A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs

John M. Abowd[1] Lars Vilhuber[1] William Block[2]

[1] Labor Dynamics Institute, ILR,

[2] Cornell Institute for Social and Economic Research,

Cornell University, Ithaca, NY, USA

September 2012, PSD 2012

Motivation

Replicating 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)

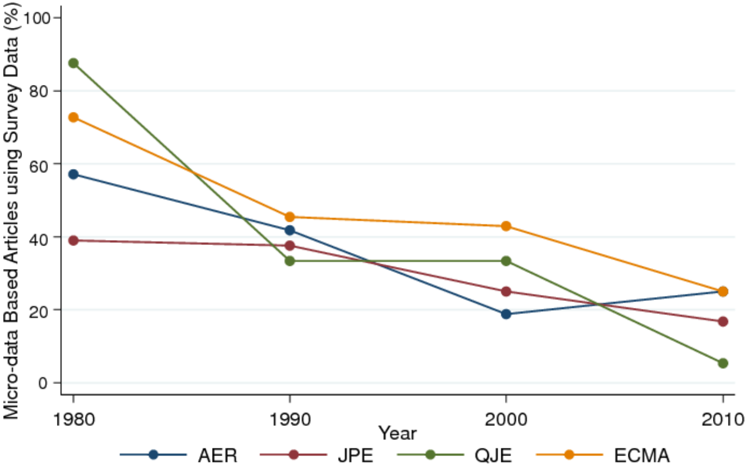
Problem will become worse

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

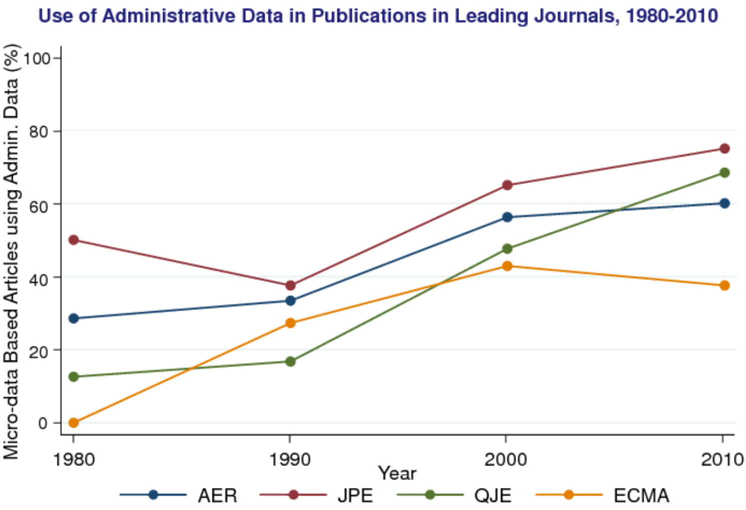
Decline in the use of classic public-use data

Use of Pre-Existing Survey Data in Publications in Leading Journals, 1980-2010



Note: "Pre-existing survey" datasets refer to micro surveys such as the CPS or SIPP and do not 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



Use of Administrative Data in Publications in Leading Journals, 1980-2010

Note: "Administrative" datasets refer to any dataset that was collected without directly surveying individuals (e.g., scanner data, stock prices, school

Sample excludes studies whose primary data sourc

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)

Stating the problem

Why we think there is a problem

Core issues

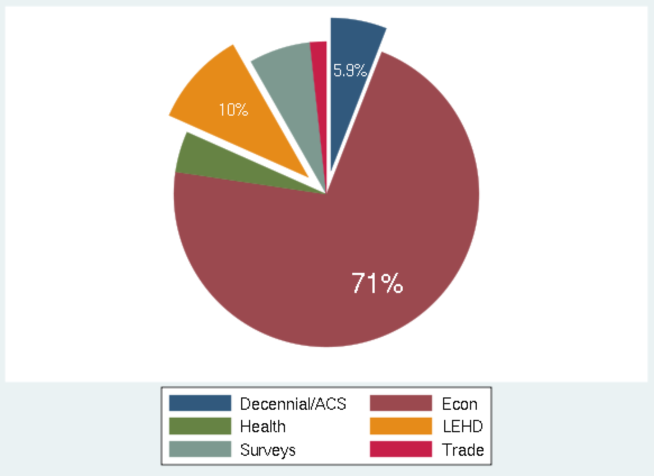
a Insufficient curation (starting with archiving)

b No way to reference data (unique identifiers)

c No consistent way to learn about the data (metadata dissemination)

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]

Core problems

- Curation

Core problems

- Curation
- Identification

Core problems

- Curation
- Identification
- Information dissemination

Core problems

- Curation
- Identification
- Information dissemination

←

- require cooperation of NSI

Core problems

- Curation ←
- Identification ←
- Information dissemination

- require cooperation of NSI
- partial solution (DOI)

Core problems

▸ Curation	←	▸ require cooperation of NSI
▸ Identification	←	▸ partial solution (DOI)
▸ Information dissemination	←	▸ core proposal

A proposed solution

Proposed solution

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”)

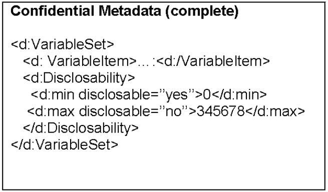
Proposed solution

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)

Extensions to DDI

Basic idea

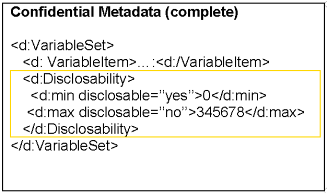


Confidential Metadata (complete)

```
<d:VariableSet>
  <d: VariableItem>. . . :<d:/VariableItem>
  <d:Disclosability>
    <d:min disclosable=''yes''>0</d:min>
   <d:max disclosable=''no''>345678</d:max>
  </d:Disclosability>
</d:VariableSet>
```

Extensions to DDI

Basic idea



Confidential Metadata (complete)

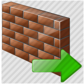
```
<d:VariableSet>
  <d: VariableItem>. . . :<d:/VariableItem>
  <d:Disclosability>
    <d:min disclosable=''yes''>0</d:min>
   <d:max disclosable=''no''>345678</d:max>
  </d:Disclosability>
</d:VariableSet>
```

Extensions to DDI

Basic idea

Confidential Metadata (complete)

```
<d:VariableSet>
  <d: VariableItem>... :<d:/VariableItem>
  <d:Disclosability>
    <d:min disclosable=''yes'')>0</d:min>
   <d:max disclosable=''no'')>345678</d:max>
  </d:Disclosability>
</d:VariableSet>
```



Derived Public Use Metadata (limited)

```
<d:VariableSet>
<d: VariableItem>... :<d:/VariableItem>
  <d:Disclosability>
    <d:min>0</d:min>
    <d:max>not disclosable</d:max>
  </d:Disclosability>
</d:VariableSet>
```

Database design

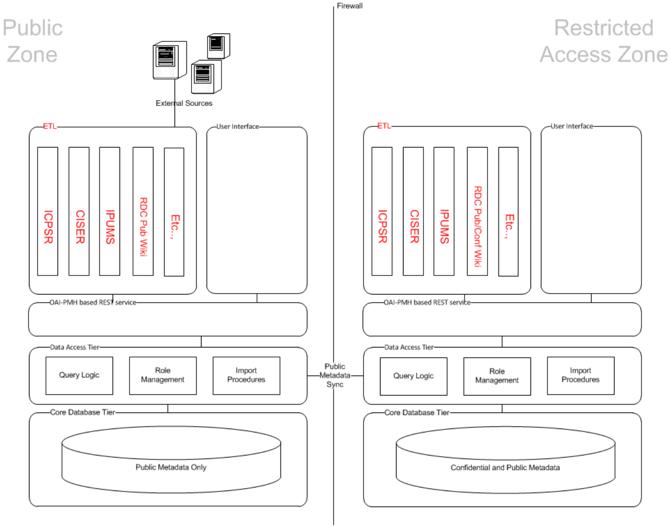
Multiple sources

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

Multiple outputs

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

Generic description



Identifiers

Unique identifiers for *articles*

```
Huberman, B. A.
Sociology of science:
Big data deserve a bigger audience
Nature, 2012, 482, 308-308
doi:10.1038/482308d
```

Unique identifiers for *data*

"DOI names are assigned to any entity for use on digital networks. They are used to provide current information, including where they (or information about them) can be found on the Internet. Information about a digital object may change over time, including where to find it, but its DOI name will not change." http://datacite.org/whatisdoi, accessed on Sept 26, 2012.

State of the implementation

DDI extension

Being formalized.

DOI assignment

Our project (NCRN) will assign DOI if not provided by curator/owner. May be validated by disclosable checksums (MD5 or similar) to verify change of files. (additional dataset-level metadata!)

Database

Design finalized, first connectors implemented, alpha-quality implementation with IPUMS, SIPP Synthetic Beta, simulated SIPP Gold Standard up and running.

The end

Thank you

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

```
$Id: Presentation-PSD-subdoc.tex 3219 2012-09-27 07
```

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