



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)*



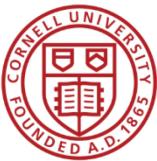
CED²AR

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)

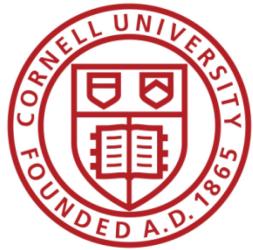
Funding by NSF Grant #1131848





Facilitating Reuse of Data

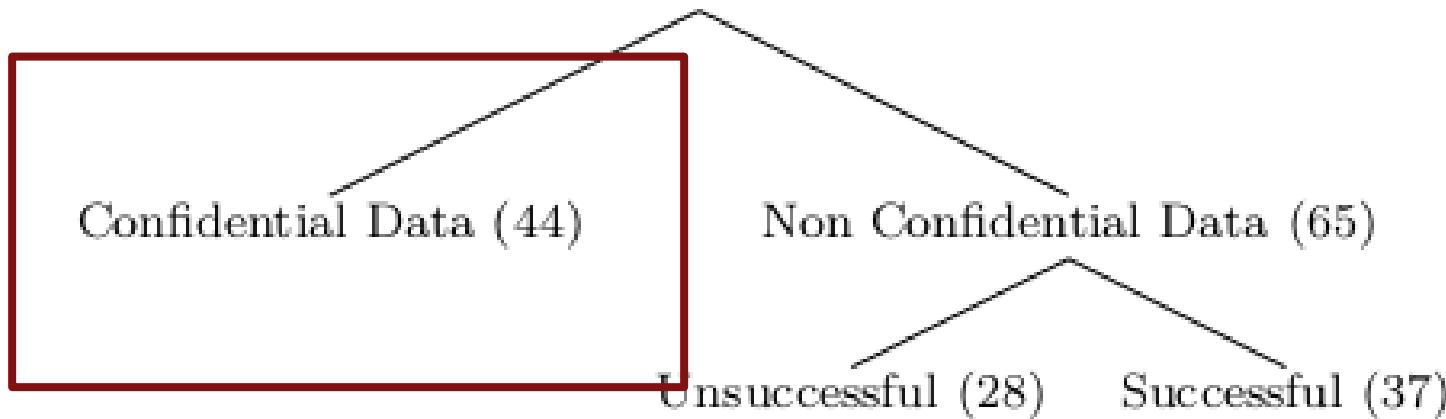
- [Online] existence of metadata record
- Permanent URL
- Availability of
 - Original data
 - Transformed data
 - Open availability
- Easy online inspection



But the biggest problem...

Figure 1: A Breakdown of the Articles

Total Articles (109)

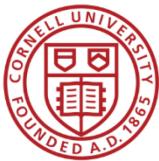


But: for confidential data...

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

Should We Just Trust These Guys?





Some are quite commendable

 nessstar

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)
 Canadian 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 spouse/partner (groups of 5).

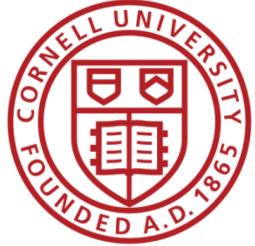
LITERAL QUESTION
 Age group of respondent's spouse/partner (groups of 5).

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.



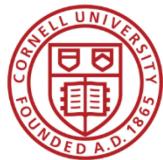
Our original goal

Facilitate documentation of confidential data
Leverage researcher knowledge



Our Approach

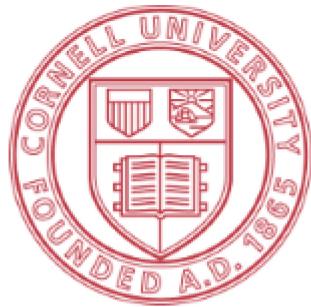
- Rely on open standards, namely the Data Documentation Initiative (DDI) schema
- Provide easy-to-use tools and interfaces to structured metadata
- Build infrastructure that enables data curators to leverage community-driven input to official documentation



How?

CED²AR

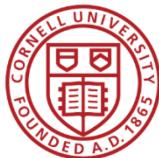
The Comprehensive Extensible Data Documentation and Access Repository





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?

CED²AR

Official Server - The Comprehensive Extensible Data Documentation
and Access Repository

[Search Variables](#)[Browse Variables ▾](#)[Browse by Codebook](#)[Documentation](#)[About](#)

Filter
Codebooks

Search



Searching all codebooks. No filters active.

NBER CES

National QWI

SSB

[Advanced Search](#)

SynLBD

Show variables

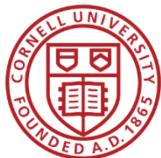
Compare
Variables

No variables selected

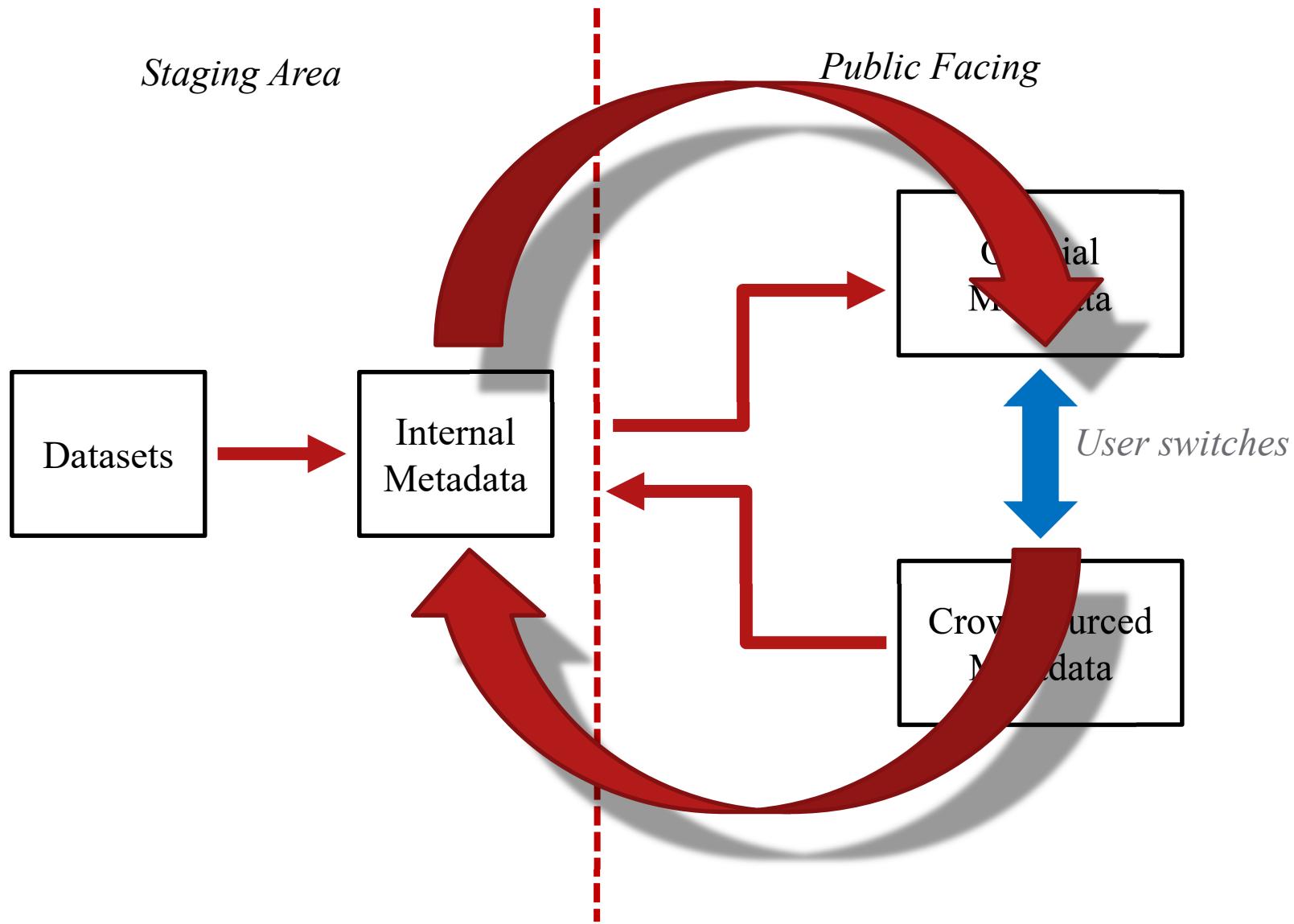


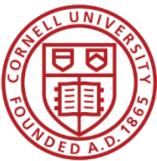
© 2012-2015, Cornell Institute for Social and Economic Research

[Report a Bug](#)[Email us](#)[Copyright Information](#)[NCRN GitHub](#)



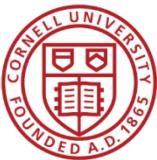
Basic Information Flow





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

Abstract

Save

The SIPP Synthetic Beta (SSB) is a Census Bureau product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the Survey of Income and Program Participation (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Form W-2 records and SSA records of retirement and disability benefit receipt, and were produced by Census Bureau staff economists and statisticians in collaboration with researchers at Cornell University, the SSA and the IRS. The purpose of the SSB is to provide access to linked data that are usually not publicly available due to confidentiality concerns.

To overcome these concerns, Census has synthesized, or modeled, all the variables in a way that changes the record of each individual in a manner designed to preserve the underlying covariate relationships between the variables. The only variables that were not altered by the synthesis process and still contain their original values are gender and a link to the first reported marital partner in the survey. Seven SIPP panels (1990, 1991, 1992, 1993, 1996, 2001, 2004) form the basis for the SSB, with a large subset of variables available across all the panels selected for inclusion and harmonization across the years. Administrative data were added and some editing was done to correct for logical inconsistencies in the IRS/SSA earnings and benefits data.

p

This field supports ASCII math See [FAQ](#) for details.

provide access to linked data that are usually not publicly available due to confidentiality concerns. To overcome these concerns, Census has synthesized, or modeled, all the variables in a way that changes the record of each individual in a manner designed to preserve the underlying covariate relationships between the variables. The only variables that were



Internal Processing: 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

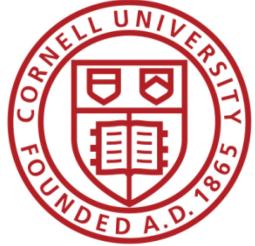
0.0% of variables have summary statistics

Title Page

Missing related studies
Missing access conditions
Missing bibliographic citation
Missing related publications

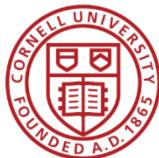
Overall Score

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.

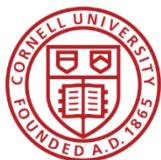
restricted

<input type="checkbox"/> Variable Name	Label	Top Access Level
<input checked="" type="checkbox"/> afdc_MN	Indicator for receipt of AFDC or TANF benefits	released
<input checked="" type="checkbox"/> afdcamt_MN	Amount of AFDC received	released
<input type="checkbox"/> birthdate	Date of Birth	released
<input type="checkbox"/> current_enroll_coll	Currently Enrolled in College	released
<input type="checkbox"/> current_enroll_hs	Currently Enrolled in HS (or less)	released

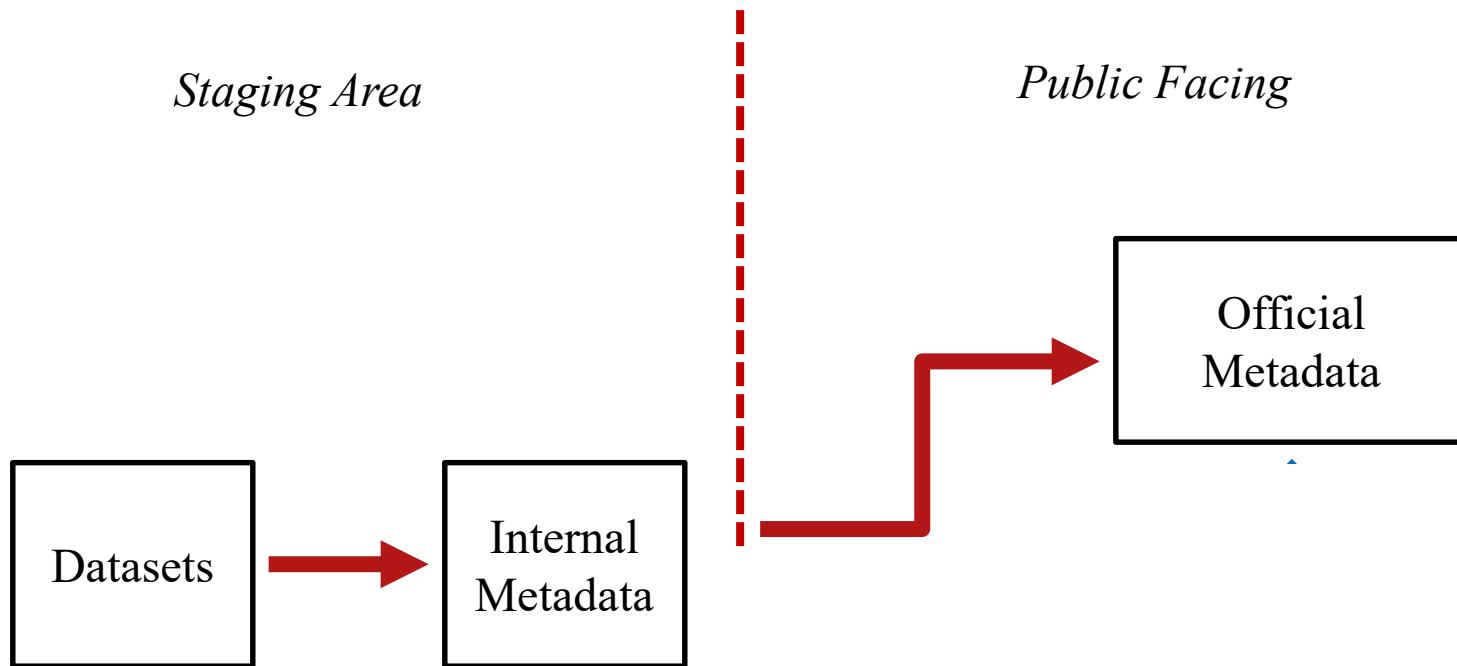


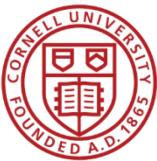
Workflow control

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

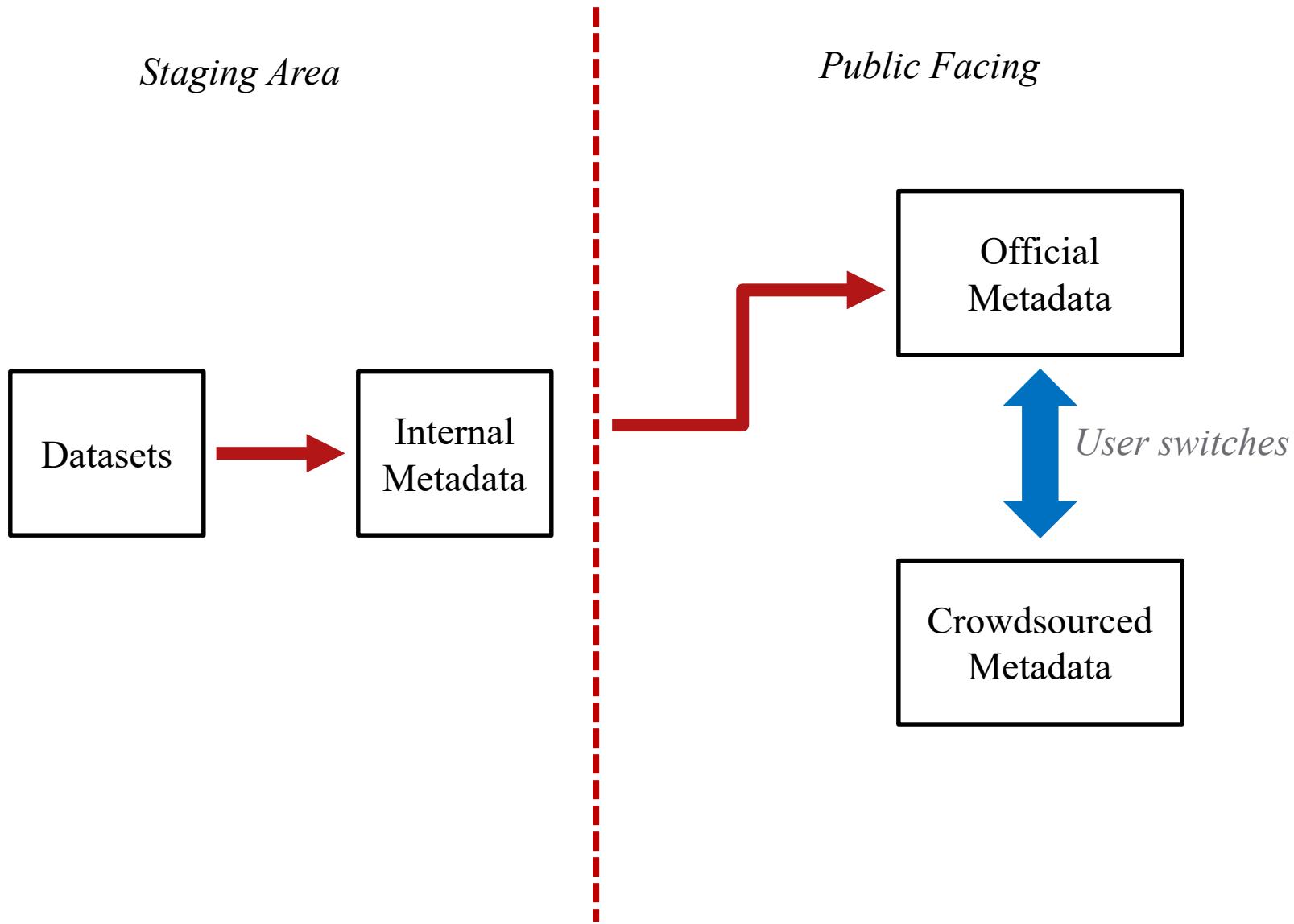


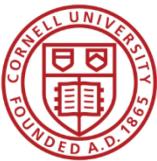
Basic Information Flow





Basic Information Flow





Official view

CED²AR

Official Server - The Comprehensive Extensible Data Documentation and Access Repository

Search Variables Browse Variables ▾



You are viewing the official / *crowdsourced contributions*.

CED2AR

/ SIPP Synthetic Beta

SIPP Synthetic Beta v6.02

[View Variables](#) (123 variables)

Last update to metadata: 2015-11-24 10:05:15 (upload date)

Document Date: November 12, 2015

Codebook prepared by: Cornell NSF-Census Research Network

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

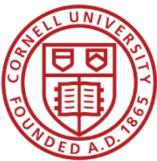
Data Distributed by:

Labor Dynamics Institute

<http://www2.vrdc.cornell.edu/news/data/sipp-synthetic-beta-file/>

United States Department of Commerce. Bureau of the Census.

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



Crowdsourced view

CED²AR

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



You are viewing crowdsourced metadata. View the [official version](#).

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.

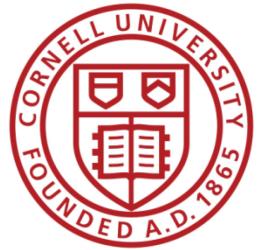
Data Distributed by:

Labor Dynamics Institute

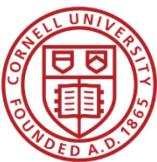
<http://www2.vrdc.cornell.edu/news/data/sipp-synthetic-beta-file/>

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> 



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  

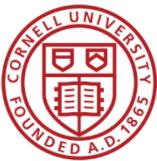
Person-level annual earnings that were taxed by FICA; these variables include earnings only up to the FICA taxable maximum and cover the years 1951-2011.

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)



Type numeric

Notes



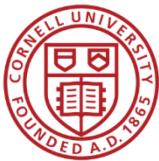
Having a 0 value on totearn_ser_yyyy could mean a couple of things: 1) this individual had no FICA-covered earnings in that year; 2) this individual had no labor income at all in this tax year; 3) this individual worked for an employer that failed to report earnings in this year (that is to say, this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

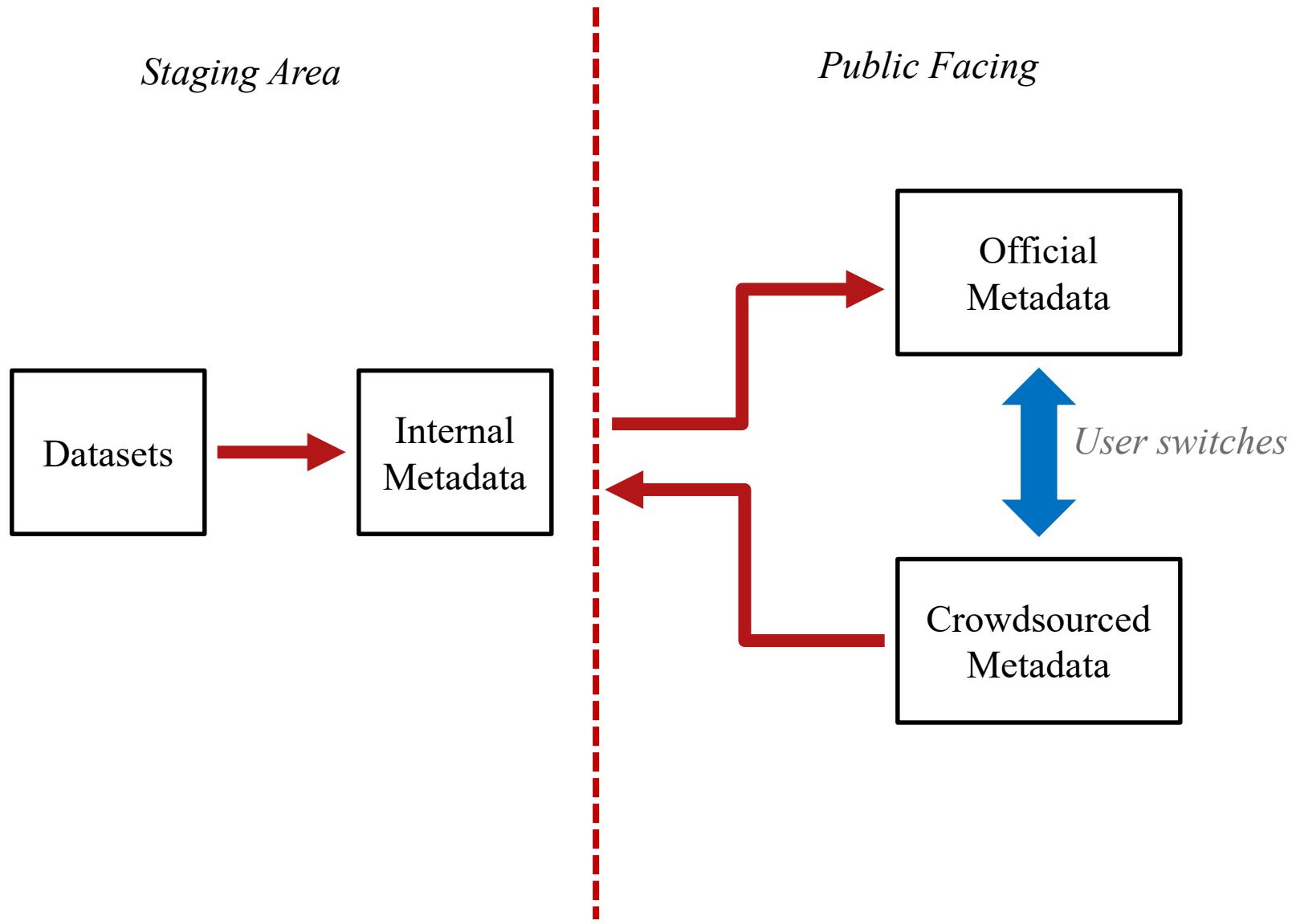
p

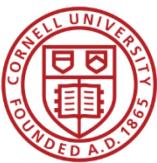
This field supports ASCII math. See [FAQ](#) for details.

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

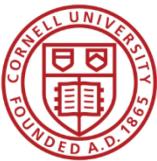
Full Description

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

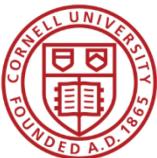
- Curators are given an interface to merge crowdsourced documentation with official

Merge Variables

The following variables have changed:

cur_endmar
birthdate

[Continue](#)



Combining Knowledge: Merging

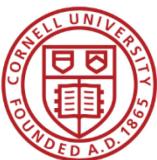
current_enroll_coll

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	<input type="checkbox"/> Use crowdsourced <input type="checkbox"/> Use original Currently Enrolled
Codebook	SIPP Synthetic Beta v6
Concept	
Concept Vocabulary	
Concept Vocabulary URI	
Type	numeric
Files	



Combining Knowledge: Merging

Crowdsourced Documentation

Official 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

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

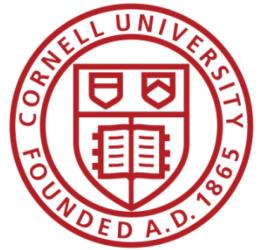
The SIPP Synthetic Beta (SSB) is a Census Bureau product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the Survey of Income and Program Participation (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Form W-2 records and SSA



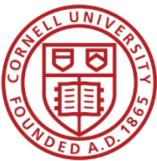
Combining Knowledge: Citations

- Contributors can be tracked for each of their changes

Modified Variables				
<u>Variable Name</u>	<u>Date Changed</u>	<u>Commit Message</u>	<u>User</u>	<u>Origin</u>
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...



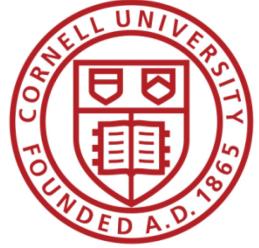
What works well

- **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 cLunkY – 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



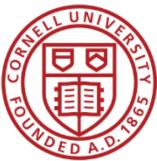
In active use

- 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



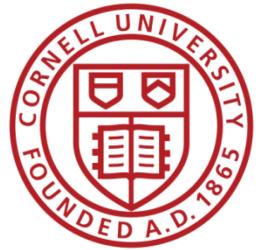
Kicking the tires

- 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)

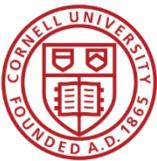


Kicking the tires

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

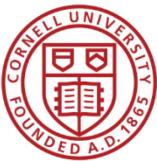


Next steps



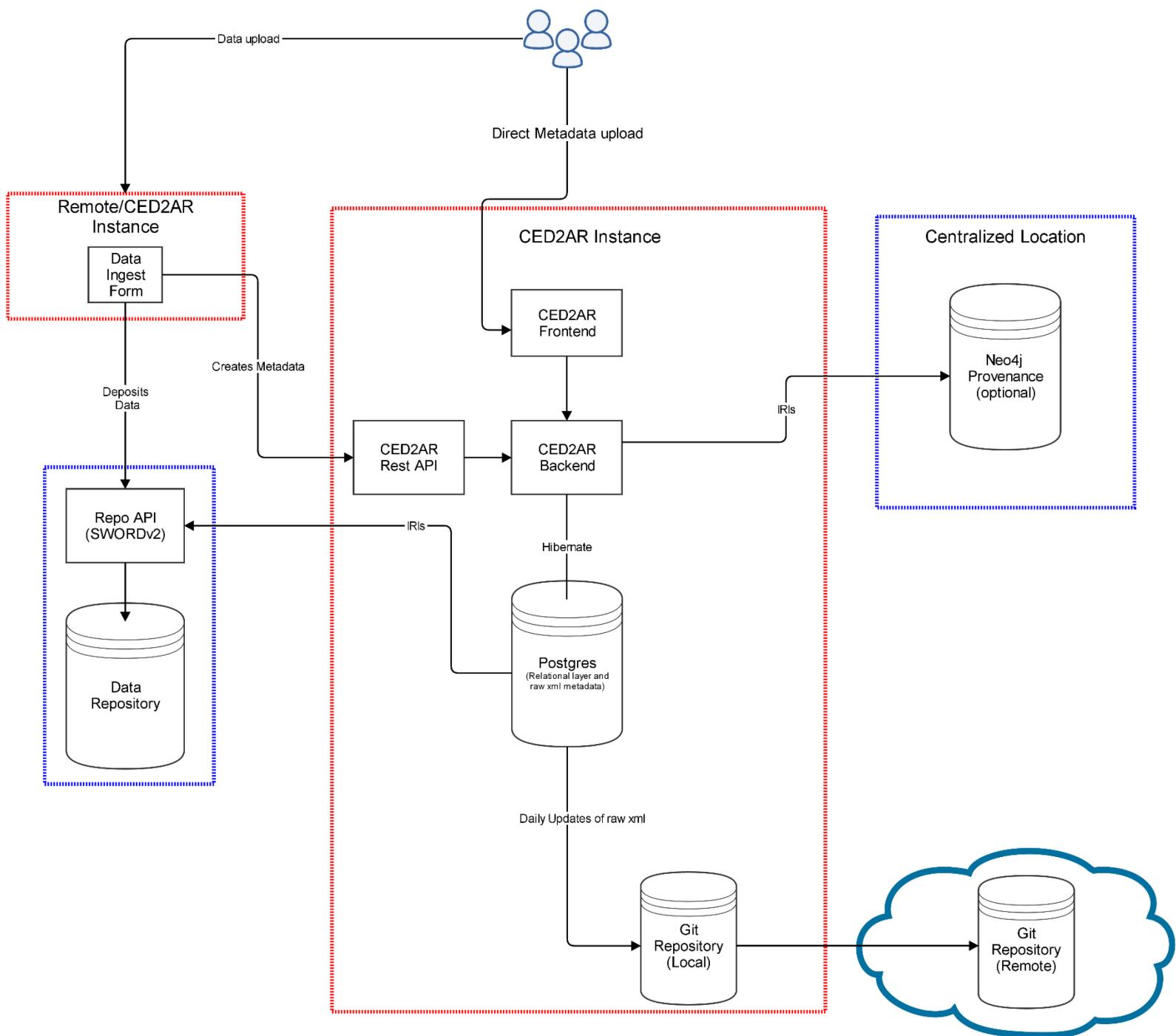
Making CED²AR v2 robust

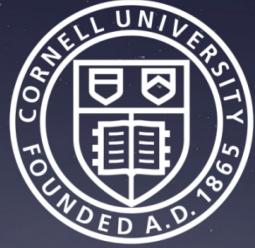
- Addition of UTF-8 editing support (Spanish, French, Portuguese, etc.)
- Additional fields (link to survey questions, anything within DDI-C)
- Bug fixes (search, UI, versioning)



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





Thank you!
Questions?



ced2ar-pub-dev-l@cornell.edu