Data Management of Confidential Data: the CED^{2}AR prototype

Lars Vilhuber[1] John M. Abowd[1] William Block[2]
Carl Lagoze[3] Jeremy Williams[2]

[1]Labor Dynamics Institute, ILR,

[2]Cornell Institute for Social and Economic Research,

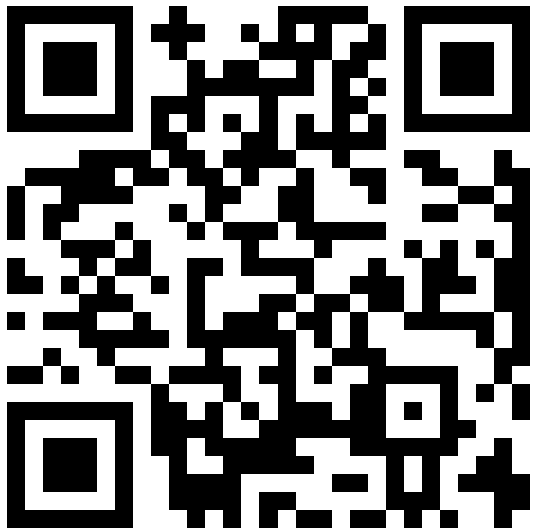
[3]University of Michigan

August 2013, Joint Statistical Meetings

Introduction

NCRN

- This work is part of the NSF Census Research Network (NCRN) - Cornell Node ("Integrated Research Support, Training and Data Documentation")
- Funded by NSF Grant #1131848.
- For more information, see www.ncrn.cornell.edu.



Introduction

Overview of work

- Basic program outlined in Abowd, Vilhuber, and Block (PSD 2012) and Lagoze, Block, Williams, Abowd, and Vilhuber, International Data Curation Conference (2013)
- PROV extension described in more detail in Lagoze, Williams, Vilhuber (Metadata and Semantics Research Conference, 2013 - proposed)

Motivation

Replication of research results

Critical element of science

- Replication of methods, data inputs, computational environment is a critical element of the scientific approach
- Journals, funding agencies (in the U.S.) have been moving to making archiving of inputs to scientific results more robust, even mandatory

Not a new problem

Econometrica

"In its first issue, the editor of Econometrica (1933), Ragnar Frisch, noted the importance of publishing data such that readers could fully explore empirical results. Publication of data, however, was discontinued early in the journal's history. [...] The journal arrived full-circle in late 2004 when Econometrica adopted one of the more stringent policies on availability of data and programs.

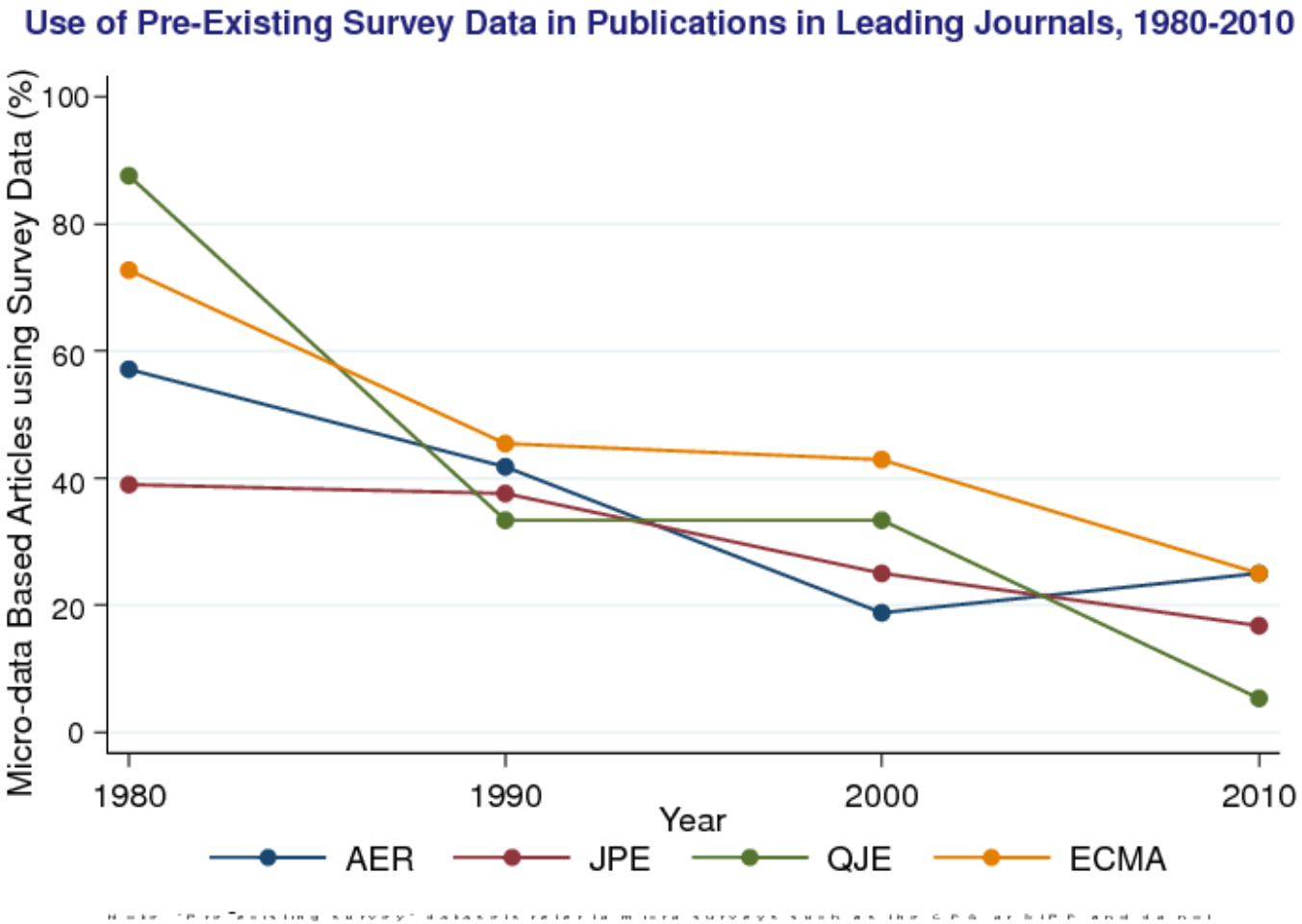
http://www.econometricsociety.org/submissions.asp#4 as cited in Anderson et al (2005)

Problem will become worse

Increased use of restricted-access data

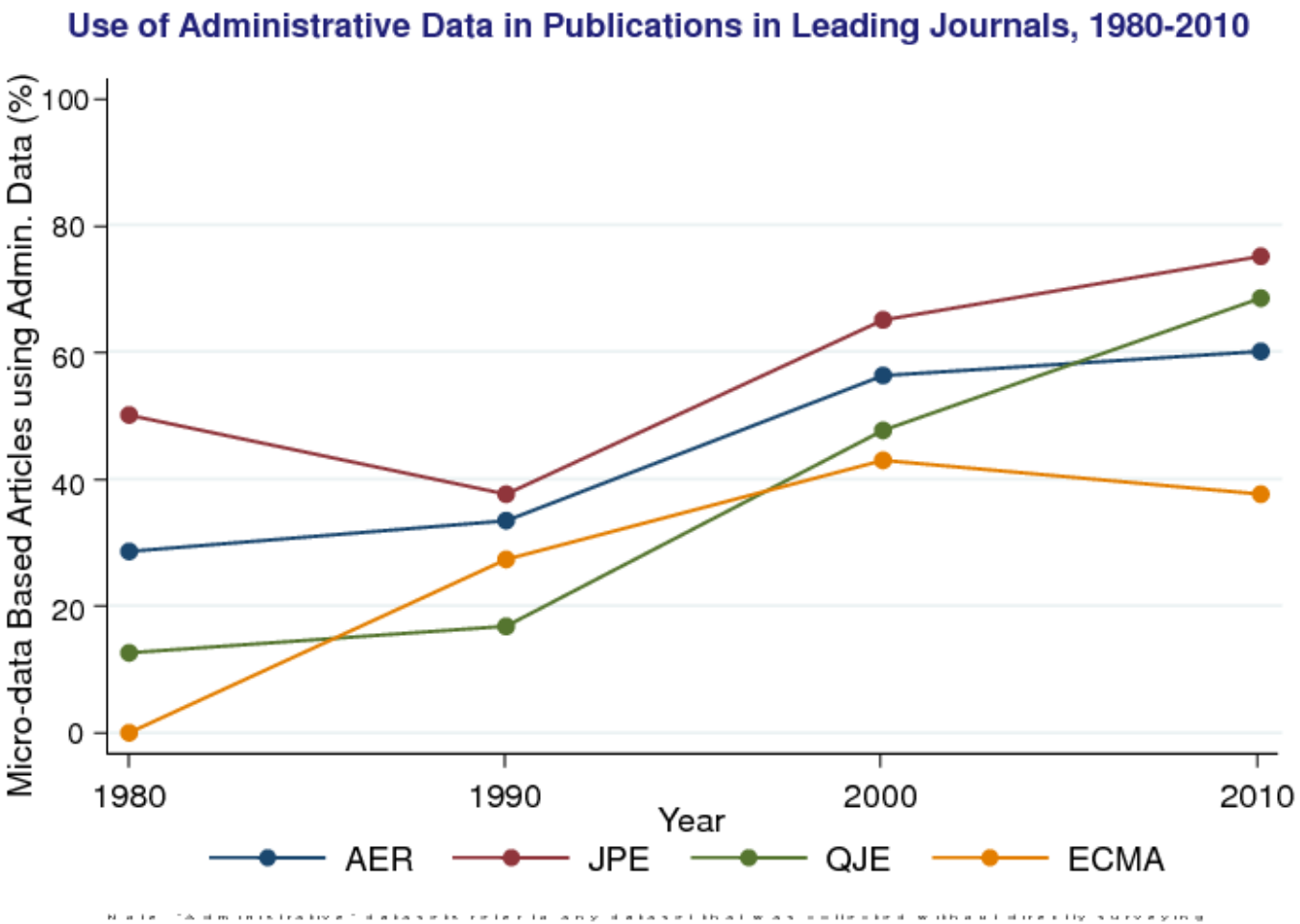
- Today's young scholars pursue research programs that mandate inherently identifiable data
  - Geospatial relations,
  - Exact genome data,
  - Networks of all sorts,
  - Linked administrative records
- These researchers acquire authorized, generally unfettered, restricted access to the confidential, identifiable data and perform their analyses in secure environments.
- Archiving (curation) of input data is complicated
- Knowledge discovery is complicated

Decline in the use of classic public-use data



include surveys designed by researchers for their study. Sample excludes studies whose primary data source is from developing countries.

Increase in the use of administrative data in economics



individuals (e.g., scanner data, stock prices, school district records, social security records).
Sample excludes studies whose primary data source is from developing countries.

Not limited to economics

Nature, 2012

"Many of the emerging 'big data' applications come from private sources that are inaccessible to other researchers. The data source may be hidden, compounding problems of verification, as well as concerns about the generality of the results."

(Huberman, Nature 482, 308 (16 February 2012) doi:10.1038/482308d)

Other domains

- Biology (genetics data, chemical compounds)
- Computer science (search records, single-firm examples)

Why we think there is a problem

Core issues

a. Insufficient curation (starting with archiving)
b. No way to reference data (unique identifiers)
c. No consistent way to learn about the data (metadata dissemination)
d. Weak or non-existent provenance tracing

Generalized problem

Multiple data sources in the US

- U.S. Census Bureau (RDC) ▸ more
- Internal Revenue Service (confidential, public-use) ▸ more
- Bureau of Labor Statistics (confidential, public-use data) ▸ more

Present elsewhere?

- Canada:
  - Centre for Data Development and Economic Research (CDER: RDC-like for business data) ▸ more
  - better: Canadian RDC network ▸ more
- France: better: Réseau Quetelet ▸ more
- Germany?

CED^{2}AR: A proposed solution

Comprehensive Extensible Data Documentation and Access (CED^{2}AR)

Core

We develop the core of a method for solving the data archive and curation problem that confronts the custodians of restricted-access research data and the scientific users of such data. Our solution recognizes the dual protections afforded by physical security and access limitation protocols.

Requirements

Royal Society (2012)

- Accessible (a researcher can easily find it);
- Intelligible (to various audiences);
- Assessable (are researchers able make judgements about or assess the quality of the data);
- Usable (at minimum, by other scientists).

Proposed solution

Extensible framework

- Based on existing standards (Data Documentation Initiative, DDI) with extension to accomodate disclosure protection mechanisms

Proposed solution

Extensible framework

- Based on existing standards (Data Documentation Initiative, DDI) with extension to accomodate disclosure protection mechanisms
- Connectors (import/export) to other sources and standards

Proposed solution

Extensible framework

- Based on existing standards (Data Documentation Initiative, DDI) with extension to accomodate disclosure protection mechanisms
- Connectors (import/export) to other sources and standards
- To be filled by multiple sources of metadata (some the curators/owners, others “crowd-sourced”)

Proposed solution

Extensible framework

- Based on existing standards (Data Documentation Initiative, DDI) with extension to accomodate disclosure protection mechanisms
- Connectors (import/export) to other sources and standards
- To be filled by multiple sources of metadata (some the curators/owners, others “crowd-sourced”)
- Interim solution for those datasets without unique identifiers (Digital Object Identifier, DOI)

Database design

Multiple sources

- Data-curator-provided metadata (possibly regularly updated, PRUNED)
- Alternate sources (IPUMS data to describe Decennial Census)
- **User-provided metadata (wiki) (planned fall 2013)**

Multiple outputs

- Local query (**working**)
- Remote federation or export
- Synchronization back to data-curator (data enclave!)

Provenance

The provenance problem

"data provenance, one kind of metadata, pertains to the derivation history of a data product starting from its original sources" [...] "from it, one can ascertain the quality of the data base and its ancestral data and derivations, track back sources of errors, allow automated reenactment of derivations to update the data, and provide attribution of data sources"

Simmhan, Plale, and Gannon, "A survey of data provenance in e-science," ACM Sigmod Record, 2005

Provenance (cont)

PROV model

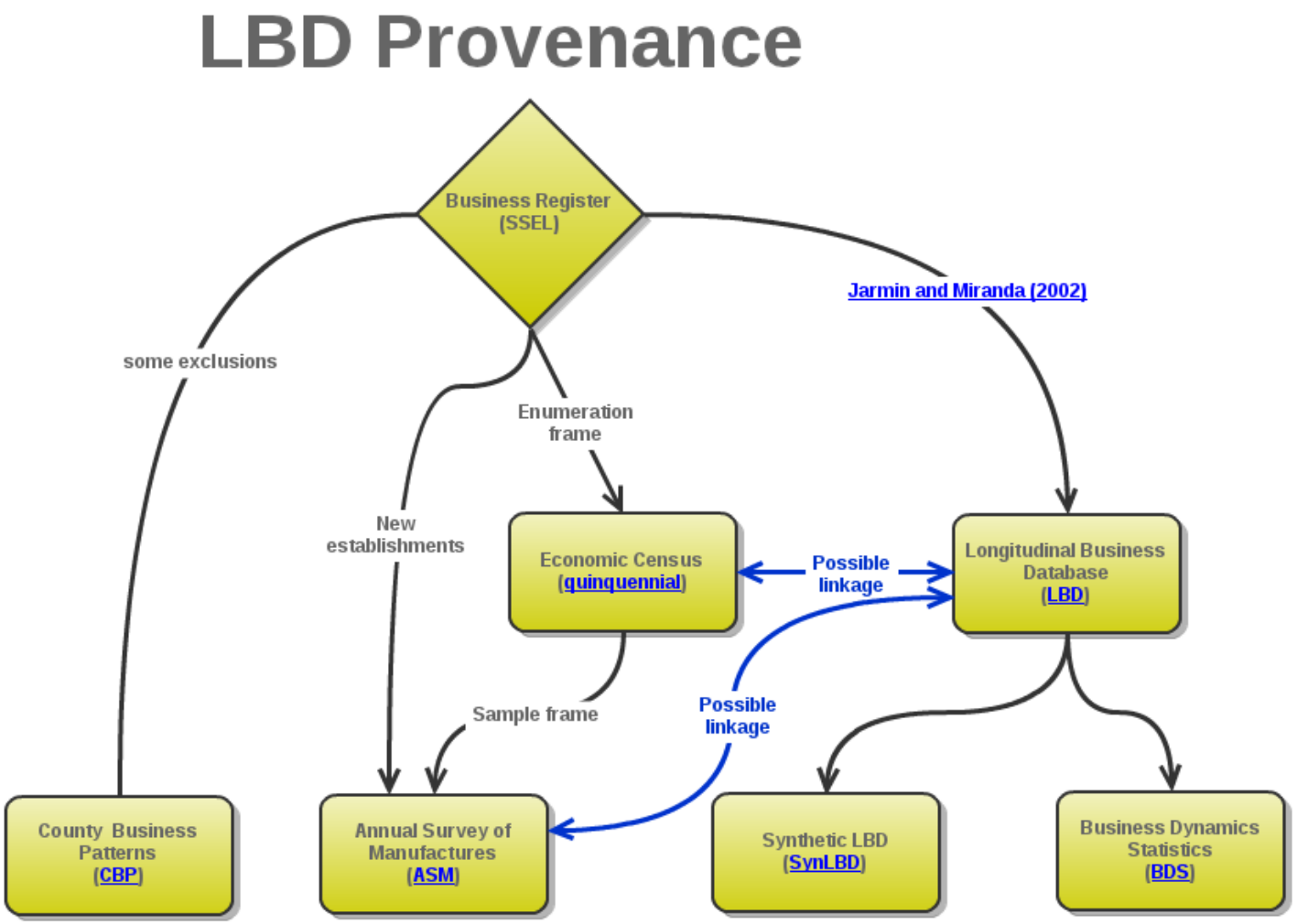
W3C PROV Model based in the notions of

1. **entities** that are physical, digital, and conceptual things in the world;
2. **activities** that are dynamic aspects of the world that change and create entities; and
3. **agents** that are responsible for activities.
4. a set of **relationships** that can exist be- tween them that express attribution,. delegation, derivation, etc.

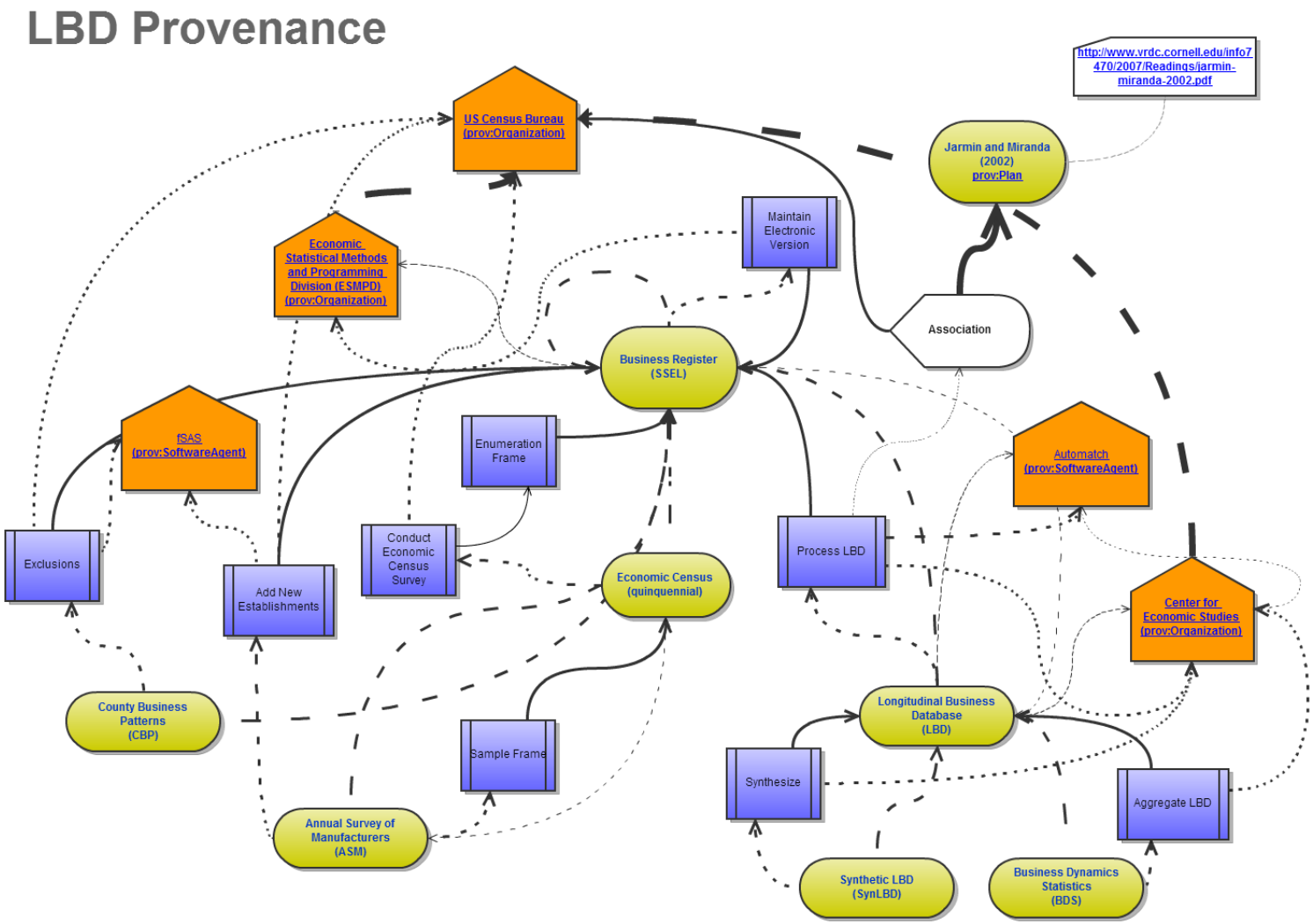
PROV and Metadata

Not (currently) a “native” component of DDI

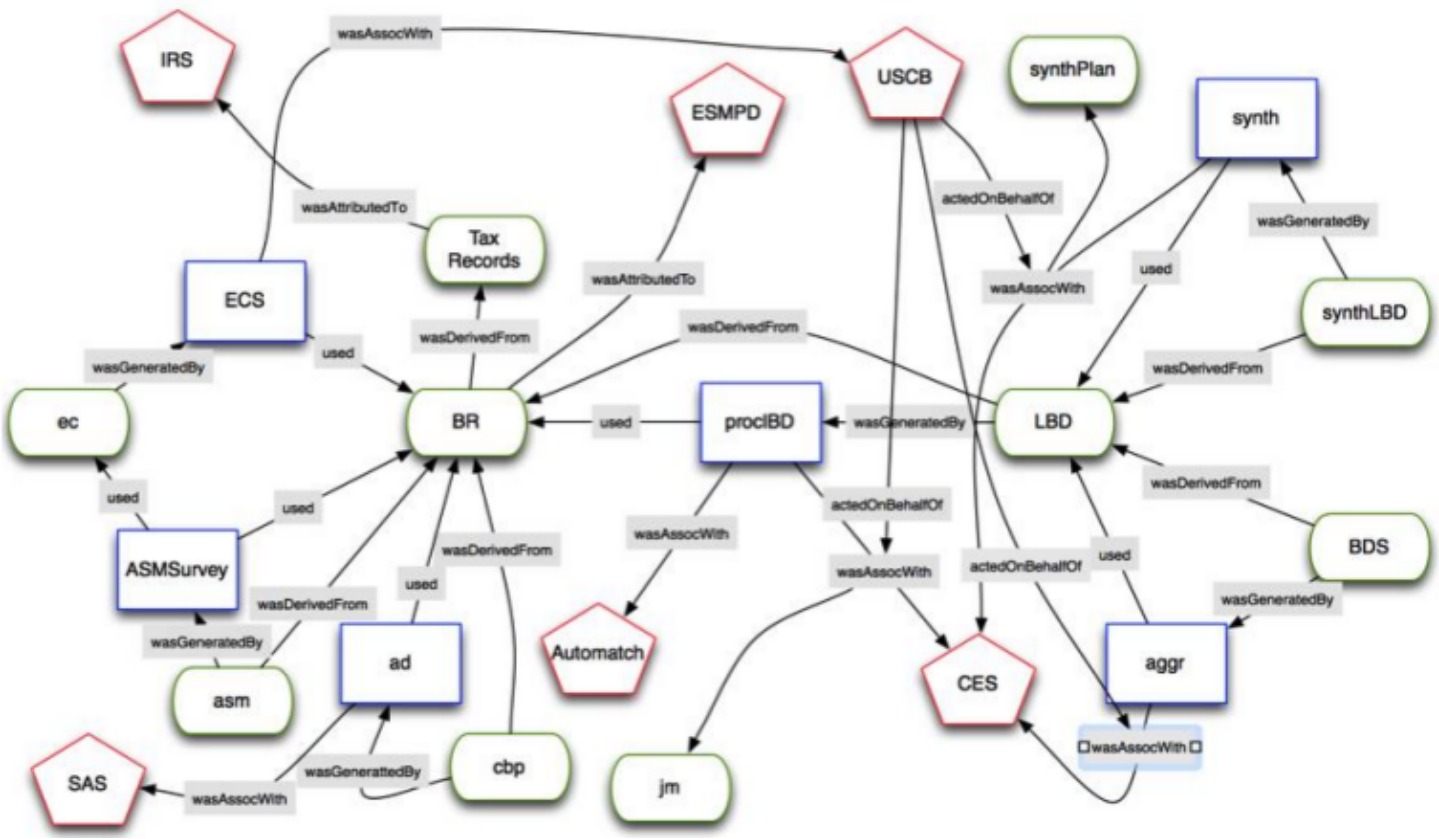
Incorporating PROV (LBD)



Incorporating PROV (LBD)



Incorporating PROV (LBD)



Incorporating PROV (LBD)

```
entity(cdr:LBD, [prov:type='cdr:dataset', prov:label="Longitudinal Business Data"])
entity(cdr:synthLBD, [prov:type='cdr:dataset', prov:label="Synthetic LBD"])
entity(cdr:BDS, [prov:type='cdr::dataset', prov:label="Business Dynamics Statistics"])
entity(cdr:BR, [prov:type='cdr:dataset', prov:label="Business Register"])
entity(cdr:cbp, [prov:type='cdr:dataset', prov:label="County Business Patterns"])
entity(cdr:asm, [prov:type='cdr:dataset', prov:label="Annual Survey of Manufacturers"])
entity(cdr:ec, [prov:type='cdr:dataset', prov:label="Economic Census"])
entity(cdr:jm, [prov:type='prov:Plan', prov:label="Jarmin Miranda 2002"])
entity(cdr:synthPlan, [prov:type='prov:Plan', prov:label="synthetic plan"])
entity(cdr:tax, [prov:type='cdr:dataSet', prov:label="IRS Tax Records"])

agent(cdr:USCB, [prov:type='prov:Organization, prov:label="US Census Bureau"])
agent(cdr:CES, [prov:type='prov:Organization, prov:label="Center for Economic Studies"])
agent(cdr:IRS, [prov:type='prov:Organization, prov:label="Internal Revenue Service"])
agent(cdr:autoMatch, [prov:type='prov:SoftwareAgent'])
agent(cdr:SAS, [prov:type='prov:SoftwareAgent'])
agent(cdr:ESMPD, [prov:type='prov:SoftwareAgent',
    prov:label="Economic Statistical Methods and Programming Division"])

activity(cdr:synth, [prov:label="anonymize"])
activity(cdr:aggr, [prov:label="aggregate"])
activity(cdr:procLBD, [prov:label="process LBD"])
activity(cdr:ad, [prov:label="aggregation/disclosure protection"])
activity(cdr:asmSurvey, [prov:label="ASM Survey"])
activity(cdr:ecs, [prov:label="economic census survey"])
```

Work on PROV

More details forthcoming

See Lagoze, Williams, Vilhuber “Encoding Provenance Metadata for Social Science Datasets”, submitted to Metadata and Semantics Research Conference (soon)

State of the implementation

DDI extension

Being incorporated.

DOI assignment

Our project (NCRN) will assign/register DOI if not provided by curator/owner

Database

Design finalized, database populated with metadata for newest SIPP Synthetic Beta. Wiki additions Fall 2013

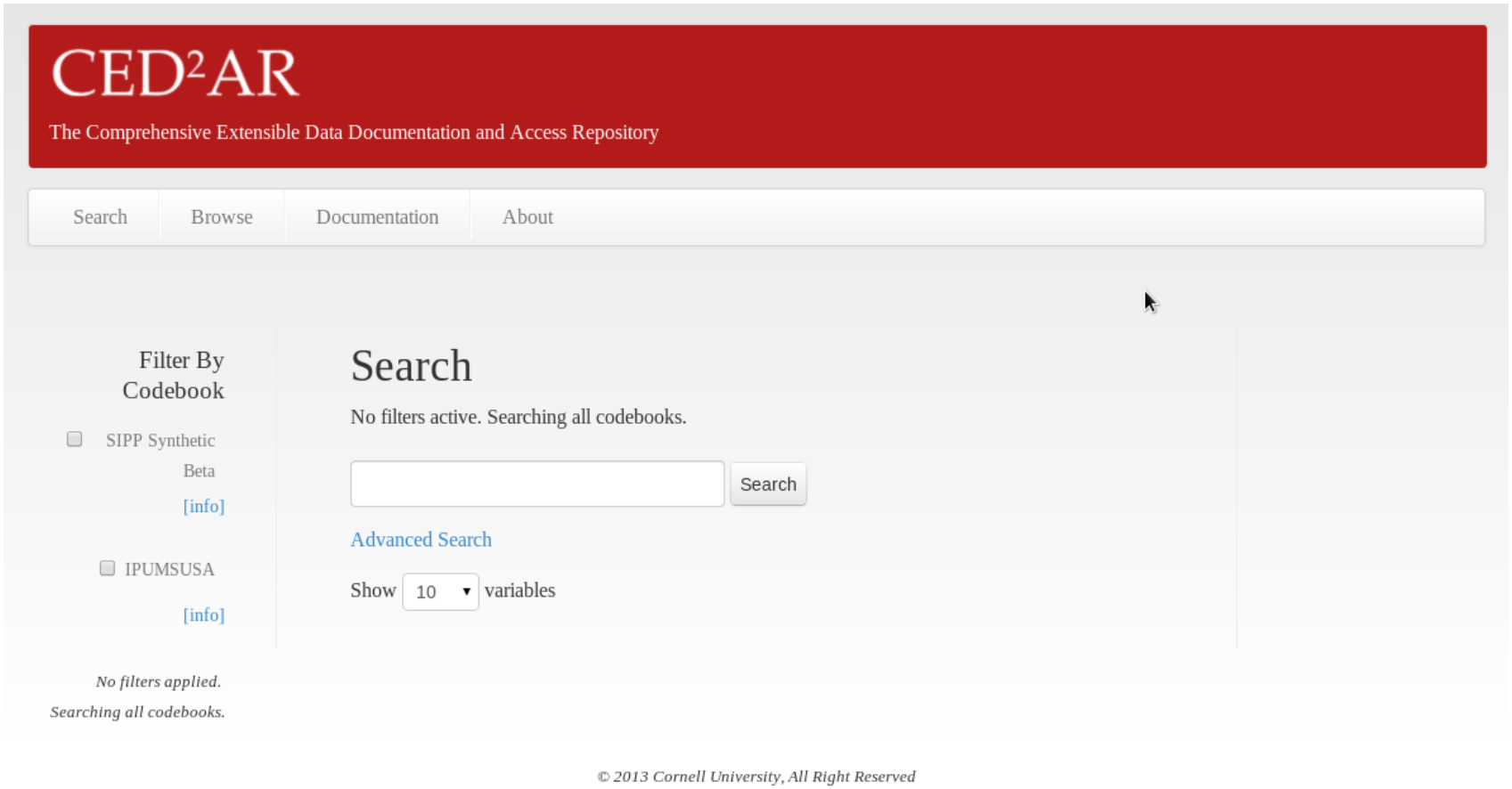
UI

Version 1.1 of the UI being completed (more robust, scalable). Wiki additions in Fall 2013

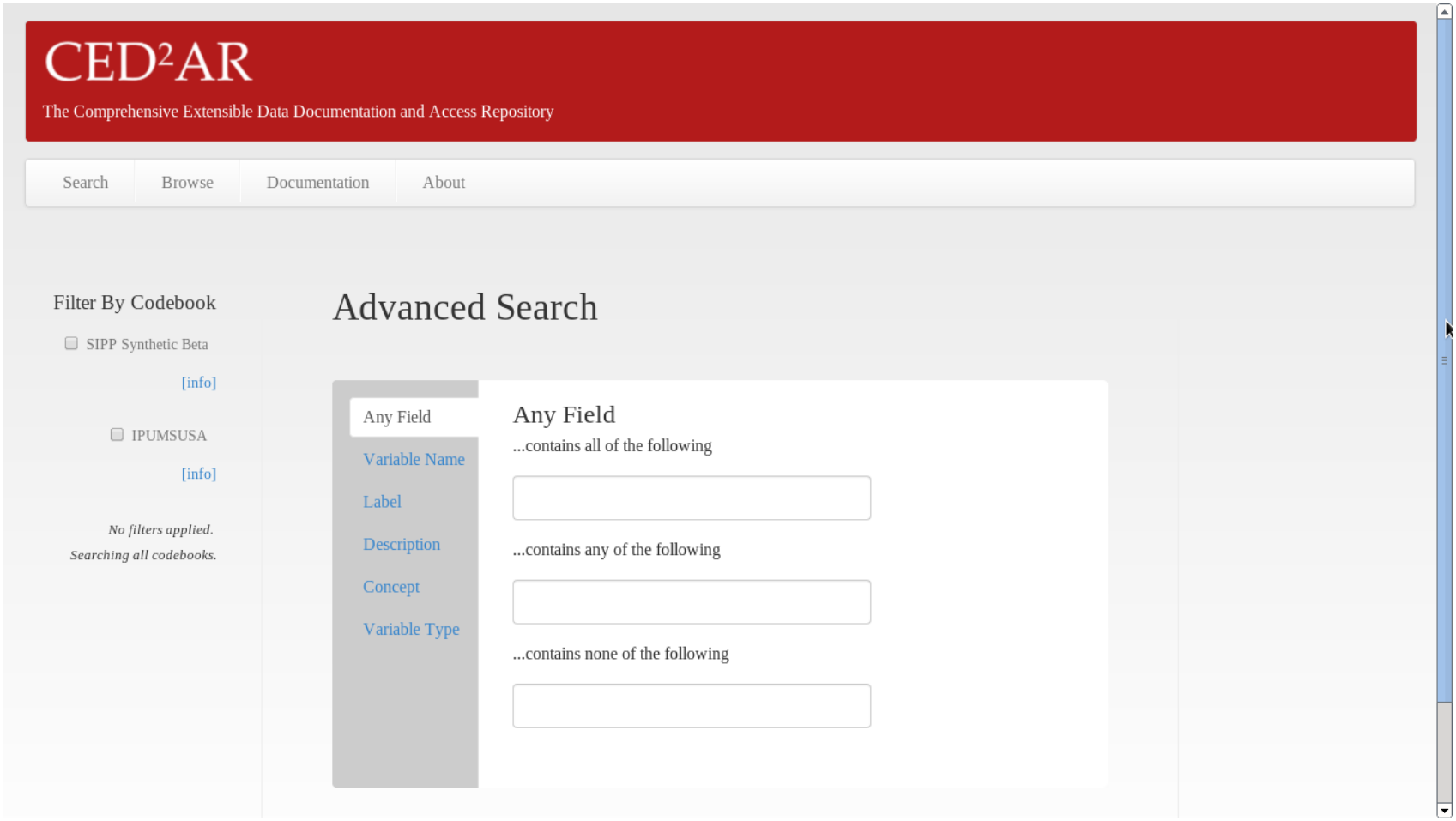
Provenance

PROV extension, integration Winter 2013/14

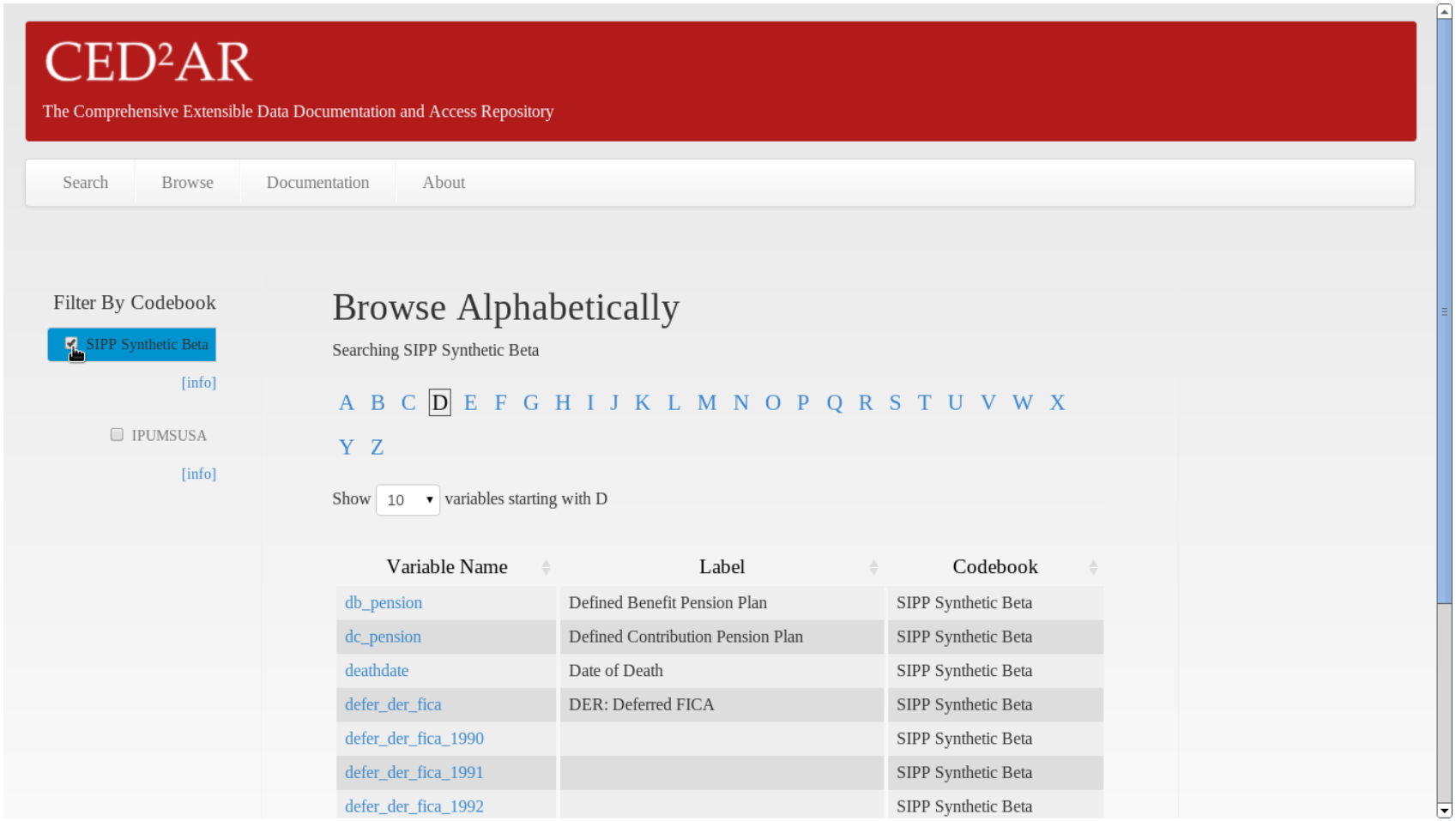
Screenshot


CED²AR
The Comprehensive Extensible Data Documentation and Access Repository
Search
Browse
Documentation
About
Filter By Codebook
SIPP Synthetic Beta
[info]
IPUMSUSA
[info]
No filters applied.
Searching all codebooks.
Search
No filters active. Searching all codebooks.
Search
Advanced Search
Show 10 variables
© 2013 Cornell University, All Right Reserved

Screenshot


CED²AR
The Comprehensive Extensible Data Documentation and Access Repository
Search
Browse
Documentation
About
Filter By Codebook
SIPP Synthetic Beta
[info]
IPUMSUSA
[info]
No filters applied.
Searching all codebooks.
Advanced Search
Any Field
Variable Name
Label
Description
Concept
Variable Type
Any Field
...contains all of the following
...contains any of the following
...contains none of the following

Screenshot


CED²AR
The Comprehensive Extensible Data Documentation and Access Repository
Search
Browse
Documentation
About
Filter By Codebook
SIPP Synthetic Beta
[info]
IPUMSUSA
[info]
Browse Alphabetically
Searching SIPP Synthetic Beta
A B C D E F G H I J K L M N O P Q R S T U V W X
Y Z
Show 10 variables starting with D
Variable Name
Label
Codebook
db_pension
Defined Benefit Pension Plan
SIPP Synthetic Beta
dc_pension
Defined Contribution Pension Plan
SIPP Synthetic Beta
deathdate
Date of Death
SIPP Synthetic Beta
defer_der_fica
DER: Deferred FICA
SIPP Synthetic Beta
defer_der_fica_1990
SIPP Synthetic Beta
defer_der_fica_1991
SIPP Synthetic Beta
defer_der_fica_1992
SIPP Synthetic Beta

The end

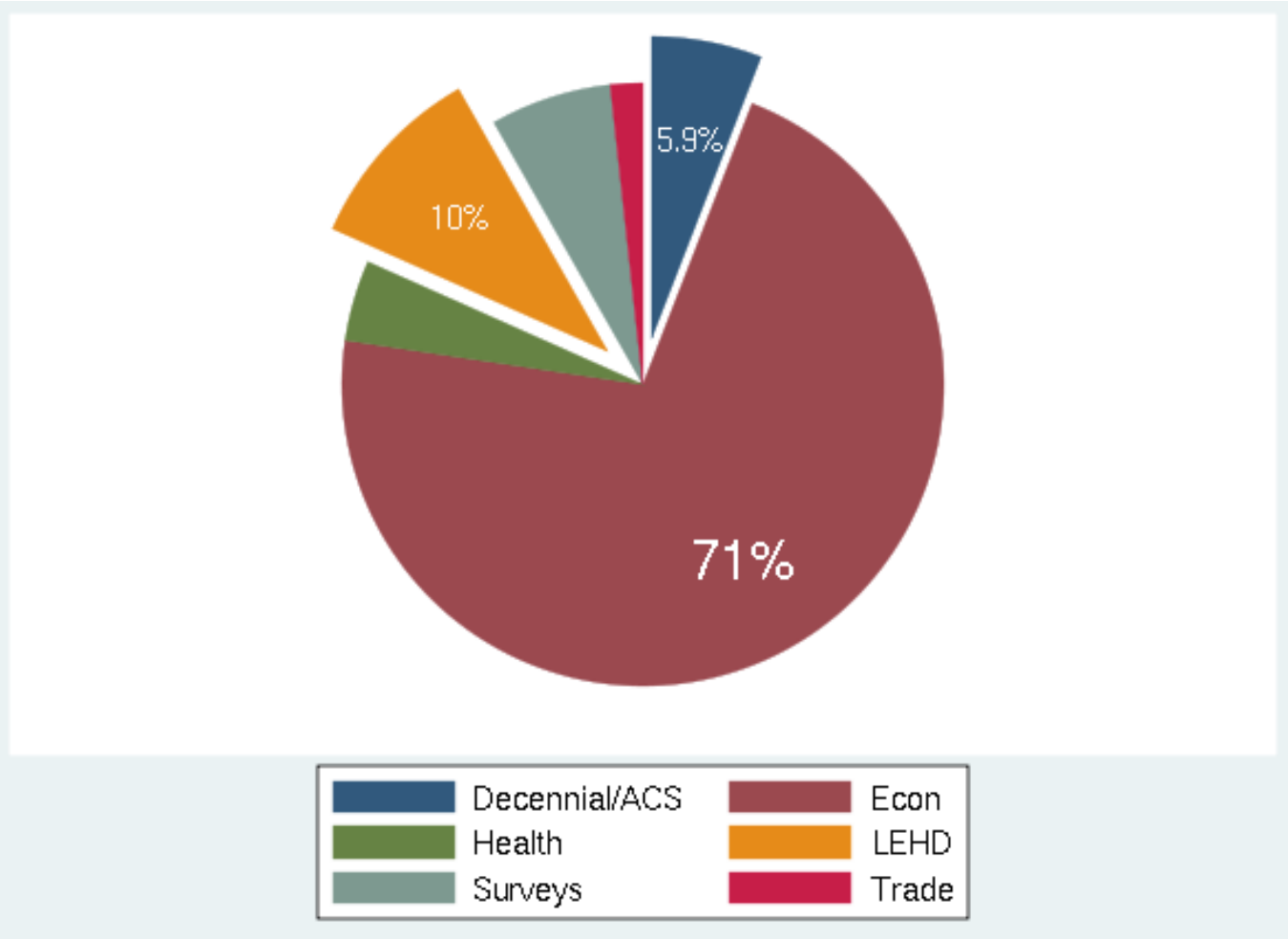
Thank you

- [3] for more details
- Labor Dynamics Institute
- VirtualRDC @ Cornell
- NCRN Cornell website

$Id: Presentation-JSM-subdoc.tex 4240 2013-08-04 02

Dataset usage in Census RDC

1,505 project-dataset pairs



Many projects use multiple datasets.

Economic (business) datasets

- 71% of datasets are business (economic) datasets
- Primarily establishment-based records from the Economic Censuses and Surveys, the Business Register, and the Longitudinal Business Database (LBD)
- They form the core of the modern industrial organization studies [5, 9] as well as modern gross job creation and destruction in macroeconomics [4, 6].
- But there are no public-use micro-data for these establishment-based products
- Exception: recently-released Synthetic LBD [2, 7]
- Currently no active curation (of derived datasets) [a], no way to reference [b], convoluted way to learn about the data structure [c*]

LEHD data

Linked employer-employee data

- Longitudinal and cross-sectional detail
- New confidentiality protection methodologies [1, 8] have unlocked large amounts of data for public-use: highly detailed local area tabulations exist based on the LEHD data
- But: no public-use micro-data exist for this longitudinal job frame or any of its derivative files.
- Confidential data are dynamic (quarterly changes)
- Currently some active curation (archiving, 10-yr!) [a*], no way to reference (publicly) [b*], convoluted way to learn about the data structure [c*]

Not unique to Census Bureau

Internal Revenue Service/ Social Security Administration

- New projects (Chetty et al, 2012; von Wachter and co-authors) have created and/or used linked longitudinal data at the IRS or the Social Security Administration.
- Neither agency has long-run experience at the statistical data curation function [a], (meta)data dissemination [b,c].
- Although both IRS and SSA have produced statistical tables for a long time.

Not unique to Census Bureau

Bureau of Labor Statistics

- Long history of making time-series available
- Limited access to microdata at the BLS
- Unknown curation [a]
- Even for public-use data, no way to reference specific releases [b]
- No well-established way to learn about microdata [c]

Canadian Centre for Data Development and Economic Research

Statistics Canada Statistique Canada

Canada

Statistics Canada

www.statcan.gc.ca

Français | Home | Contact Us | Help | Search | canada.gc.ca

Home > The Canadian Centre for Data Development and Economic Research (CDER) >

The Canadian Centre for Data Development and Economic Research

- Application process and guidelines
- Proposal requirements
- Application for accreditation
- Data sets
- Pricing policy
- Microdata research contract
- Frequently asked questions
- Contact information

Data Sets

A number of business micro databases can be accessed at CDER. Key databases are listed below. For more documentation on each of the databases, or documentation on other databases, please contact CDER at cder@statcan.gc.ca.

Annual Survey of Manufacturing
Annual Survey of Manufacturing – Export and Import Registry Database
Canada Border Service Agency Customs Database
Capital and Investment Program
Longitudinal Employment Analysis Program
Longitudinal Worker File
National Accounts Longitudinal Microdata File
T2-LEAP
T2-LEAP-Export and Import Registry Database
Survey of Financing of Small and Medium Enterprises
Survey of Innovation and Business Strategies
Workplace Employee Survey

Annual Survey of Manufactures (ASM)

The ASM is a survey that covers all manufacturing locations together with associated head offices, sales offices and auxiliary units which have been classified to the manufacturing industries. Details collected include principal industrial st
cost of materials and supplies used, c

Canadian Research Data Centres

RDC projects and publications
Conferences
FAQ

top banner, then select the "Advanced Search" option and in the field "Include pages with all these words" type in the text url:rdc and add any key word. For example, "url:rdc census" which will result in all pages on the Research Data Centres Program website that contain the keyword "census".

Surveys available in the RDCs

The following data sets are currently available at the RDCs. For additional sources of data please refer to Statistics Canada **Products and Services**.

To read a short **description** about a specific survey used at the RDCs, click on the survey details.

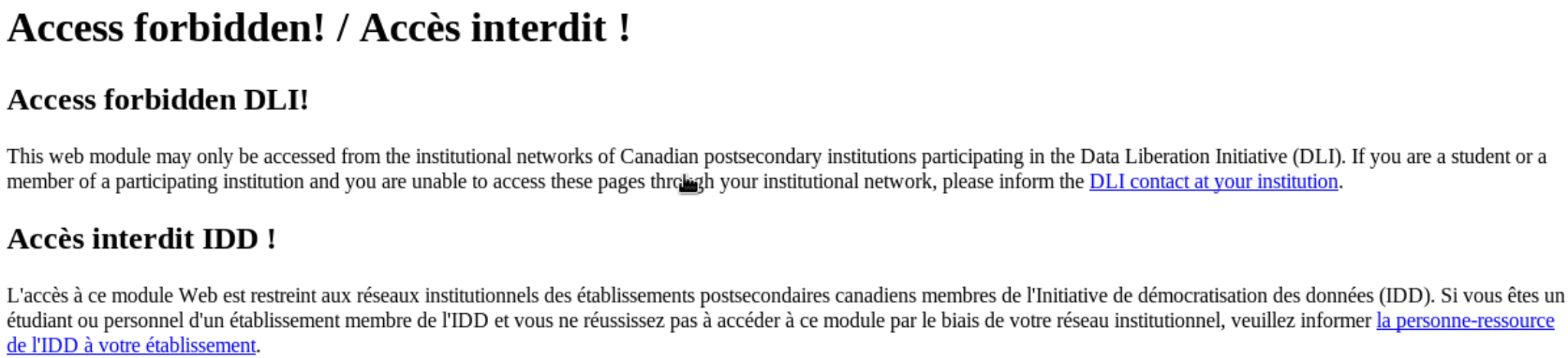
To access **detailed documentation** on a specific survey used at the RDCs, click on the appropriate cycle or year. Many of the surveys below have multiple cycles. The links below will take you to the most recent cycle or wave released. Please select "Other reference period" in the "Definitions, Data Sources and Methods Pages" for links to documentation for the earlier cycles.

Record Number	Survey Name	Acronym
5108	Aboriginal Children's Survey	ACS
3250	Aboriginal Peoples Survey	APS
3879	Adult Education and Training Survey	AETS
3207	Canadian Cancer Registry	CCR
3226	Canadian Community Health Survey - Annual Component	CCHS
5015	Canadian Community Health Survey – Mental Health	CCHS
5049	Canadian Community Health Survey - Nutrition	CCHS
5146	Canadian Community Health Survey – Healthy Aging	CCHS
5071	Canadian Health Measures Survey Biobank	CHMS
4440	Canadian Tobacco Use Monitoring Survey	CTUMS
	Census of Population - Additional documentation	
4508	Ethnic Diversity Survey - User Guide - Codebook	EDS
3504	Survey of Family E	

Canadian Research Data Centres

... but also not perfect

Attempt to access data information on General Social Survey



Access forbidden! / Accès interdit !

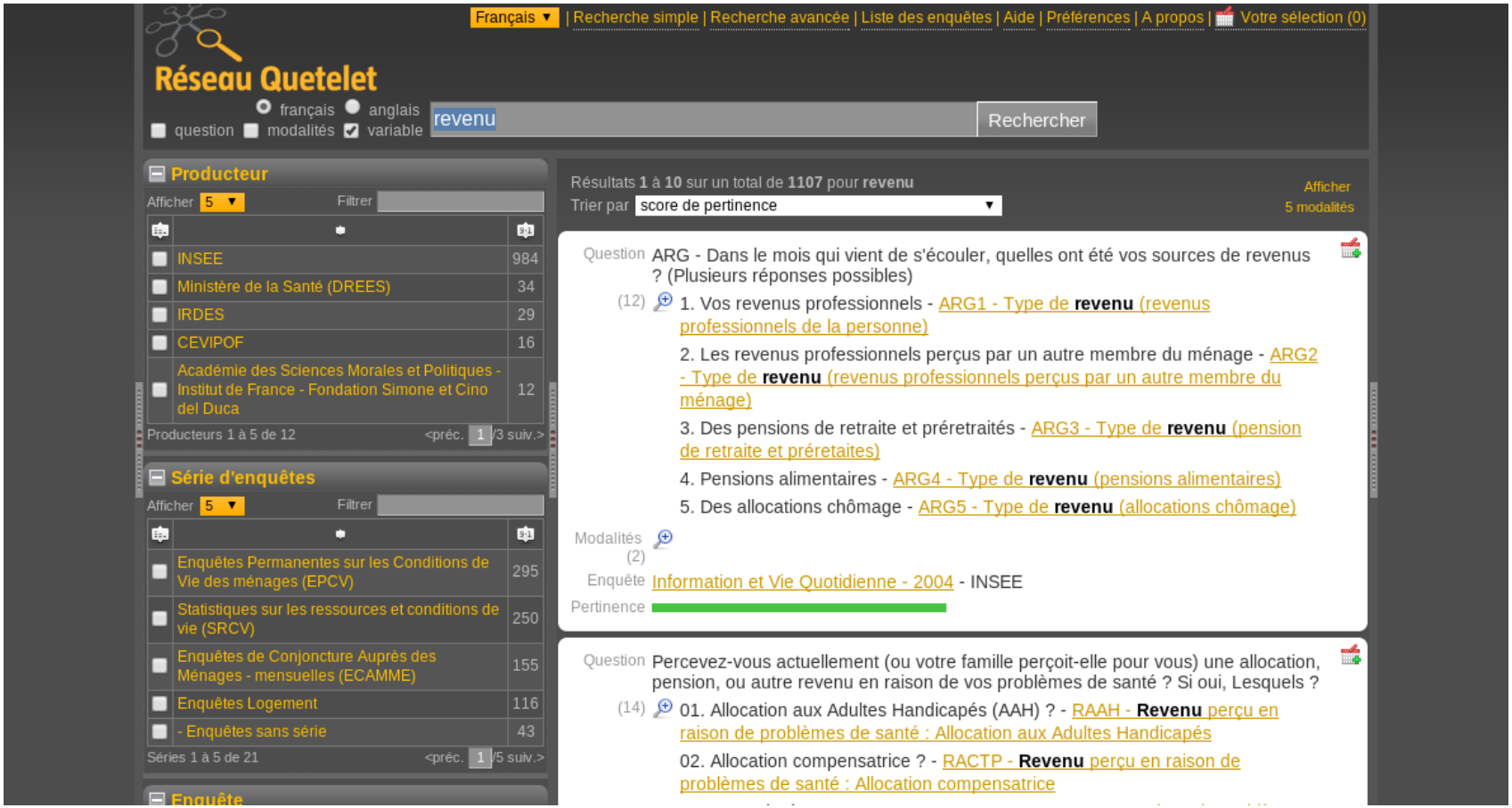
Access forbidden DLI!

This web module may only be accessed from the institutional networks of Canadian postsecondary institutions participating in the Data Liberation Initiative (DLI). If you are a student or a member of a participating institution and you are unable to access these pages through your institutional network, please inform the DLI contact at your institution.

Accès interdit IDD !

L'accès à ce module Web est restreint aux réseaux institutionnels des établissements postsecondaires canadiens membres de l'Initiative de démocratisation des données (IDD). Si vous êtes un étudiant ou personnel d'un établissement membre de l'IDD et vous ne réussissez pas à accéder à ce module par le biais de votre réseau institutionnel, veuillez informer la personne-ressource de l'IDD à votre établissement.

Réseau Quetelet




J. M. Abowd, K. Gittings, K. L. McKinney, B. E. Stephens, L. Vilhuber, and S. Woodcock, "Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series," Federal Committee on Statistical Methodology, Tech. Rep., January 2012. [Online]. Available: http://www.fcsm.gov/events/papers2012.html


J. M. Abowd and L. Vilhuber. (2010) Synthetic data server. [Online]. Available: http://www.vrdc.cornell.edu/sds/


J. M. Abowd, L. Vilhuber, and W. Block, "A proposed solution to the archiving and curation of confidential scientific inputs," in *Privacy in Statistical Databases*, ser. Lecture Notes in Computer Science, J. Domingo-Ferrer and I. Tinnirello, Eds., vol. 7556. Springer, 2012, pp. 216–225. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33627-0_17


S. J. Davis, J. C. Haltiwanger, and S. Schuh, *Job creation and destruction*. Cambridge, MA: MIT Press, 1996.


T. Dunne, M. J. Roberts, and L. Samuelson, "The Growth and Failure of U.S. Manufacturing Plants," *Quarterly Journal of Economics*, vol. 104, no. 4, pp. 671–698, 1989.


J. Haltiwanger, R. S. Jarmin, and J. Miranda, "Who creates jobs? Small vs. large vs. young," Center for Economic Studies, U.S. Census Bureau, Working Papers 10-17, Aug. 2010. [Online]. Available: http://ideas.repec.org/p/cen/wpaper/10-17.html


S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd, "Towards unrestricted public use business microdata: The synthetic longitudinal business database," *International Statistical Review*, vol. 79, no. 3, pp. 362–384, December 2011. [Online]. Available: http://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html


A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," *International Conference on Data Engineering (ICDE)*, 2008.


G. S. Olley and A. Pakes, "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, vol. 64, no. 6, pp. 1263–1297, November 1996. [Online]. Available: http://www.jstor.org/stable/2171831