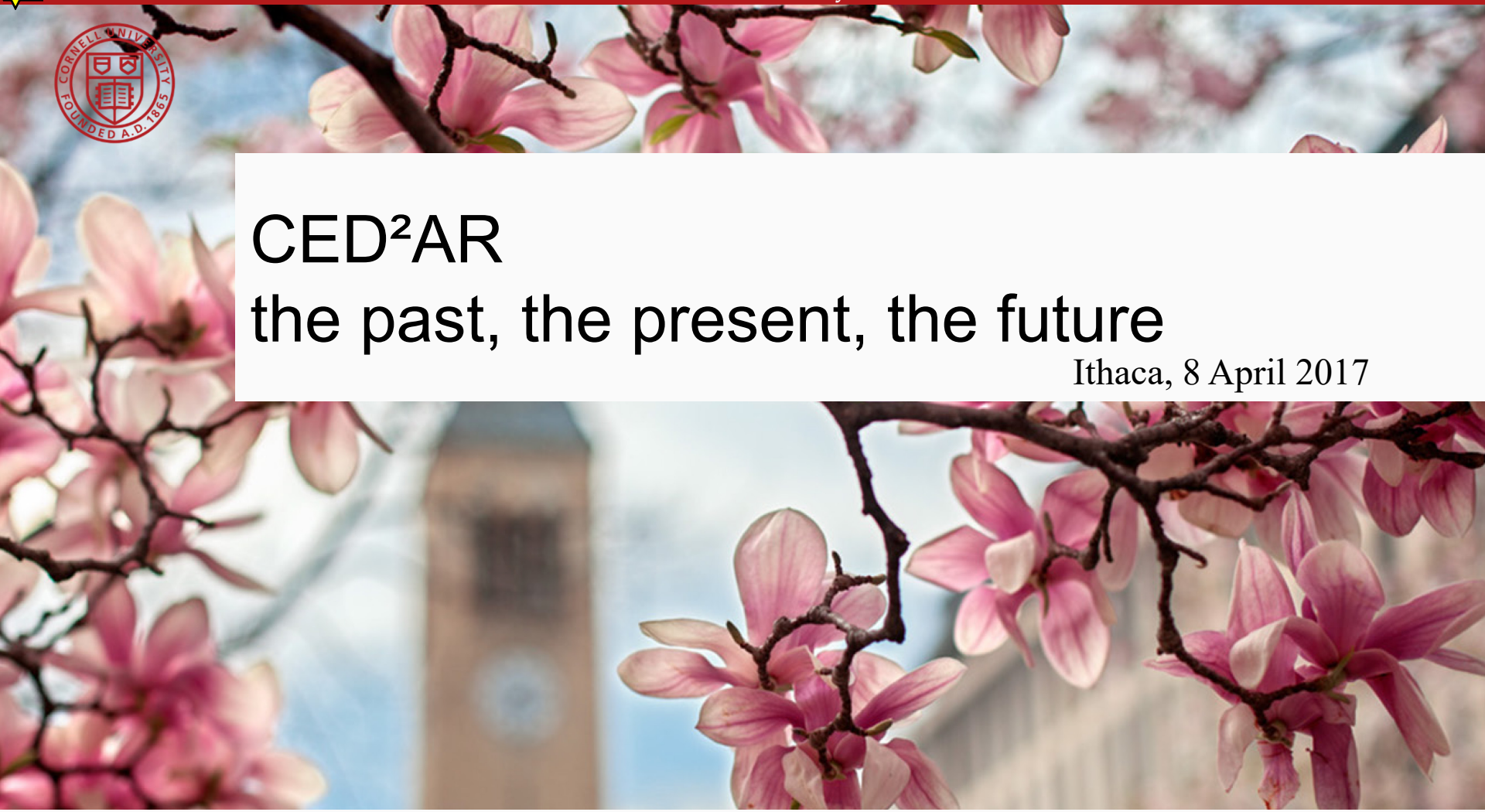
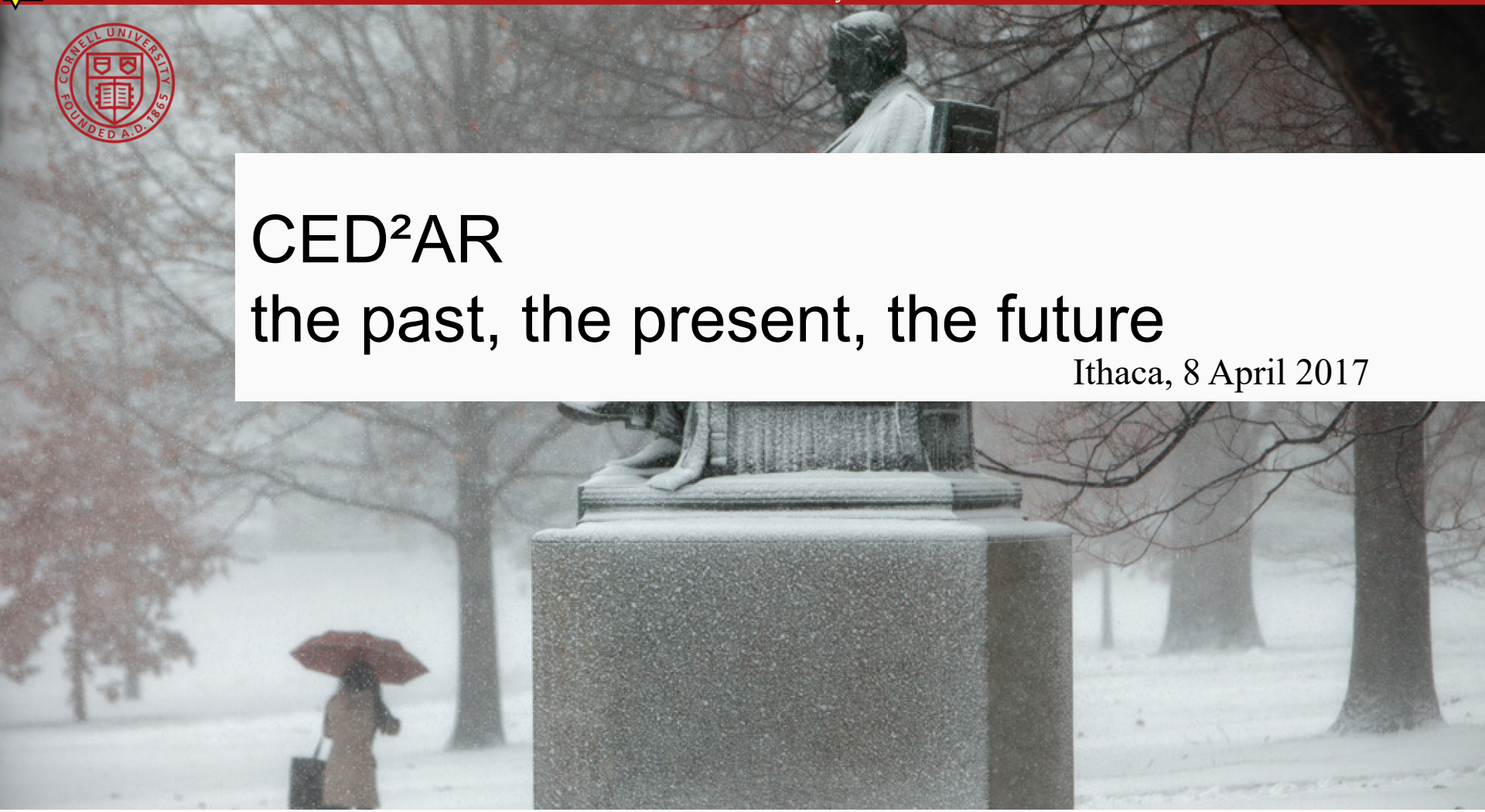


CED²AR
the past, the present, the future

Ithaca, 8 April 2017

Lars Vilhuber, Kyle Brumsted, Charles Simmer, Brandon Barker, William Block
Benjamin Perry, Venkata Kambhampaty,
(Cornell University)



CED²AR
the past, the present, the future

Ithaca, 8 April 2017

Lars Vilhuber, Kyle Brumsted, Charles Simmer, Brandon Barker, William Block
Benjamin Perry, Venkata Kambhampaty,
(Cornell University)

Acknowledgements

Other contributors include

- Jeremy Williams (Cornell University)
- Carl Lagoze (University of Michigan)
- John Abowd (Cornell University)
- Jared Lyle (ICPSR)
- Sanda Ionescu (ICPSR)
- Matthew Richardson (ICPSR)

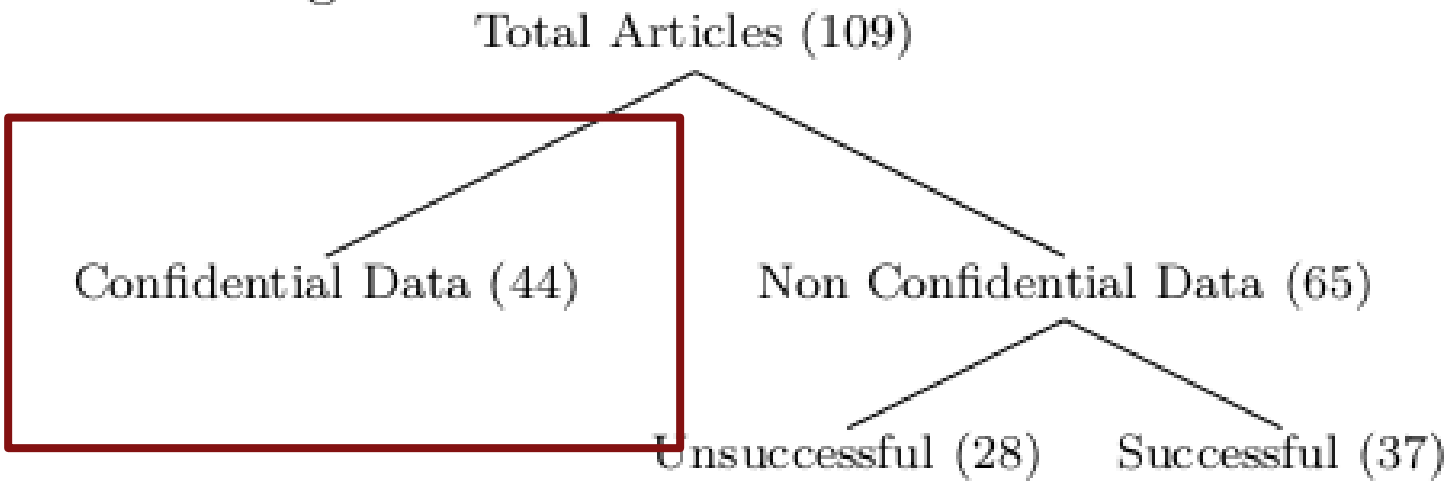
Funding by NSF Grant #1131848

Facilitating Reuse of Data

- [Online] existence of metadata record
- Permanent URL
- Availability of
  - Original data
  - Transformed data
  - Open availability
- Easy online inspection

But the biggest problem…

Figure 1: A Breakdown of the Articles

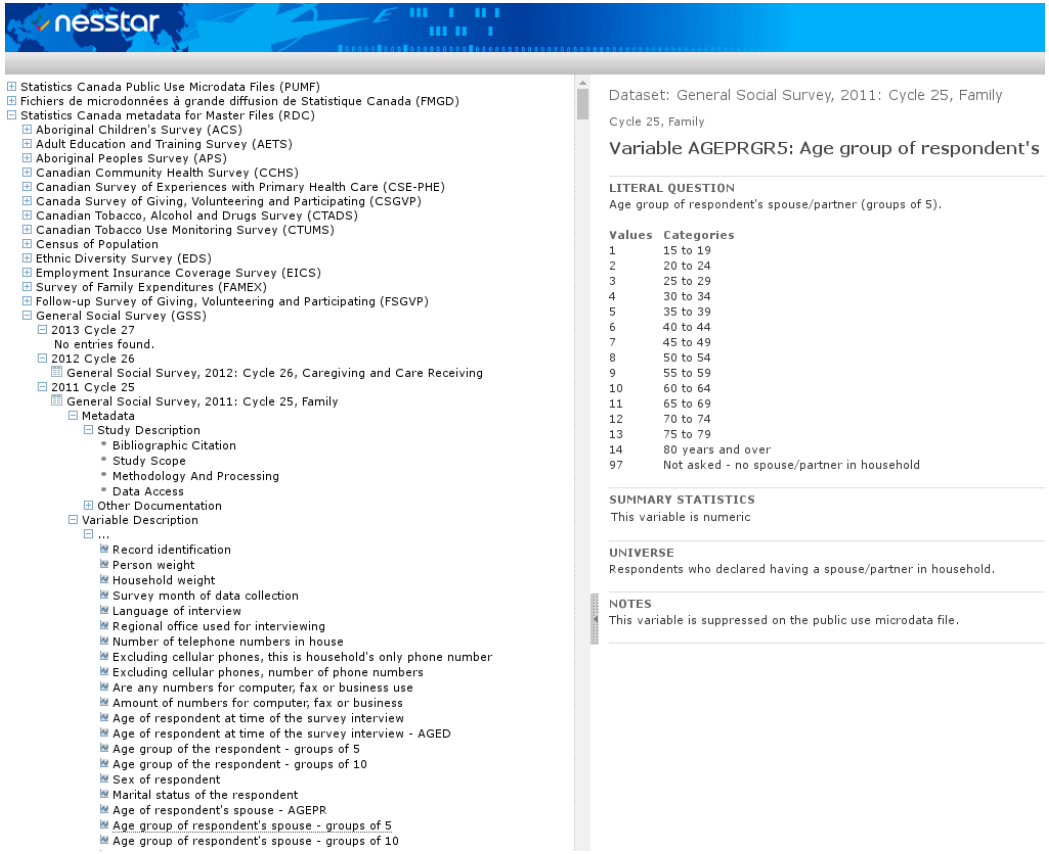


But: for confidential data…

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

Should We Just Trust These Guys?

Some are quite commendable



Our original goal

Facilitate documentation of confidential data

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?

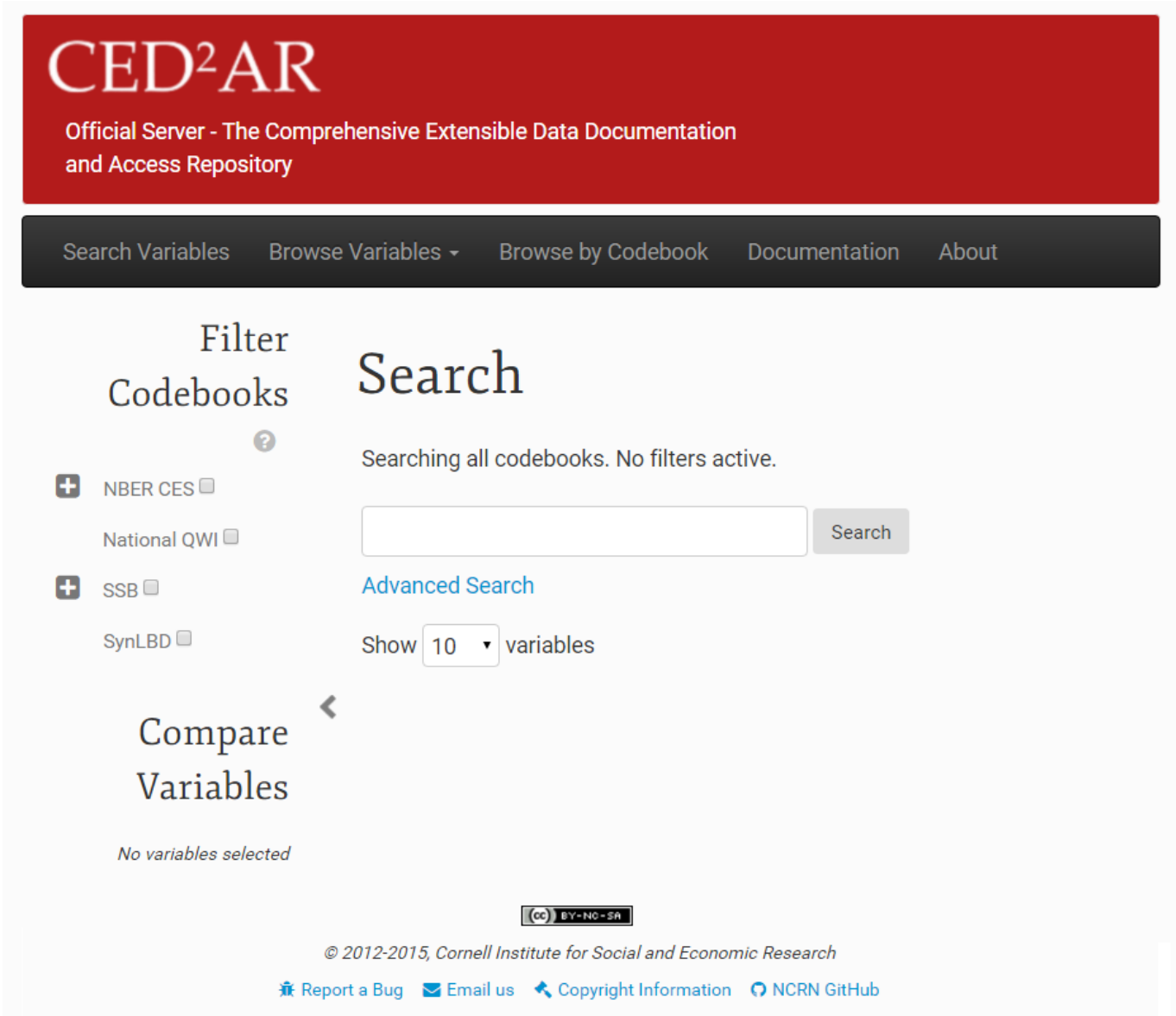
CED²AR

The Comprehensive Extensible Data Documentation and Access Repository

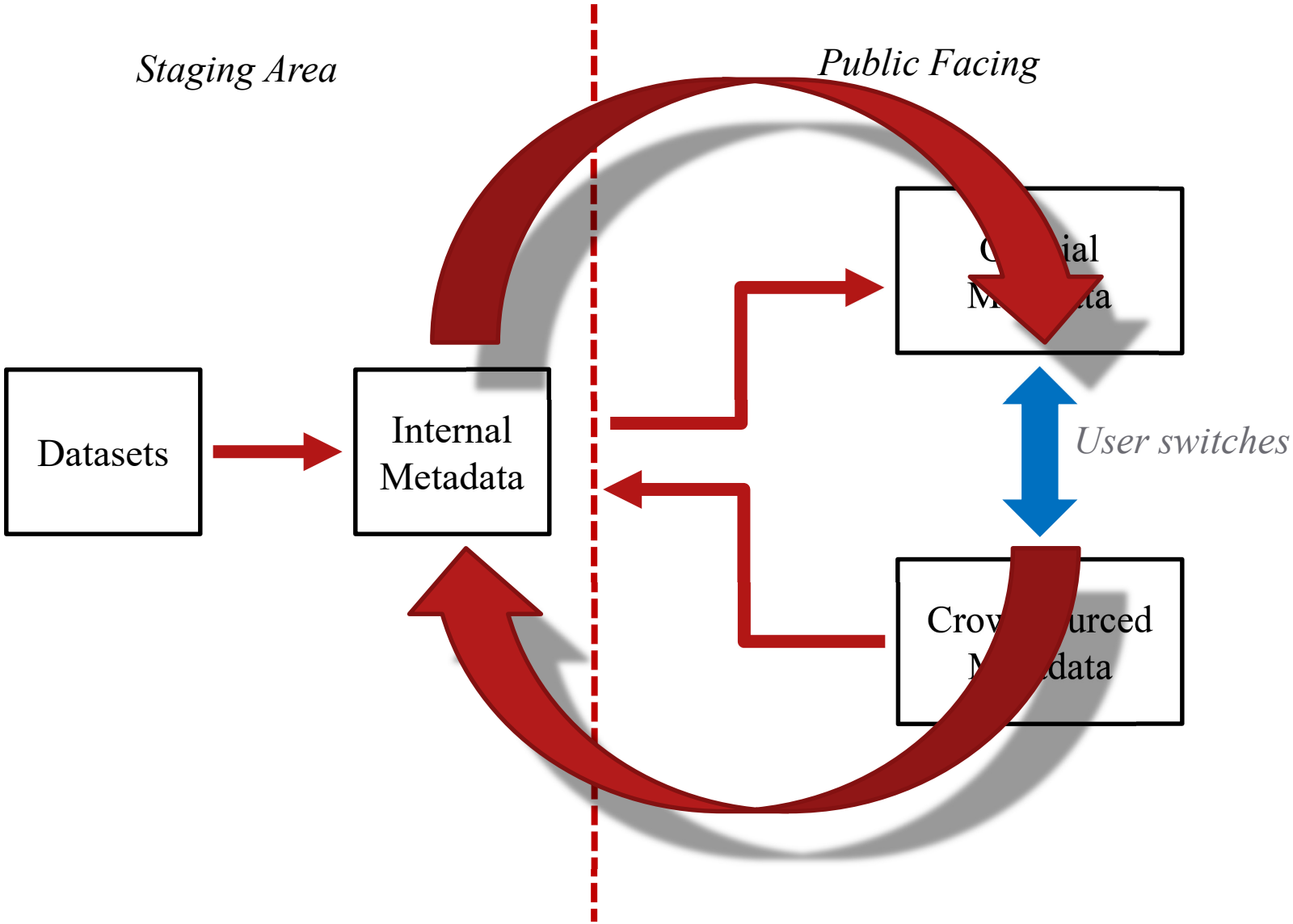
Progress towards our goals

- Facilitate documentation of confidential data
  - Proposed enhancement of DDI-C to allow for fine-grained access control to information (including variable names)
  - Software to leverage the standard
- Leverage researcher knowledge
  - Mechanism to allow for crowdsourced or collaborative editing of DDI-C metadata
- Metadata curation software
  - Web-based, extendable DDI editor
  - Designed for documenting existing datasets (not host data)
  - Allows for "ASCIImath" notation (in app)
- Online at www2.ncrn.cornell.edu/ced2ar-web

What is CED²AR?



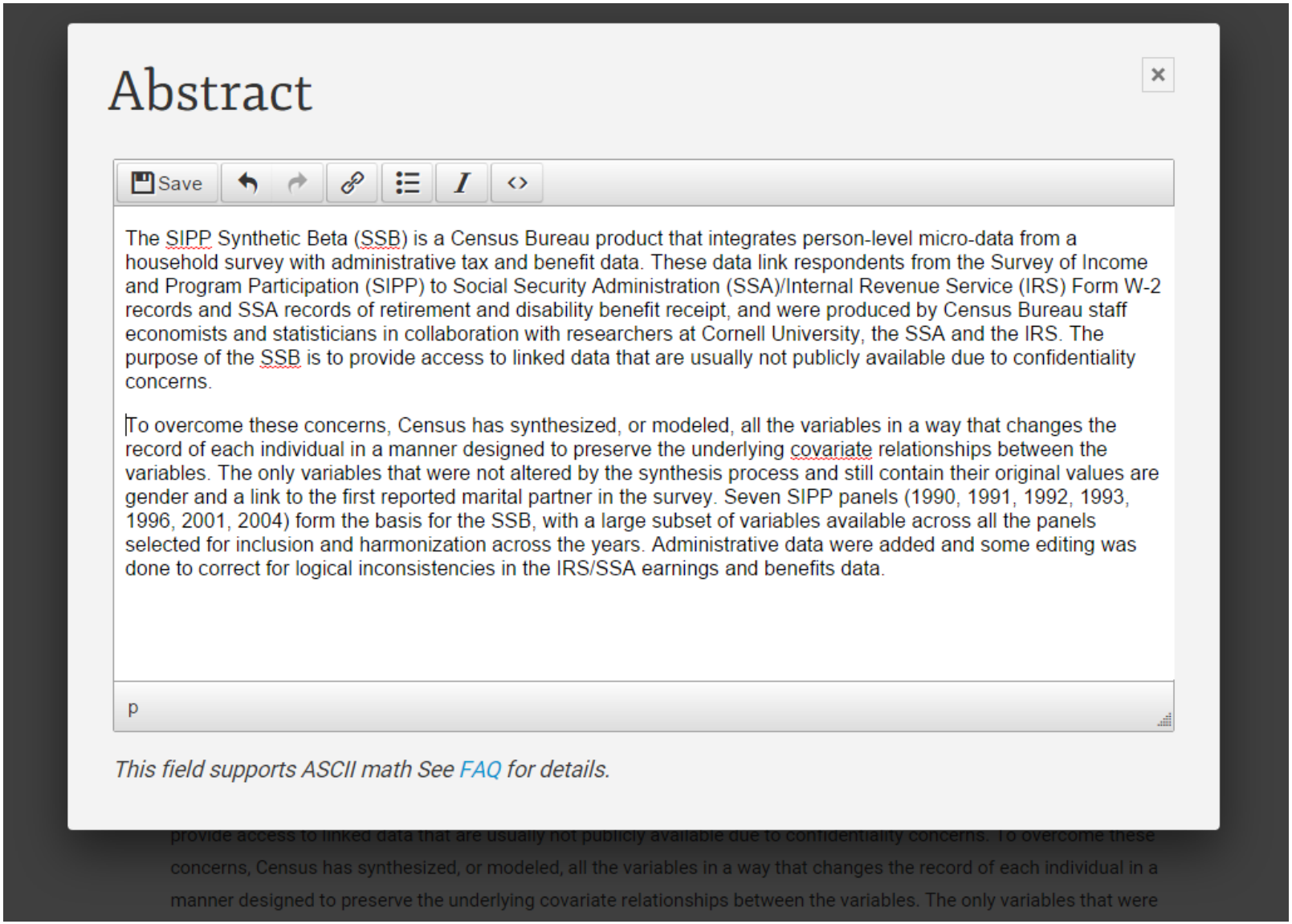
Basic Information Flow



Internal Processing

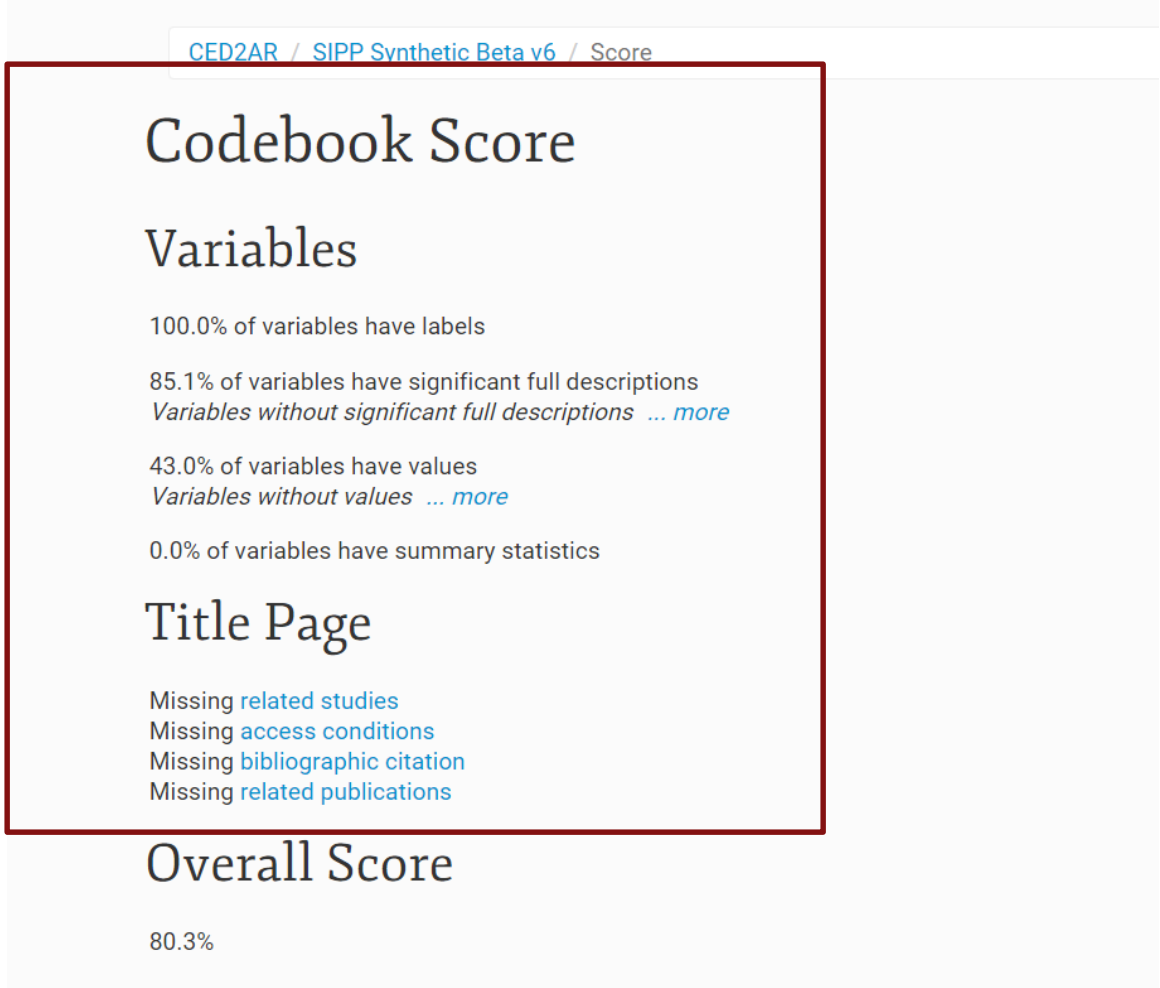
1. Creation of skeletal metadata
   - Assuming data is already curated
   - Converting data into standardized metadata
     - Tools included (for SAS, Stata, SPSS, CSV), not discussed here
2. Hand editing and subsetting
   - Adding verbose descriptions
   - Applying disclosure limitation
3. User accessible
   - These tools can be manipulated by normal users
   - They could be incorporated into existing workflows

Internal Processing: Hand Editing



Internal Processing: Scoring

- Provide feedback to improve sparse documentation

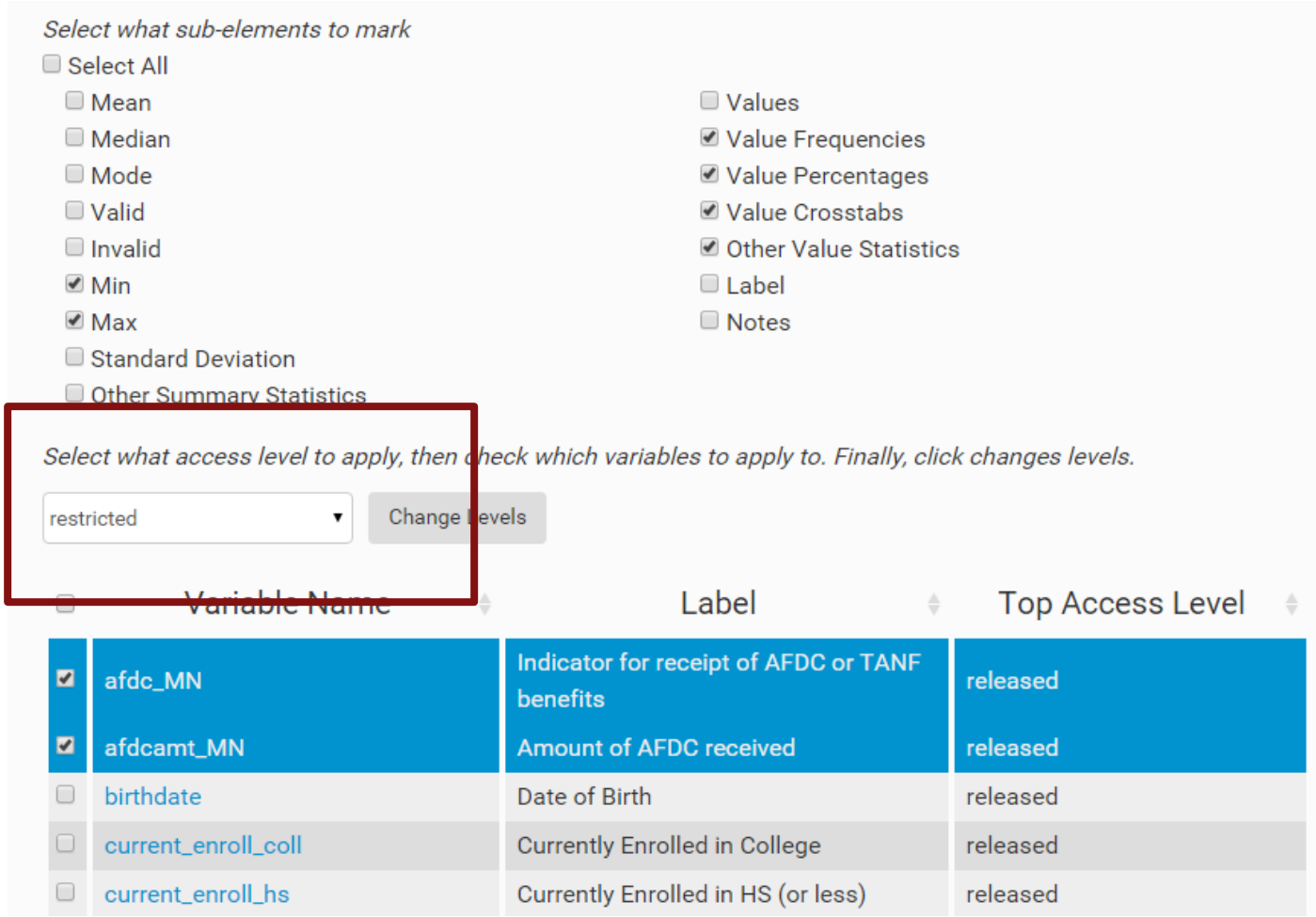


Fine-grained access controls

Important when working with confidential (meta)data

Internal Processing: Access Control

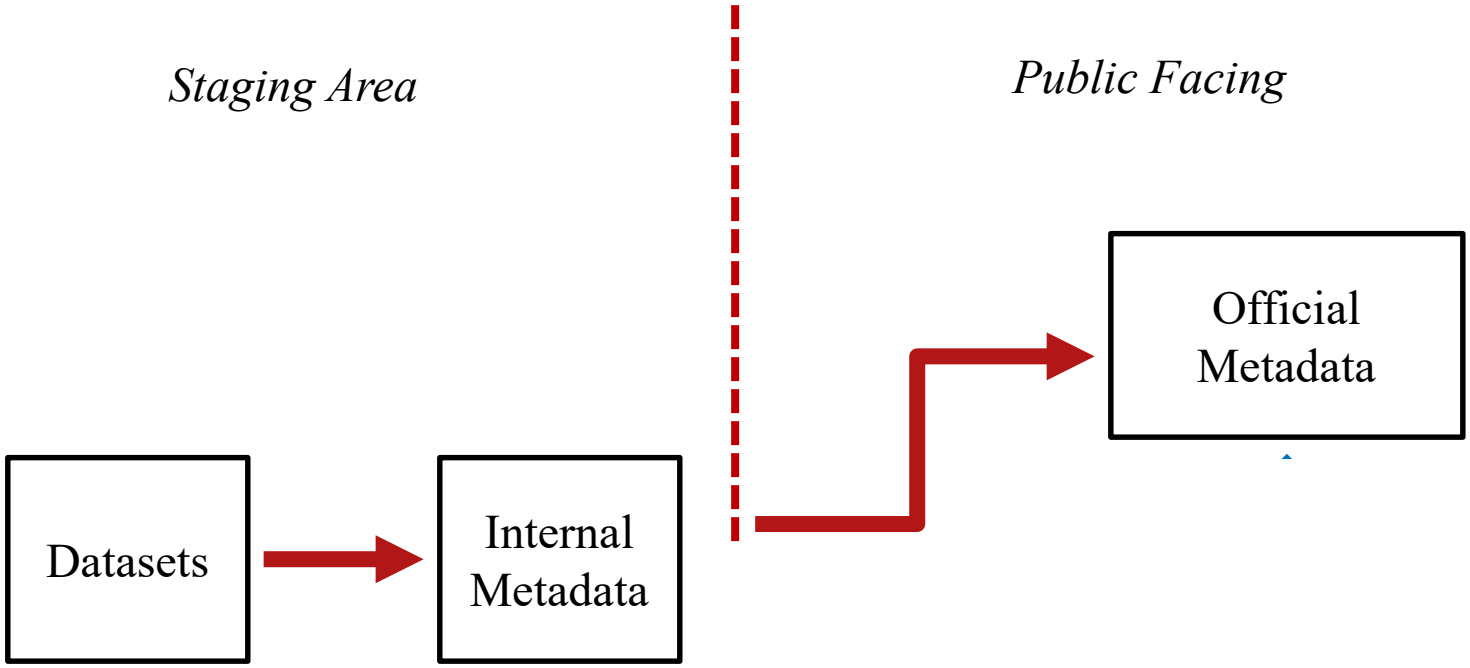
- Marking elements with different restrictions



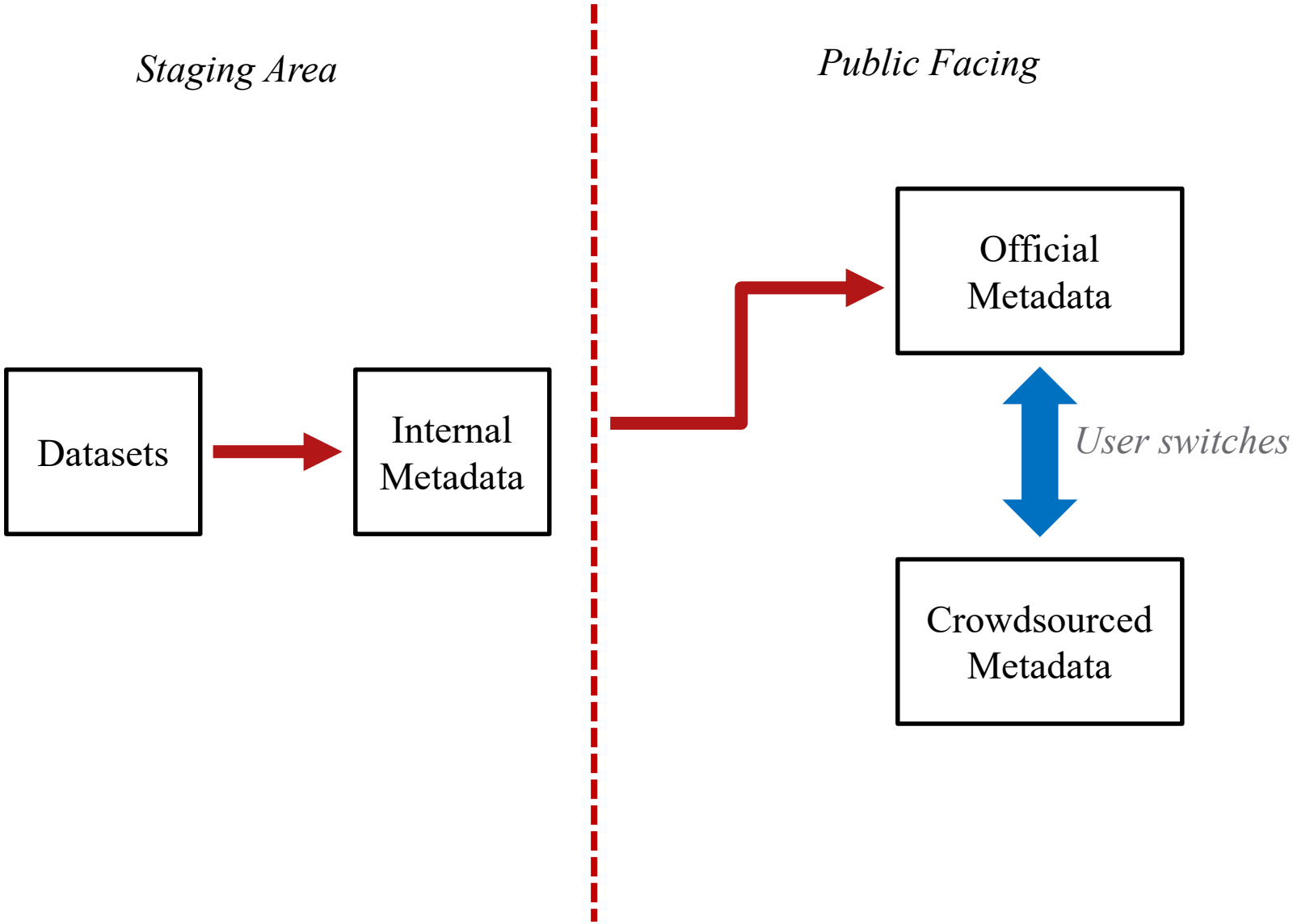
Workflow control

- Ability to view additions/subtractions
  - Between versions
  - Between crowd-sourced information and official information
- Ability to control access
  - Editing versus viewing
  - Authentication and reputation

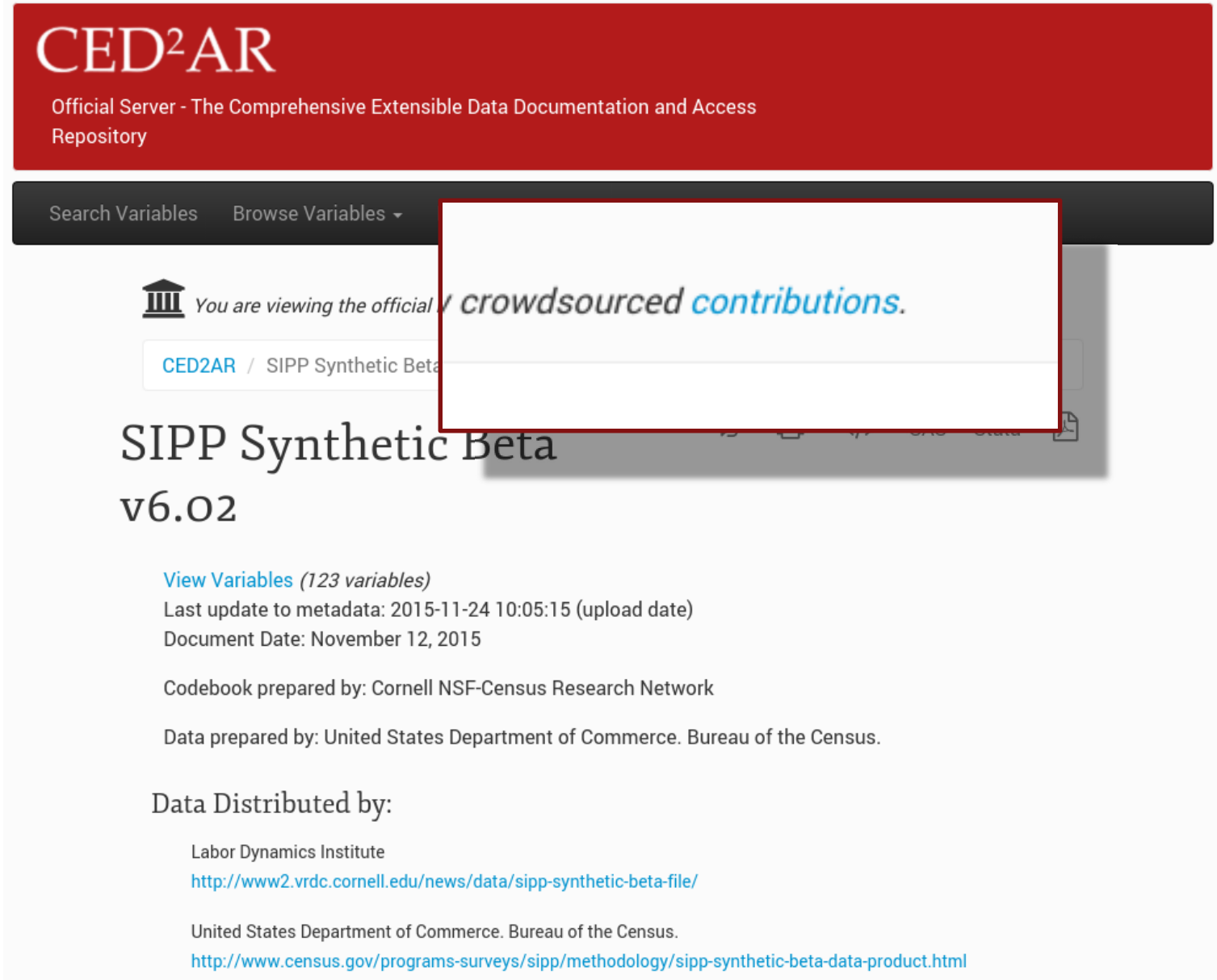
Basic Information Flow



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

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)

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.

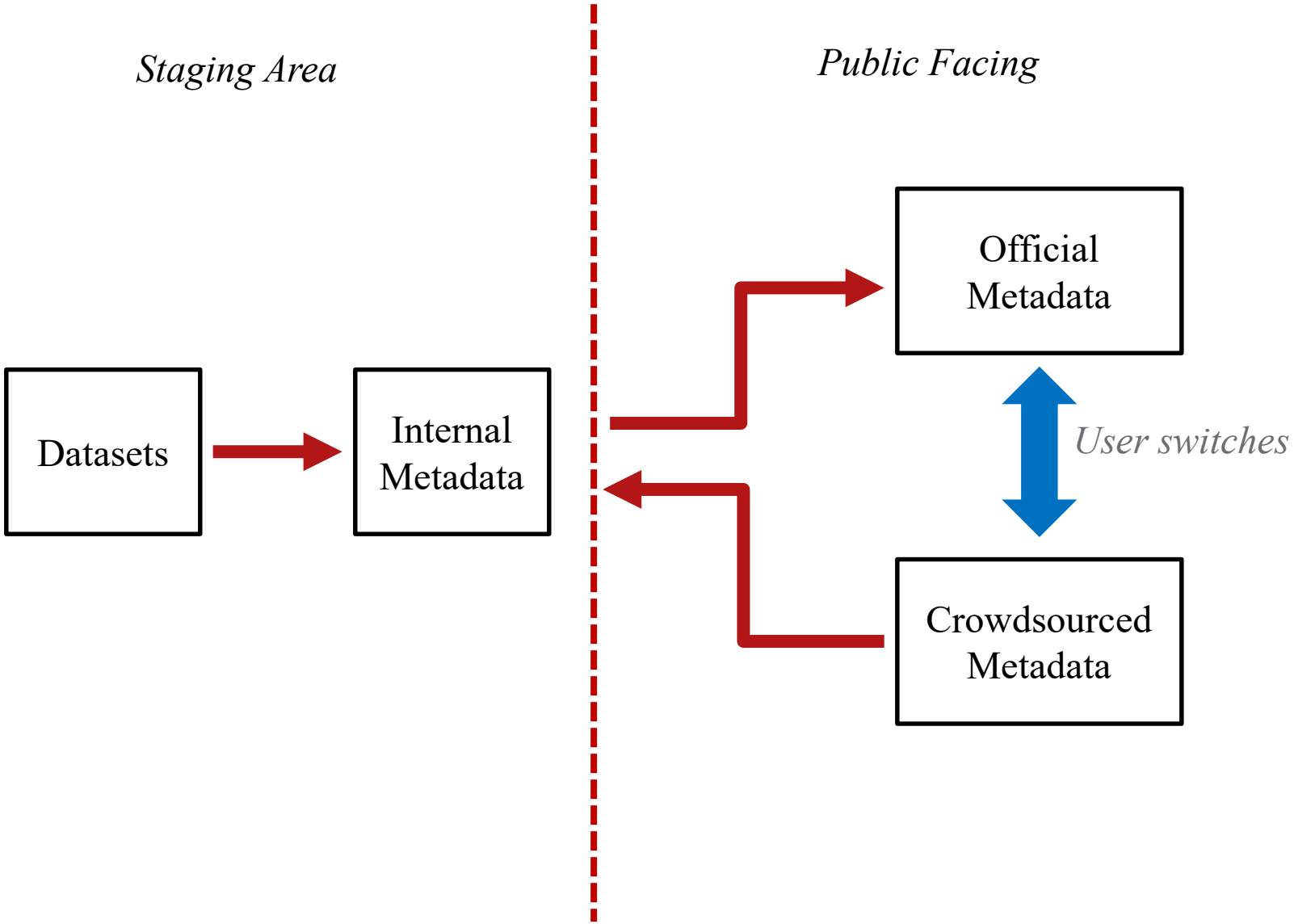
If you are looking at earnings in earlier years, particular the 1960s and earlier, there will be more people with $0 earnings because many jobs were not FICA-taxable then. Even today, there are some instances of legitimate non-FICA earnings that would not be reflected on the SER. One example of this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

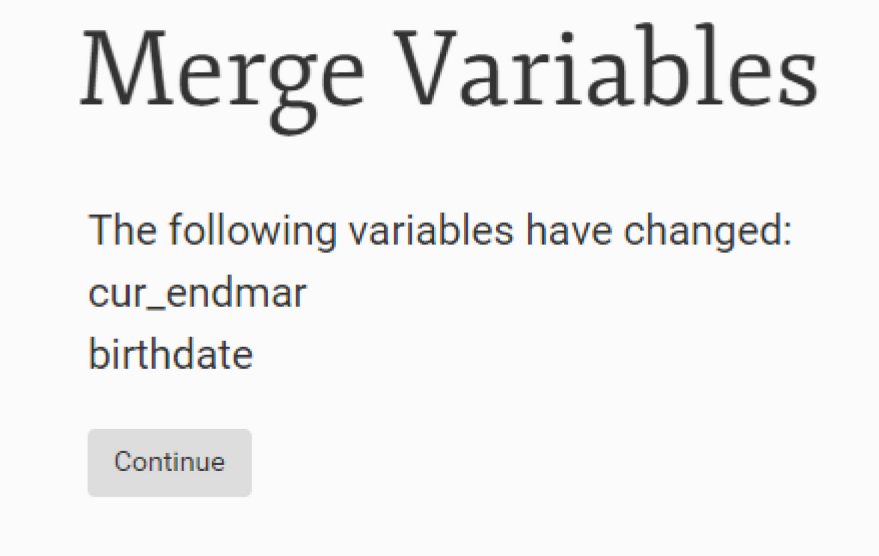
Notes (1 total)

#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*

That's the theory…

What works well

- Very portable application (self-sufficient package, deployable as desktop app without major programming experience)
- Efficient editor for non-technical users (many people like it)
- Quite efficient in ingesting large DDI files (tested by ICPSR with large examples)
- Native cross-codebook variable search and variable comparison

What doesn't work so well

- Stuck at DDI-C
- Configuration cL~~u~~*n*kY – in particular for the three-way setup
- Scalability an issue
  - Many codebooks
  - User interface not optimized
- Bugs
  - Search
  - UI issues
  - Local versioning (workgroup scenario)

Implementations and Testing

In active use

- Two data products by the Census Bureau
- A few demo public-use data products
  - Common to all: we do not host the data, only the metadata

Kicking the tires

- Census Bureau
  - Collaboration with CARRA as an outside implementation
  - Discussions with SEHSD for internal implementation (AHS, SIPP)
  - Discussions with CES for (limited) internal implementation (RDC)

Kicking the tires

- ICPSR/ openICPSR
  - Replacement for internal DDI editor?
  - Use for openICPSR (for self-serve metadata creation)
- mTBI project (Block)

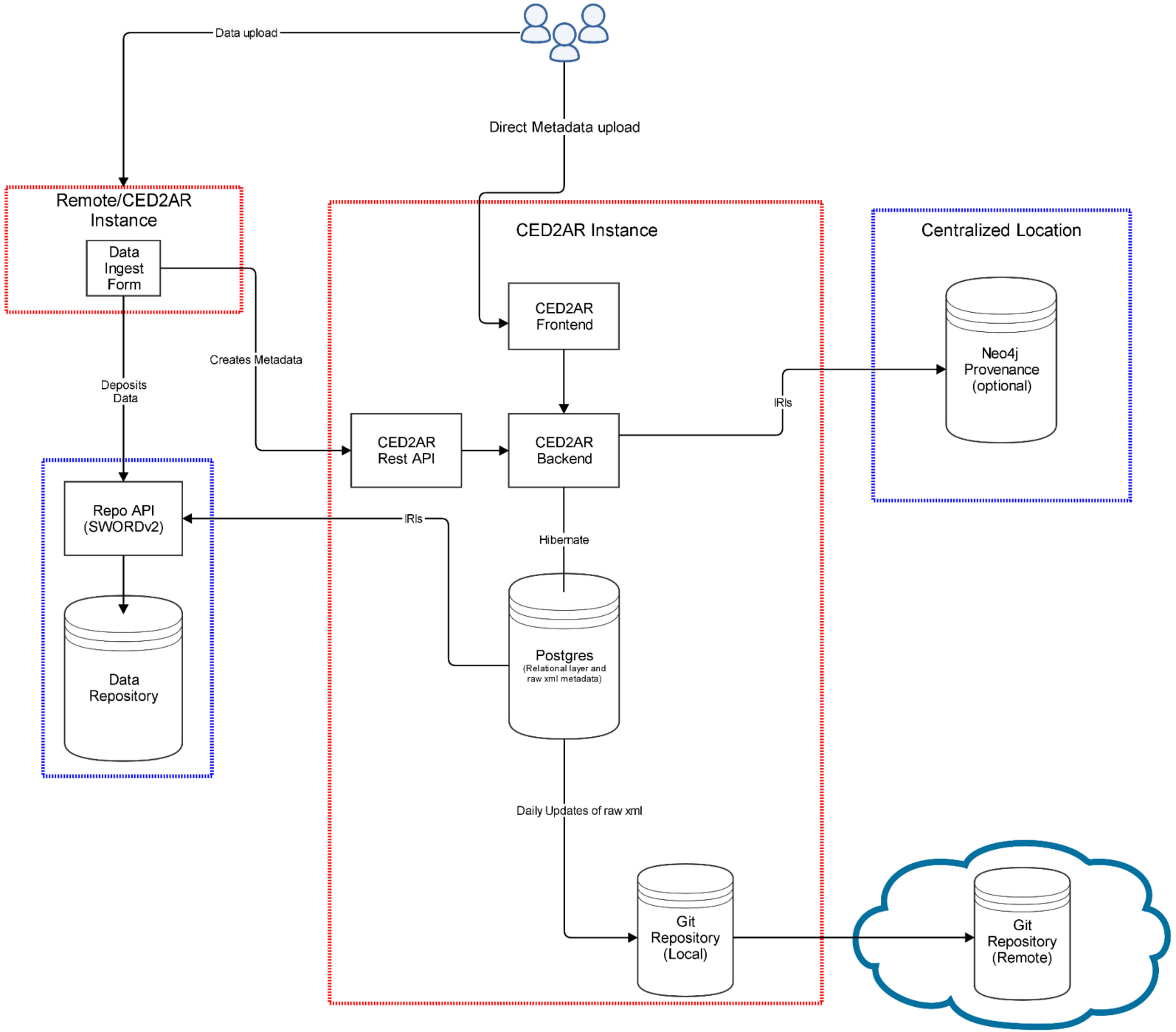
Next steps

Making CED²AR v2 robust

- Addition of UTF-8 editing support (Spanish, French, Portuguese, etc.)
- Additional fields (link to survey questions, anything within DDI-C)
- Bug fixes (search, UI, versioning)

CED²AR V3 Roadmap

- Move to relational database – more flexibility wrt schemas (DDI-L, others)
- More flexible architecture
  - standalone modules
  - Single page application (SPA) for frontend
- Redesigned (simpler) VCS


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


Thank you!
Questions?
ced2ar-pub-dev-l@cornell.edu