

Replication - issues

Lars Vilhuber¹

¹Labor Dynamics Institute, ILR, Cornell University, United States

August 2016

Replicability

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

The problem

Good intentions, costly access

“researchers could submit programs that [...] research assistants would run. Alternatively, researchers wishing to work directly with the data could come and work on the Institute’s premises. ”

The problem

Good intentions, costly access

“researchers could submit programs that [...] research assistants would run. Alternatively, researchers wishing to work directly with the data could come and work on the Institute’s premises. ”

Uncertain access

“Data [...] is proprietary and owned by the Alachua County, Florida School District. The corresponding author [...] holds the deidentified dataset [...] and will provide copies to authors who receive written permission from the Alachua County Public Schools.”

The problem

Good intentions, costly access

“researchers could submit programs that [...] research assistants would run. Alternatively, researchers wishing to work directly with the data could come and work on the Institute’s premises. ”

Uncertain access

“Data [...] is proprietary and owned by the Alachua County, Florida School District. The corresponding author [...] holds the deidentified dataset [...] and will provide copies to authors who receive written permission from the Alachua County Public Schools.”

No access

Some do not provide any information on access.

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)



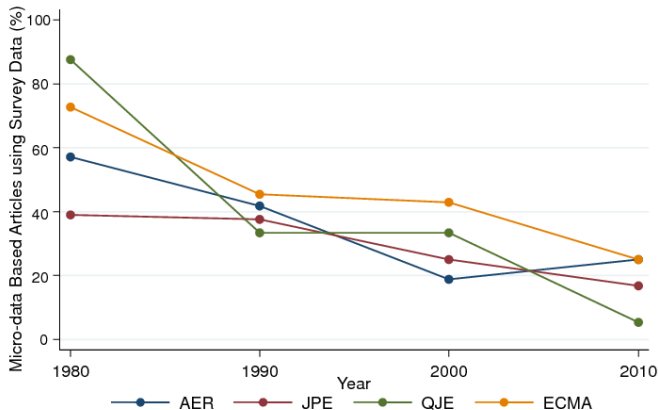
Problem will become worse

Increased use of restricted-access data

- ▶ Archiving (curation) of input data is complicated
- ▶ Knowledge discovery is complicated

Decline in the use of classic public-use data

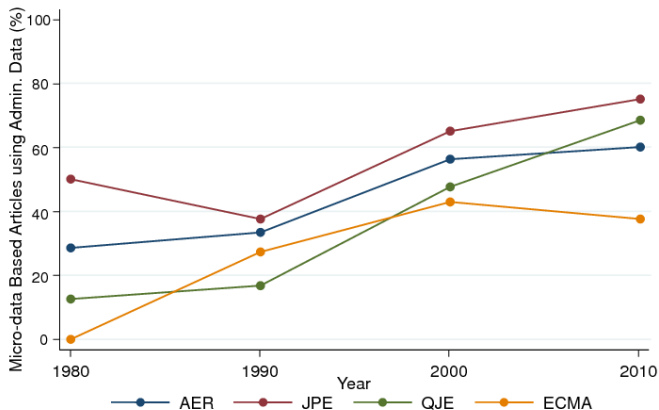
Use of Pre-Existing Survey Data in Publications in Leading Journals, 1980-2010



Note: "Pre-existing survey" datasets refer to major surveys such as the CPS, AHSIPP, and panel studies 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: The administrative datasets refer to any dataset that was collected without directly surveying individuals (e.g., scanner data, stock prices, school district records, social security records). Sample excludes studies whose primary data source is from developing countries.

Results from the LDI Replication Lab

Undergraduate research team

- ▶ Census of articles in the American Economic Journal: Applied Economics (2010, 2011, 2013)
- ▶ Each article is analyzed for availability of replication archive (as required by journal!)
- ▶ If data and programs are available, reproducibility is tested.

Some very preliminary results

Table: Replication Success

	Yes	No	Partial	Sum
2010	10	19	6	35
2011	12	20	4	36
2013	15	12	11	38
Total	37	51	21	109

Some very preliminary results

Table: Reason for Replication Failure

	Missing Data	Corrupted Data	Code Error	Missing Code	Sum
2010	15	1	1	2	19
2011	15	1	1	3	20
2013	12	0	0	0	12
Total	42	2	2	5	51

Some very preliminary results

Table: Reason for Missing Data

	Administrative			Private		Sum
	local	National	Regional	Commercial	Other	
2010	2	8	0	4	3	17
2011	2	8	4	1	0	15
2013	2	2	1	4	2	11
Total	6	18	5	9	5	43

Some very preliminary results

Table: Type of Access to Confidential Data

	Formal	Informal w/ Commitment	Informal w/o Commitment	No Info	Sum
2010	2	3	9	3	17
2011	2	0	10	3	15
2013	1	2	8	0	11
Total	5	5	27	6	43

Not limited to one journal

NIH-funded research

- ▶ article is open-access
- ▶ not clear about data access

A small anonymous example

Journal List > HHS Author Manuscripts > PMC3600



HHS Public Access

Author manuscript

Peer-reviewed and accepted for publication

About author manuscripts

Submit a manuscript

J Health Econ. Author manuscript; available in PMC 2014 Jul 1.

PMCID: PMC3600

Published in final edited form as:

NIHMSID: NIHMS388

[J Health Econ. 2014 Jul; 41: 600-600.](#)

Published online 2014 May 9. doi: [10.1016/j.jhealeco.2014.05.000](https://doi.org/10.1016/j.jhealeco.2014.05.000)

A small anonymous example

Journal List > HHS Author Manuscripts > PMC3600



HHS Public Access

key assumptions in the transition model and the health care cost model. A complete technical appendix containing details on the modeling is available online at <https://sites.google.com/site/pd/Home/programs>

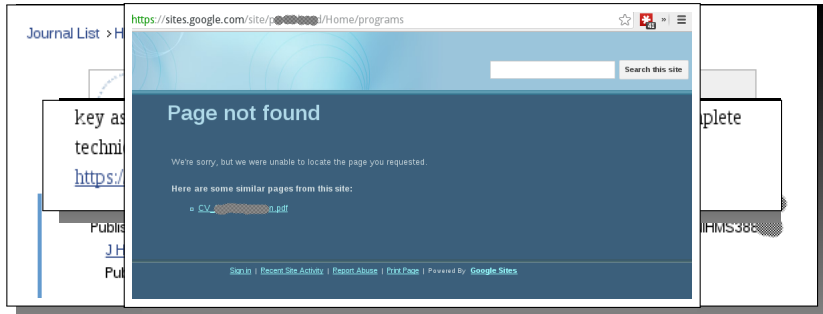
Published in final edited form as:

NIHMSID: NIHMS388

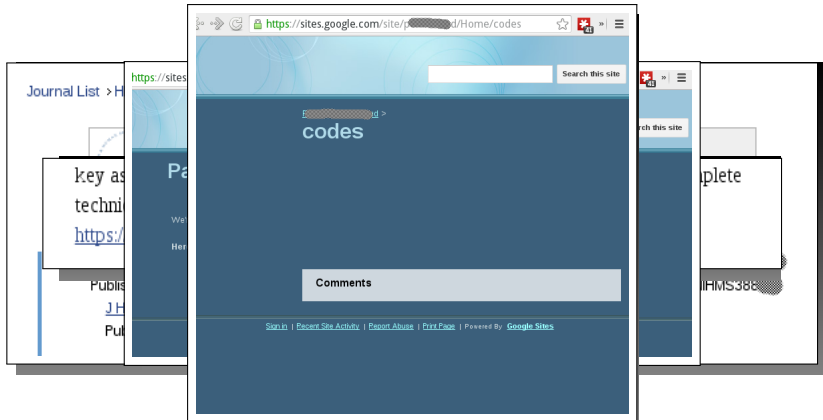
[J Health Econ. 2014 Jul; 41: 609-621.](#)

Published online 2014 May 9. doi: [10.1016/j.jhealeco.2014.05.001](https://doi.org/10.1016/j.jhealeco.2014.05.001)

A small anonymous example



A small anonymous example



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)

Non-federal confidential data

States, school districts, private companies, academic and private surveys: need a place to live to be re-used.

Options

- ▶ openICPSR <https://www.openicpsr.org/>
- ▶ Harvard Dataverse
<https://dataverse.harvard.edu/> (1,315 DV, 59,530 DS)
- ▶ Ontario Council of University Libraries:
<http://dataverse.scholarsportal.info/dvn/> (64 DV, 5,289 files)

Hinges on compatibility of data deposit rules, laws, regulations, etc.

Data citations

Examples

*Deschenes, Elizabeth Piper, Susan Turner, and Joan Petersilia. **Intensive Community Supervision in Minnesota, 1990-1992: A Dual Experiment in Prison Diversion and Enhanced Supervised Release** [Computer file].*

ICPSR06849-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2000. doi:10.3886/ICPSR06849

*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**", doi:10.7910/DVN/27923, Harvard Dataverse [Distributor], V2*

[src]

Data citations

Examples

*Deschenes, Elizabeth Piper, Susan Turner, and Joan Petersilia. **Intensive Community Supervision in Minnesota, 1990-1992: A Dual Experiment in Prison Diversion and Enhanced Supervised Release** [Computer file].*

ICPSR06849-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2000. doi:10.3886/ICPSR06849

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", doi:10.7910/DVN/27923, Harvard Dataverse [Distributor], V2

[src]

So we know how to deposit and cite data...

So we know how to deposit and cite data...

... except nobody does it...

We didn't do it...

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](https://doi.org/10.1016/j.jeconom.2010.09.008)

PMCID: P

NIHMSID: N

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

The Quarterly Workforce Indicators (QWI) are local labor market data produced and released every

We didn't do it...

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](https://doi.org/10.1016/j.jeconom.2010.09.008)

PMCID: N

NIHMSID: N

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

5. Abowd JM, Vilhuber L. The sensitivity of economic statistics to coding errors in personal identifiers 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

Abstract

The Quarterly Workforce Indicators (QWI) are local labor market data produced and released every

We didn't do it...

Abowd and Vilhuber (2011)

J Econom. Author manuscript; available in PMC 2012 Mar 1.

PMCID: N

Published in final edited form as:

NIHMSID: N

J Econom. 2011 Mar 1; 161(1): 82–99.

doi: [10.1016/j.jeconom.2010.09.008](https://doi.org/10.1016/j.jeconom.2010.09.008)

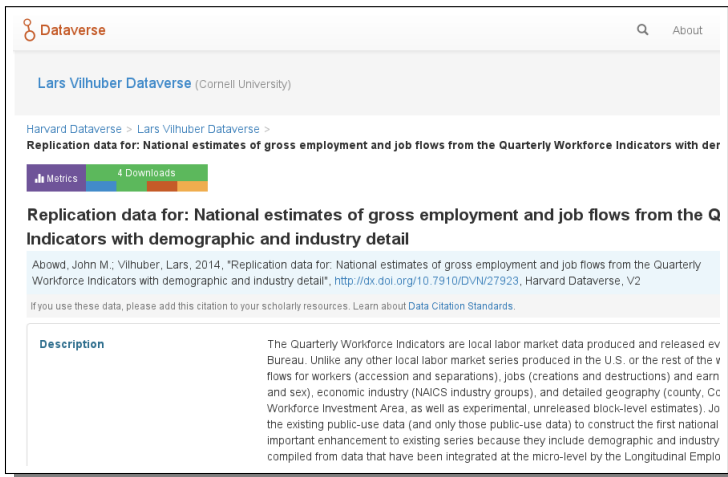
No confidential data were used in this paper. All public-use Quarterly Workforce Indicators data can be accessed from <http://www.vrdc.cornell.edu/news/data/qwi-public-use-data/>. The national indicators developed in this paper were accessed from <http://www.vrdc.cornell.edu/news/data/qwi-national-data/>. 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 end of the paper 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.

Abstract

The Quarterly Workforce Indicators (QWI) are local labor market data produced and released every

Then we archived it better...

... at Harvard Dataverse



The screenshot shows the Harvard Dataverse interface. At the top, the 'Dataverse' logo is on the left and a search icon with the word 'About' is on the right. Below the header, the user 'Lars Vilhuber Dataverse (Cornell University)' is listed. A breadcrumb trail reads 'Harvard Dataverse > Lars Vilhuber Dataverse >'. The main title of the dataset is 'Replication data for: National estimates of gross employment and job flows from the Quarterly Workforce Indicators with der'. Below the title is a progress bar with a 'Metrics' tab selected and a green bar indicating '4 Downloads'. The dataset description follows, starting with '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'. A citation instruction is provided: 'If you use these data, please add this citation to your scholarly resources. Learn about [Data Citation Standards](#).' The 'Description' section contains a paragraph about the Quarterly Workforce Indicators, noting they are local labor market data produced by the Bureau of Economic Analysis, unlike other series produced by the Bureau of Labor Statistics. The text is partially cut off at the end.

Dataverse

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

Then we archived it better...

... at Harvard Dataverse

 Dataverse


 About


Keyword	Employment Dynamics
Topic Classification	Economics
Related Publication	<p>John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Work 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/</p> <p>John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Work 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/</p> <p>John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Work with demographic and industry detail (with color graphs)," Center for Economic Studies, U.S. Census Bureau, Wc 11, 2010. http://ideas.repec.org/p/cen/wpaper/10-11.html</p>
Producer	Labor Dynamics Institute (Cornell University) (LDI) http://www2.vrdc.cornell.edu/news/data/qwi-national-data/ 

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

Then we archived it better...

... at Harvard Dataverse


About

Keyword	Employment Dynamics
Topic Classification	Economics
Related Publication	<p>John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Work 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/</p> <p>John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Work 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/</p> <p>John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Work with demographic and industry detail (with color graphs)," Center for Economic Studies, U.S. Census Bureau, Wc 11, 2010. http://ideas.repec.org/p/cen/wpaper/10-11.html</p>
Producer	Labor Dynamics Institute (Cornell University) (LDI) http://www2.vrdc.cornell.edu/news/data/qwi-national-data/ 

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

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 between them that express attribution, delegation, derivation, etc.

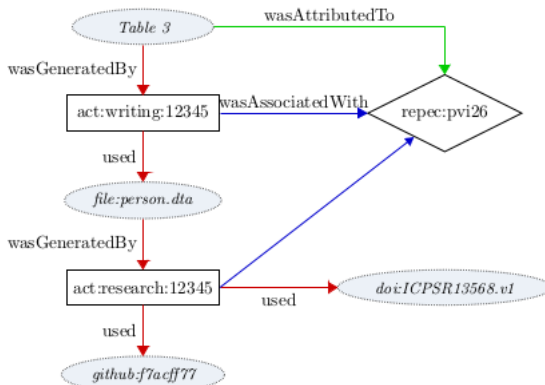
PROV and Metadata

Not (currently) a “native” component of DDI

Provenance for research

Sample research activity with full provenance

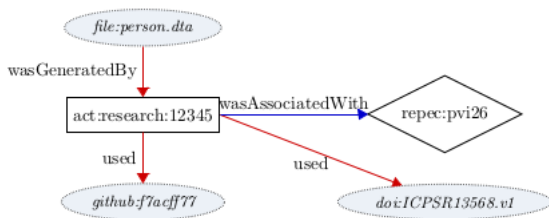
Figure 4: Sample research activity with full provenance



Provenance for research

Sample research activity with simple provenance

Figure 5: Sample research activity with simplified provenance



Merci

