensive Extensible Data Documentation and Access Repository



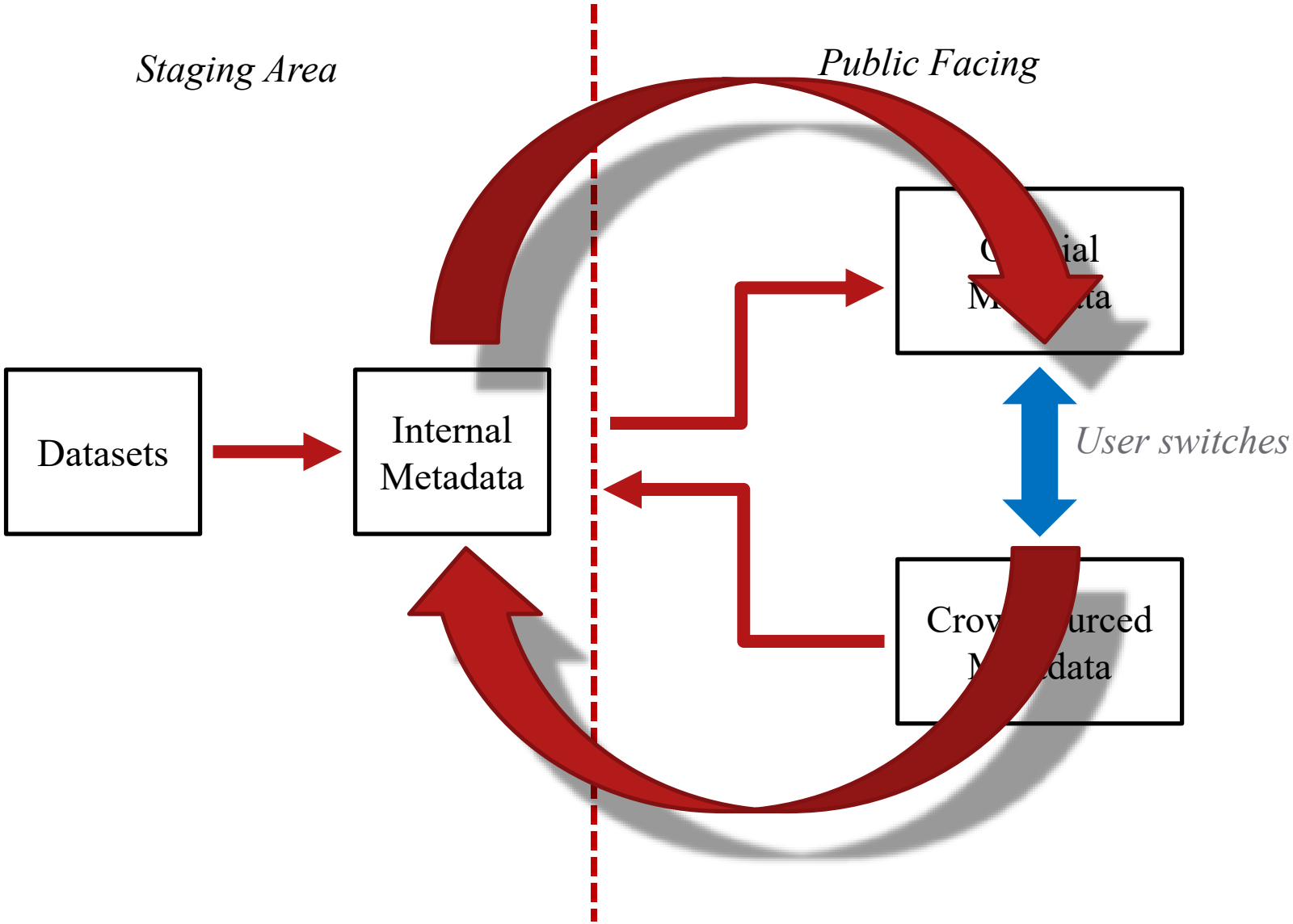
What is CED²AR?

- Metadata curation software
- Designed for documenting existing datasets
- Funded by NSF grant #1131848
- Online at www2.ncrn.cornell.edu/ced2ar-web

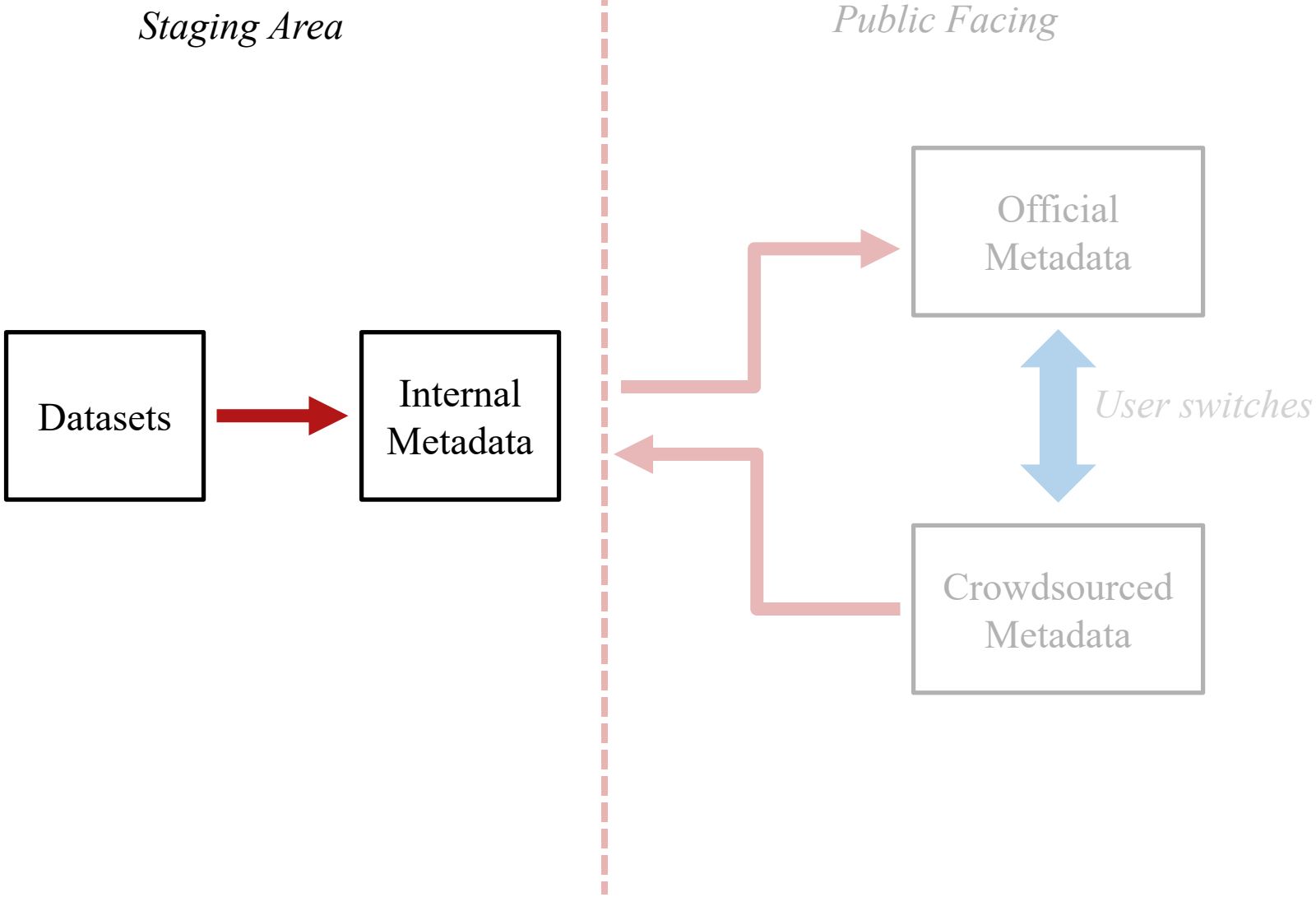
What is CED²AR?



Basic Information Flow



Basic Information Flow



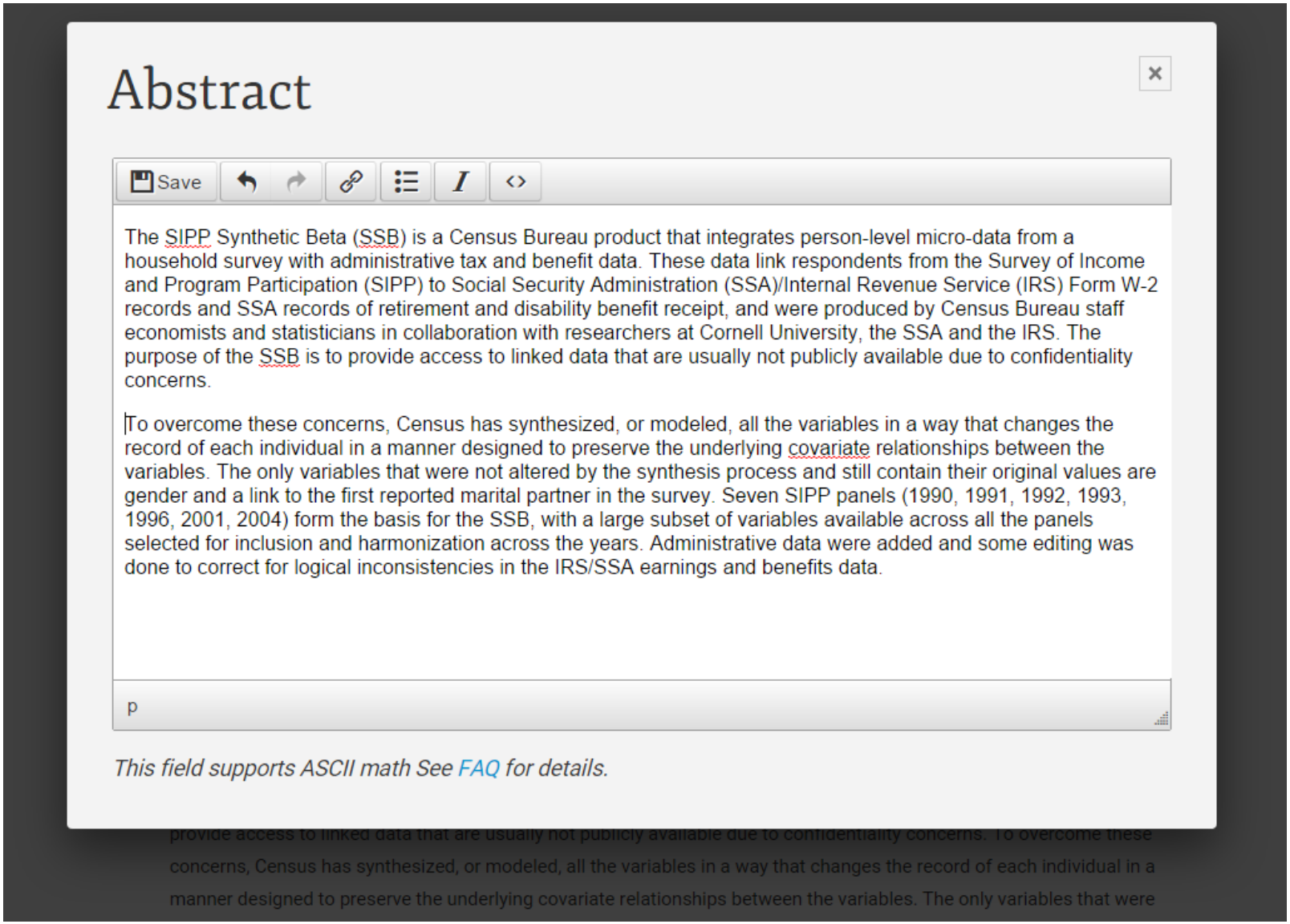
Internal Processing

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

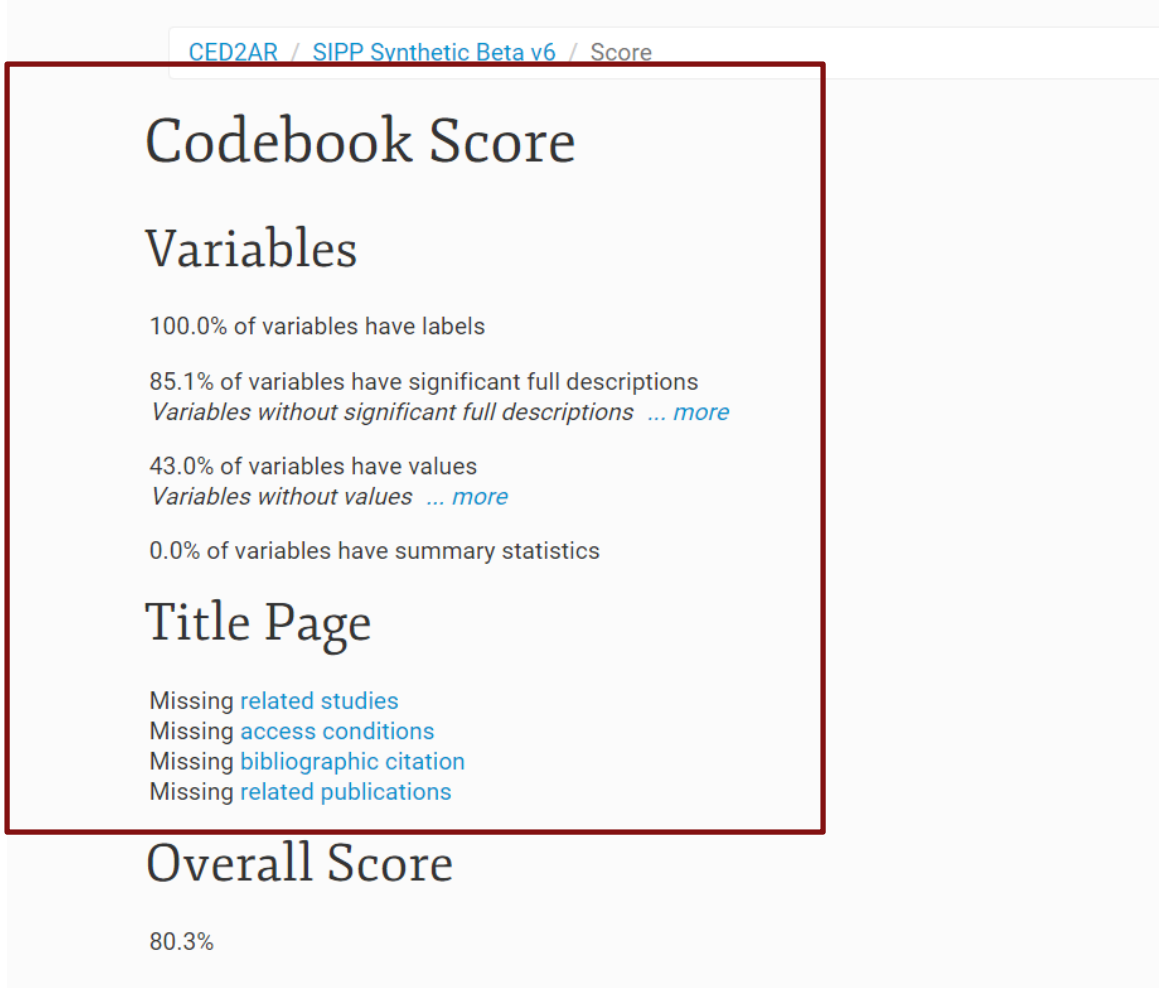
- Simple editing interface
  - Web-based, with limited rich text features
  - Math allowed (LaTeX)
- Feedback
  - Completeness of codebook?
  - Without technical jargon!
  - Can be tuned

Internal Processing: Hand Editing



Internal Processing: Scoring

- Provide feedback to improve sparse documentation

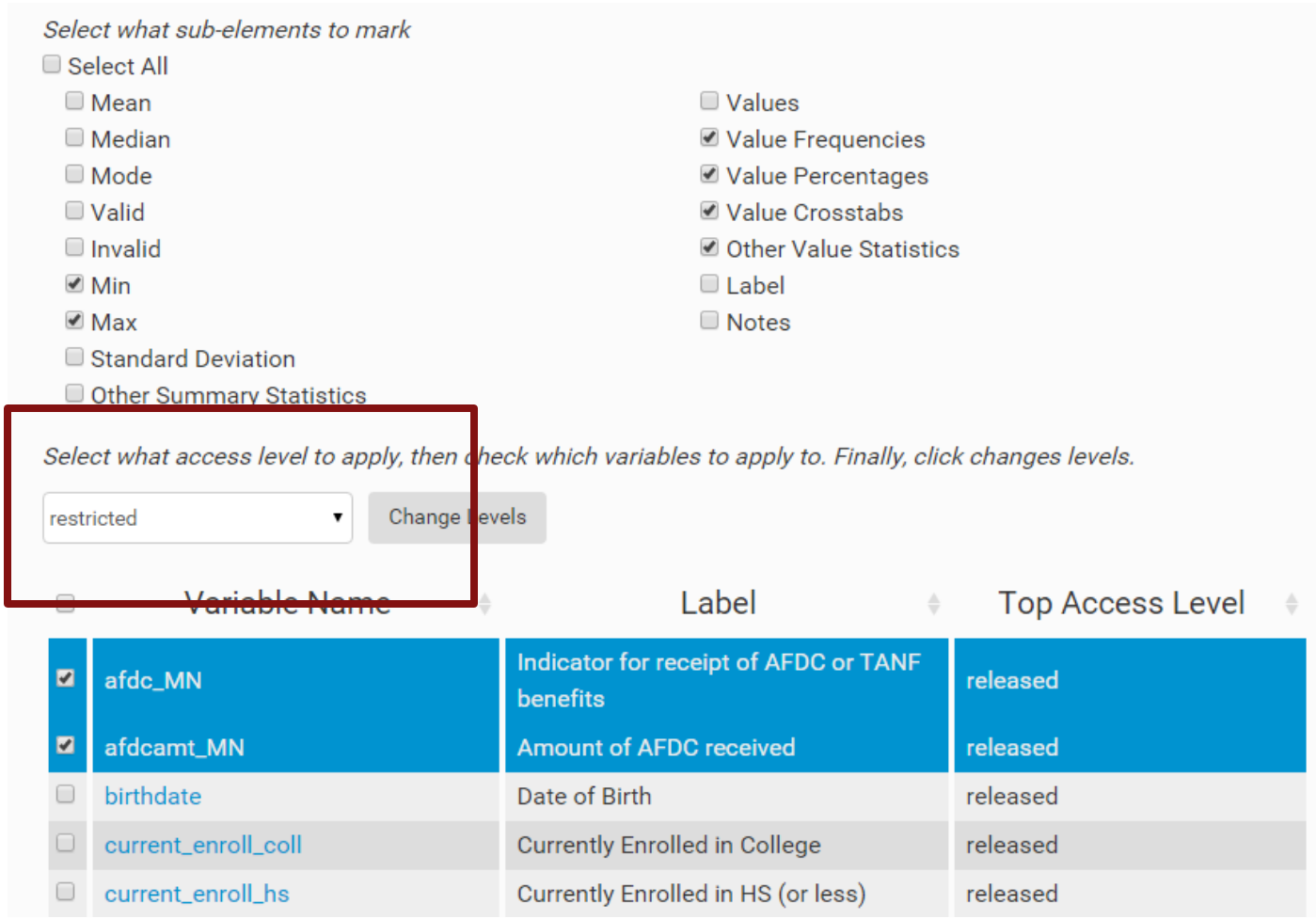


Fine-grained access controls

Important when working with confidential (meta)data

Internal Processing: Access Control

- Marking elements with different restrictions

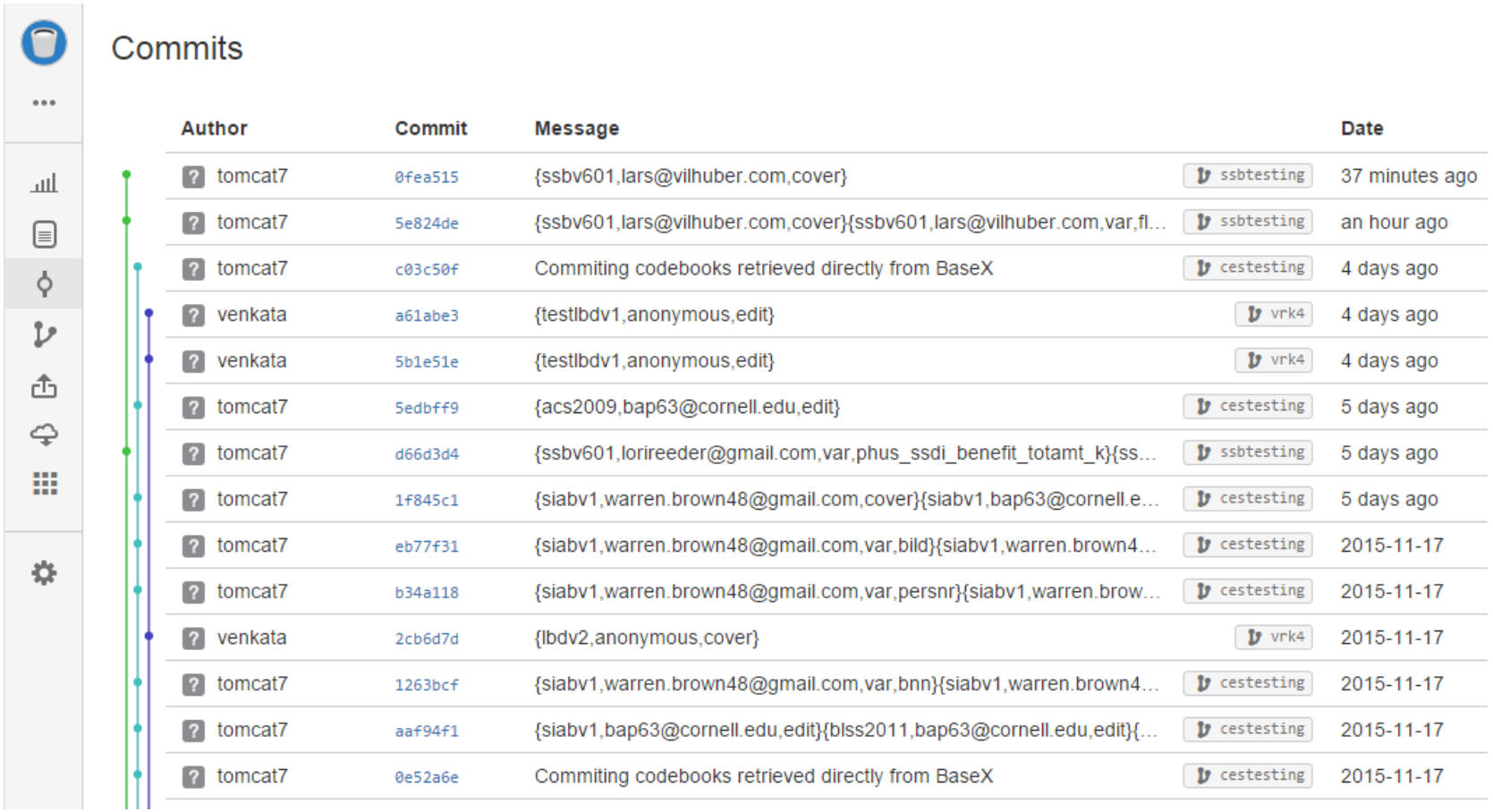


Workflow control

- 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

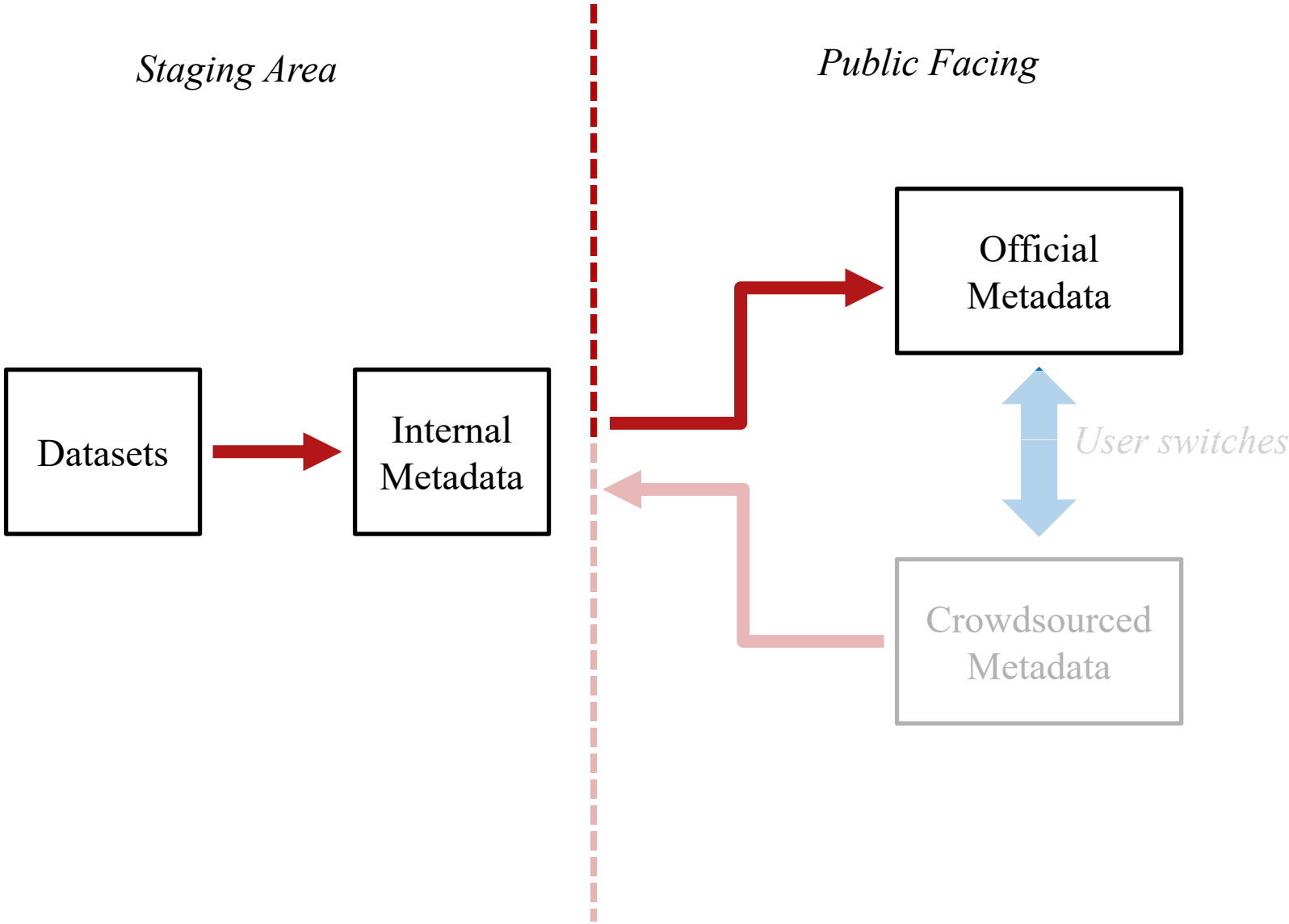
All changes are logged externally via Git



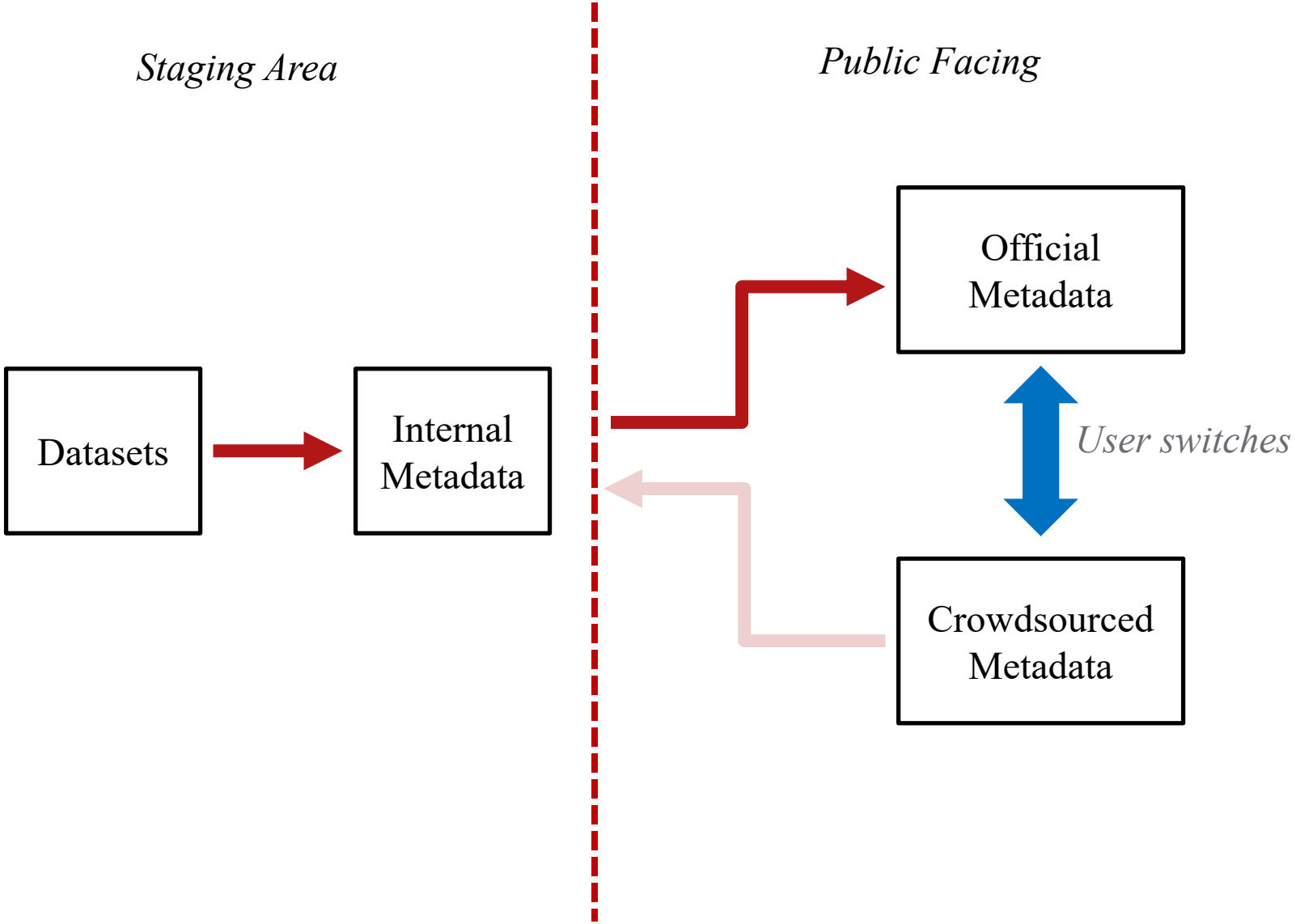
Commits

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

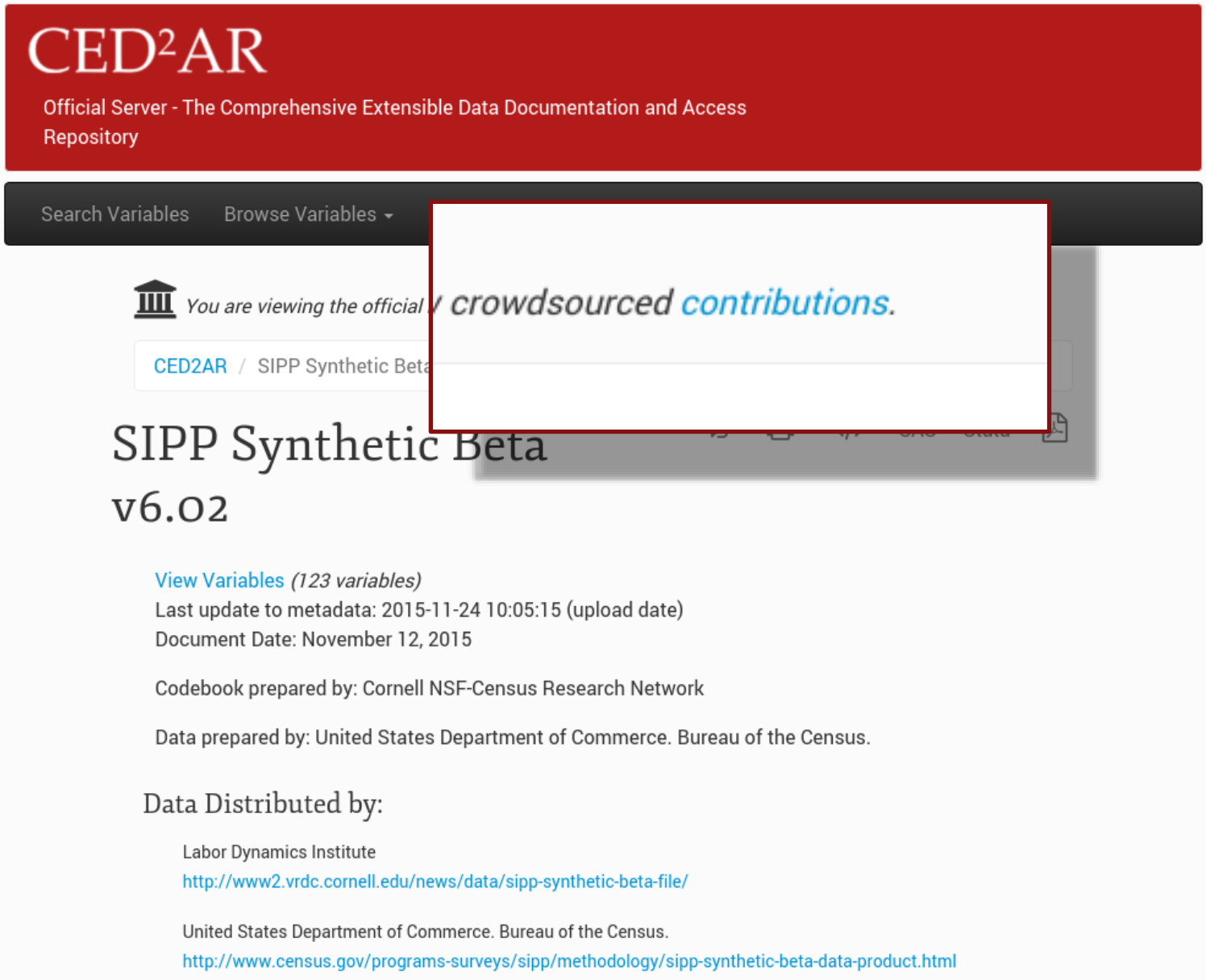
Basic Information Flow



Basic Information Flow



Official view



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/

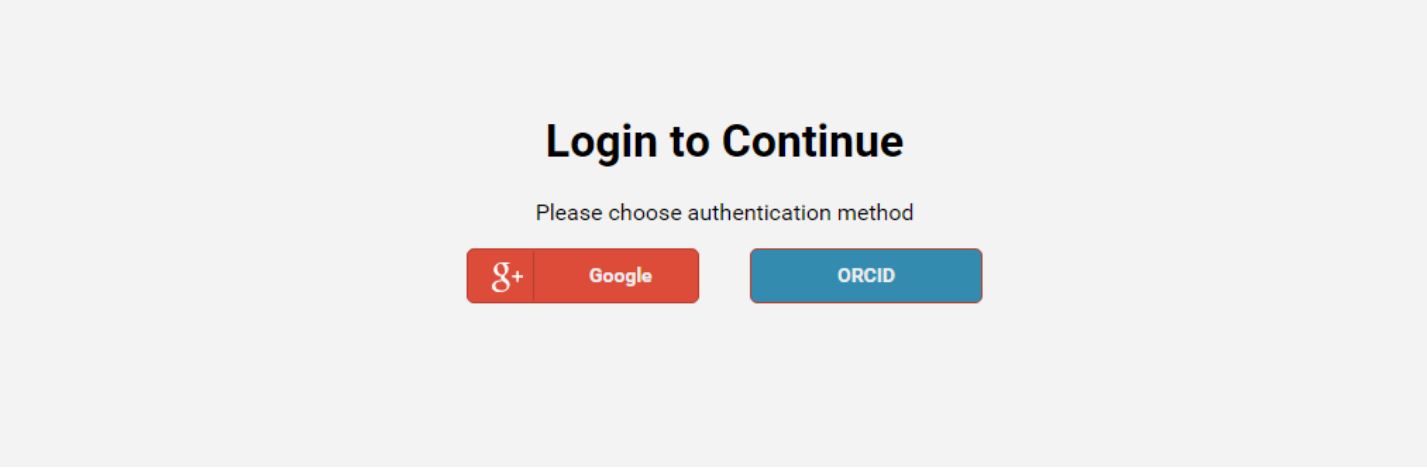
United States Department of Commerce. Bureau of the Census.
http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html

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

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

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).

+ *Add Note*

Type numeric

Notes

Save

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.

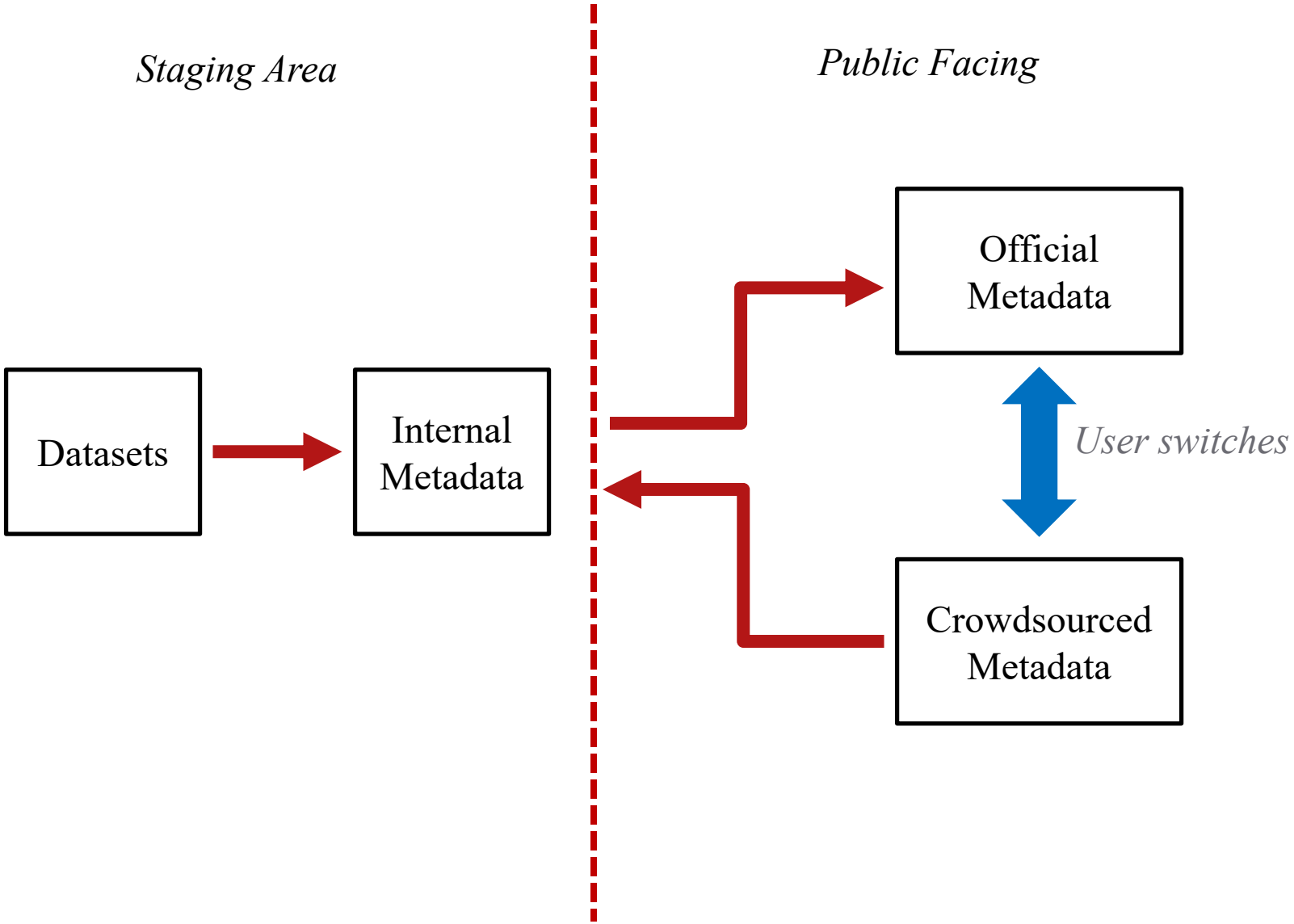
If you are looking at earnings in earlier years, particular the 1960s and earlier, there will be more people with $0 earnings because many jobs were not FICA-taxable then. Even today, there are some instances of legitimate non-FICA earnings that would not be reflected on the SER. One example of this 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).

p

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.

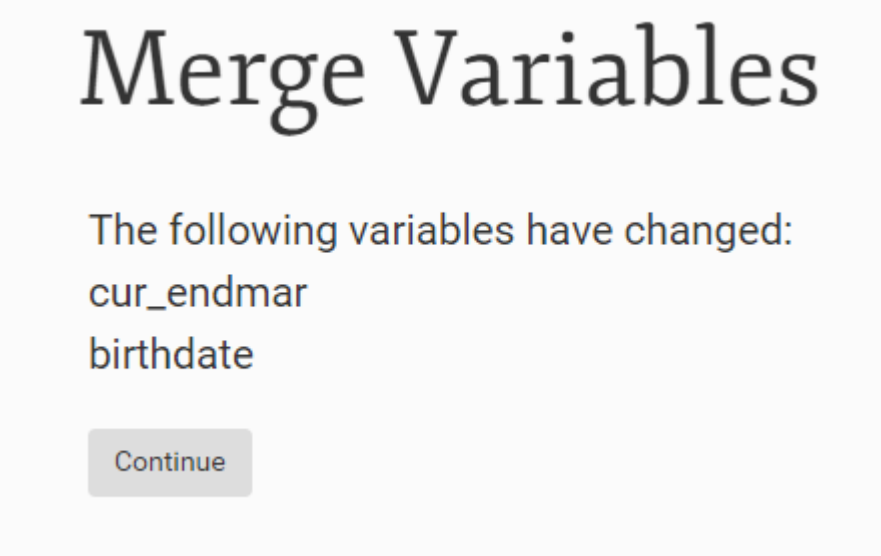
Notes (1 total)

#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



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

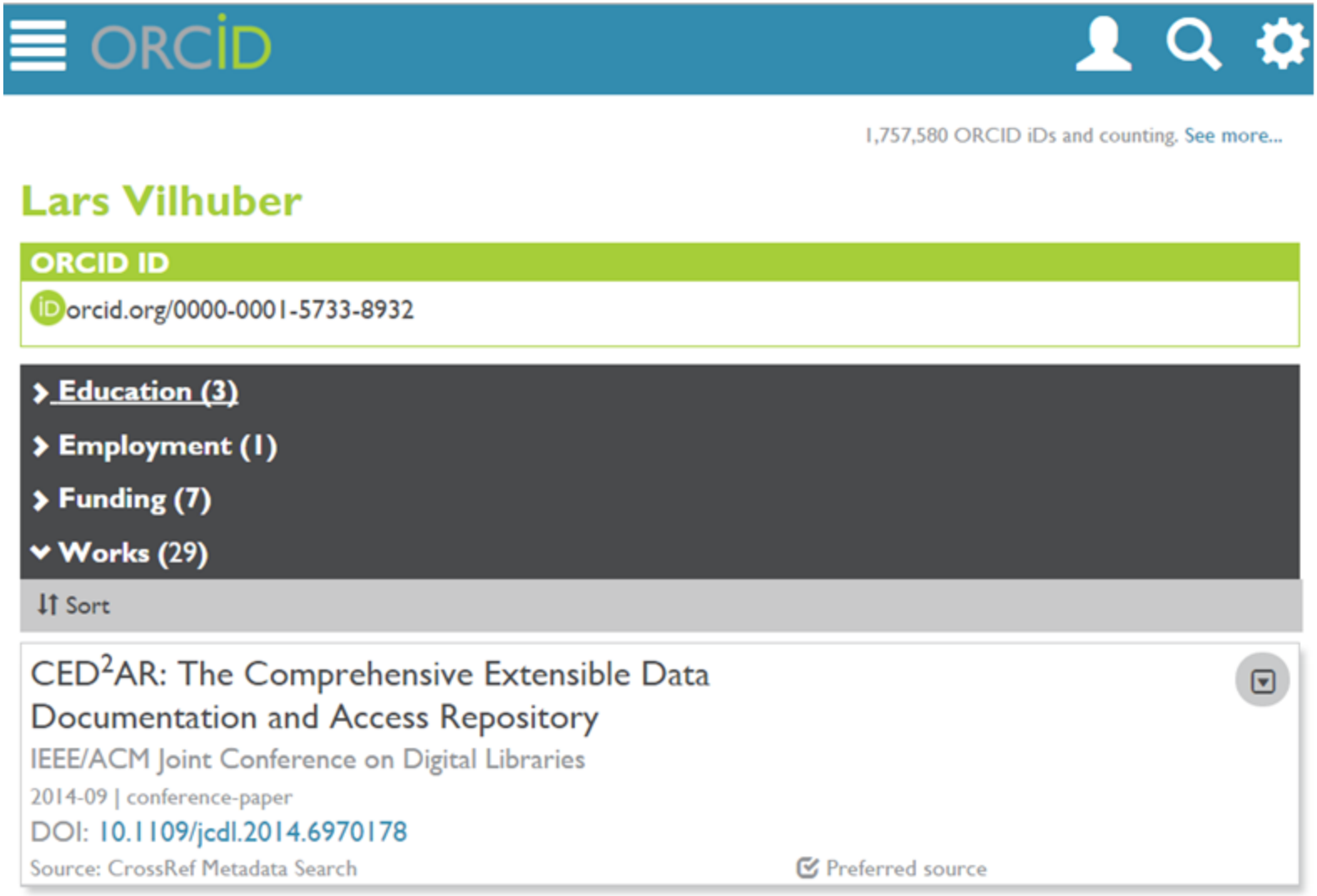
- Contributors can be tracked for each of their changes

CED2AR / SIPP Synthetic Beta v6.01 / Variable Versions

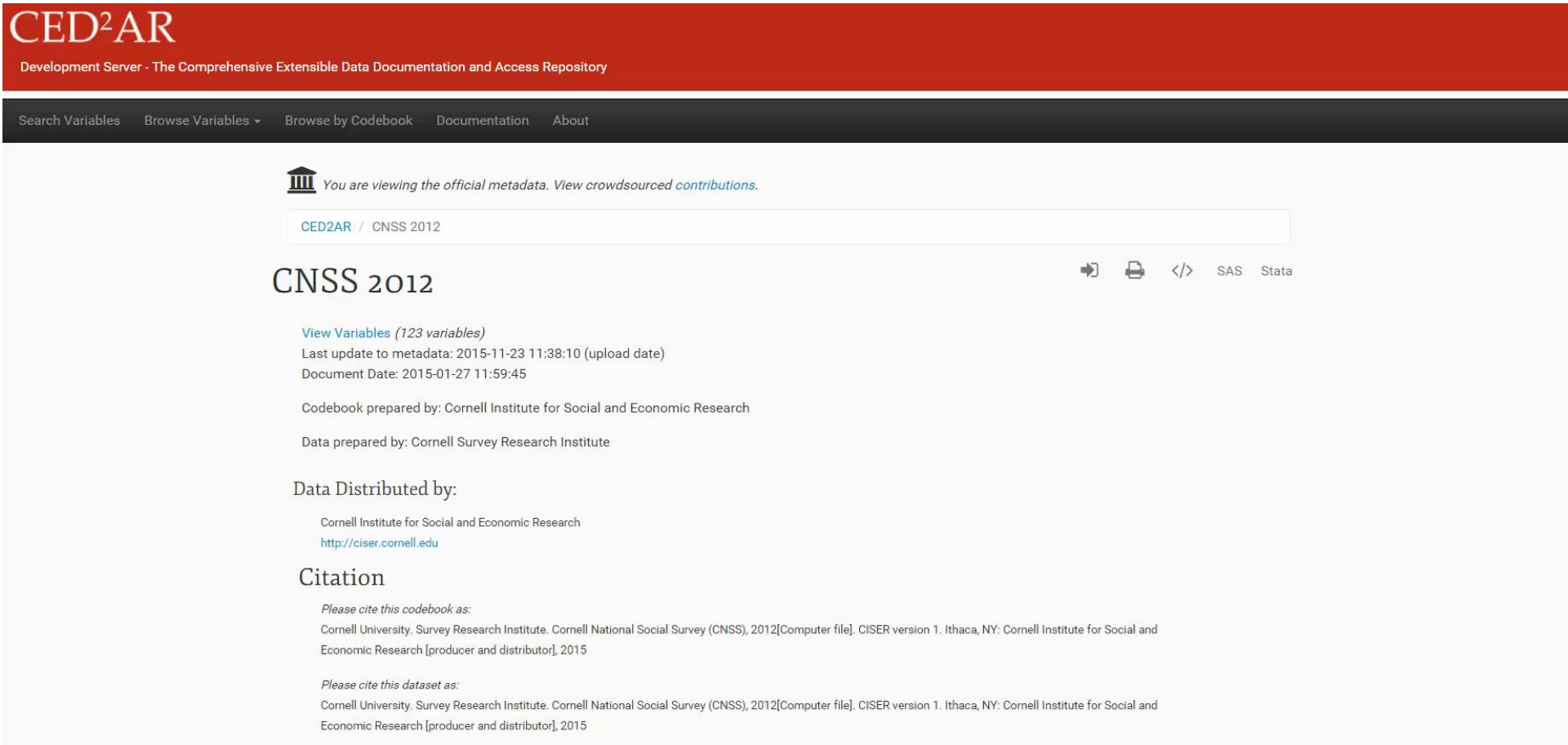
Modified Variables

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



Demo



Try for yourself: *http://demo.ncrn.cornell.edu*

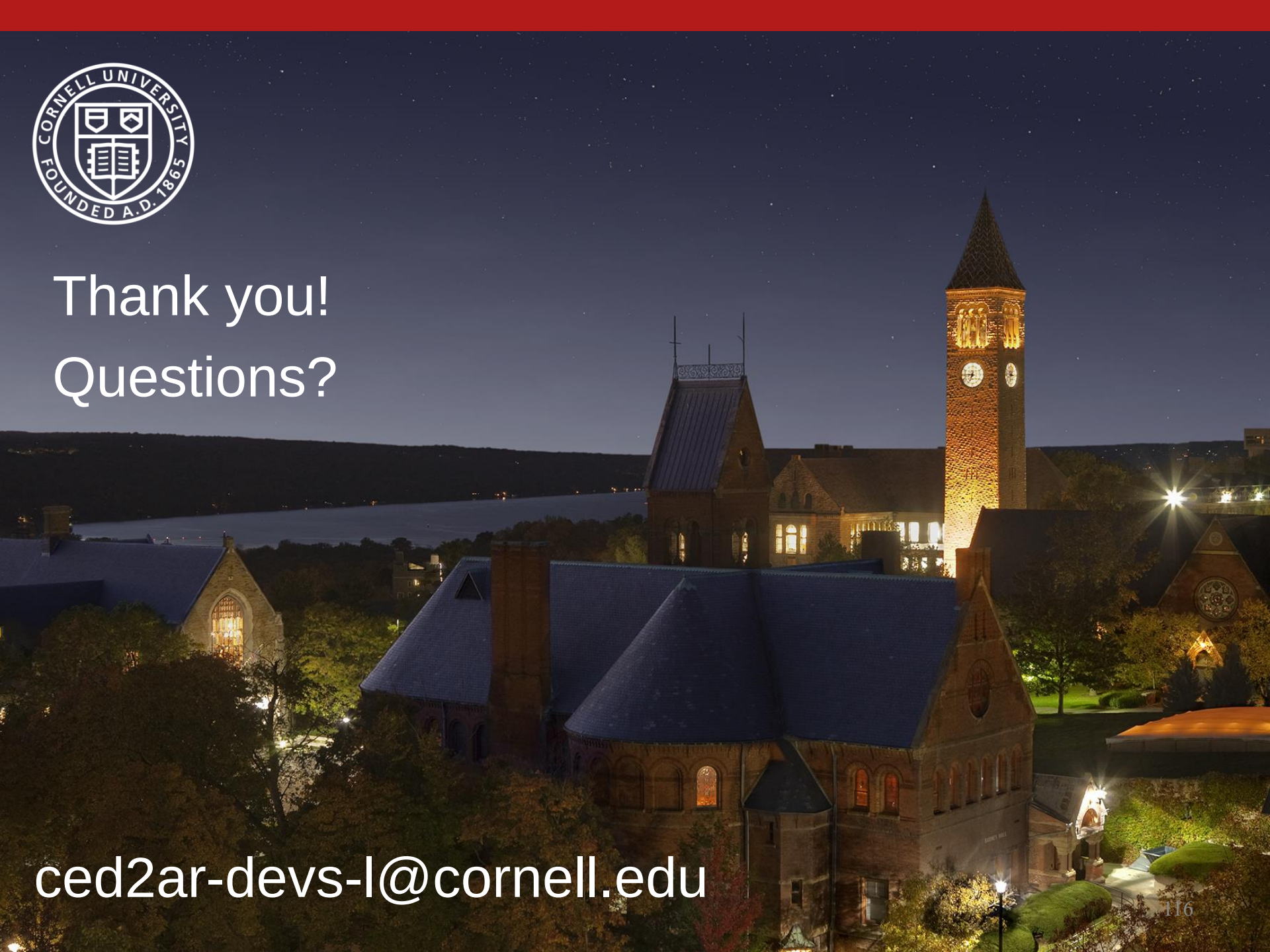
Where does this leave us?

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