An ex-post workflow documentation tool

Lars Vilhuber* Carl Lagoze† Ben Perry‡

April 24, 2015

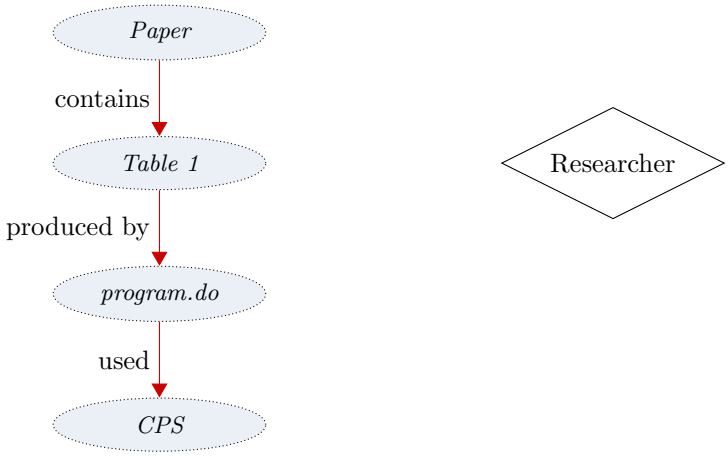
1 Introduction

Describe PROV.`http://www.w3.org/TR/prov-primer/` Describe CED^{2}AR. Describe problem.

The basic empirical scientific workflow is straightforward. Collect or acquire data, establish an analysis protocol and execute via statistical software, obtain results, and discuss results in the form of technical documents or published journal articles.

With this tool and associated implementation of a provenance-tracking document, we aim to capture the most frequent scenarios of that workflow. We start with the premise of an interested or co-erced user. For instance, take a researcher who is preparing the final manuscript for a journal that requires replication-ready archives of data and programs. If he hasn't done so up to this point, the researcher will start by collecting all the relevant pieces of data and code that make up the inputs to the paper. At its simplest, the paper has one table, that is produced by a single program (we assume that it's not produced by error-prone manual manipulations of data). That program read in some original data. See Figure 1.

Figure 1: Provenance



The goal of the software described in this paper is to "interview" the author about all the relevant components. Given the existing paper, what is the name, location, and purpose of the program that generated Table 1? What dataset served as input to the program? Where is that dataset located? And could you please describe in more detail the variables that make up your table?

*Cornell University, corresponding author. This work is funded by NSF Grant 1131848.
†University of Michigan
‡Cornell University

We choose a data-centric approach. Tying these concepts together, we leverage DDI for the data documentation (describe DDI, describe CED^{2}AR), and we leverage PROV (link to PROV) to describe the entities contributing to the program and the generated data. In its simplest instantiation, the input data is already well-catalogued, well-curated, and is browsable by DDI at a stable URL, possibly via a DOI.

We build on our prior work incorporating PROV into DDI. We now implement that specification as a software tool that can be used to visualize the workflow, and also provide all the necessary information to (potentially) upload the data and programs to a publication site.

Issues that are solved including identifying the researcher (agent), specifying roles.

The resulting DDI+PROV file is machine-readable, and complies with two widely used standards. An immediate application is the frequent task researchers encounter in restricted-access data environments. There, researchers need to document the data they are requesting release for in a way that the disclosure avoidance analysis (most often a human and not a program task) can understand. Typical documents require description of the data, of the programs, and where the data was sourced from - all elements of the process and documentation described here. (Census RDC, IAB, Synthetic Data Server).

2 Basic PROV of a simple workflow

3 Graph components

The completed file can be found as an attachment to this document (here).

3.1 Research activity

Agents

Agents undertake a research activity. Agents can be identified by external identifiers. In this example, we have identified users by their RePEc handle, and included their RePEc homepage.

```
<!--  AGENTS  -->
<prov:agent prov:id="repeca:pab175">
    <prov:type>prov:Person</prov:type>
    <foaf:givenName>John M. Abowd</foaf:givenName>
    <foaf:workInfoHomepage>https://ideas.repec.org/e/pab175.html</
       foaf:workInfoHomepage>
</prov:agent>

<prov:agent prov:id="repeca:pvi26">
    <prov:type>prov:Person</prov:type>
    <foaf:givenName>Lars Vilhuber</foaf:givenName>
    <foaf:workInfoHomepage>https://ideas.repec.org/e/pvi26.html</
       foaf:workInfoHomepage>
</prov:agent>
```

Entities

Entities are the datasets, programs, and articles that are being linked here. A subset are listed here.

```
<!--   ENTITIES     -->
<prov:entity prov:id="exn:article">
    <dc:title>A published paper</dc:title>
</prov:entity>
<prov:entity prov:id="RePEc:eee-econom-v-161-y-2011-i-1-p-82-99">
```

```
    <dc:title>National estimates of gross employment and job flows from
       the Quarterly Workforce
        Indicators with demographic and industry detail</dc:title>
    <dc:date>2011</dc:date>
</prov:entity>

<prov:entity prov:id="exn:program-unpub">
    <dc:title>An unpublished program</dc:title>
</prov:entity>
<prov:entity prov:id="file:home-spec555-path-to-program-do">
    <dc:title>Internal program</dc:title>
    <dc:date>2014</dc:date>
</prov:entity>

<prov:entity prov:id="exn:data">
    <dc:title>A data set</dc:title>
</prov:entity>
<prov:entity prov:id="file:home-spec555-path-to-file-dta">
    <dc:title>My dataset</dc:title>
    <dc:date>2014</dc:date>
</prov:entity>
```

Activities

We define a research activity to generate papers and data, and ultimately articles.

```
<!--    ACTIVITIES  -->
<prov:activity prov:id="act:research"/>
<prov:activity prov:id="act:research-12345"/>

<prov:activity prov:id="act:writing"/>
<prov:activity prov:id="act:writing-12345"/>
```

Linking them

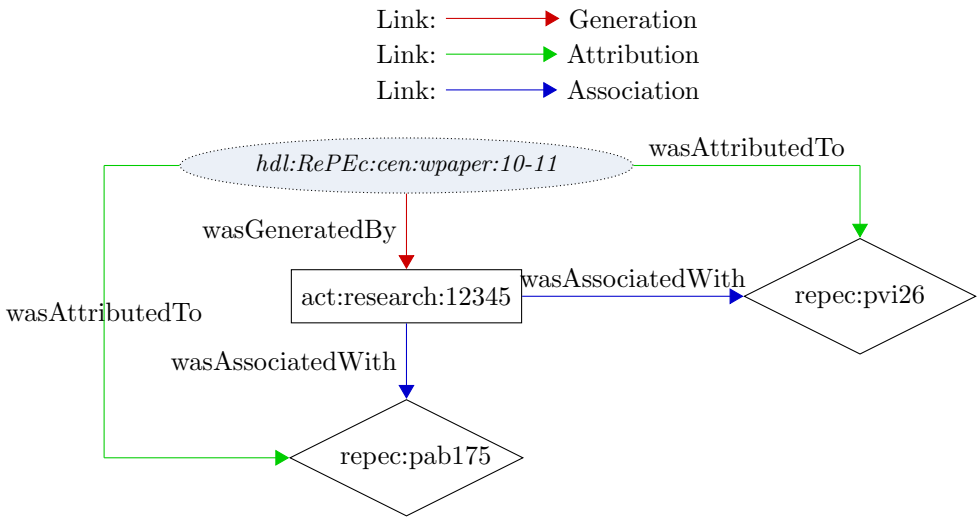
```
<!--    LINKS    -->

<prov:wasAssociatedWith>
    <prov:activity prov:ref="act:research12345"/>
    <prov:agent prov:ref="repeca:pab175"/>
    <prov:role>act:author</prov:role>
</prov:wasAssociatedWith>

<prov:wasAssociatedWith>
    <prov:activity prov:ref="act:research12345"/>
    <prov:agent prov:ref="repeca:pvi26"/>
    <prov:role>act:author</prov:role>
</prov:wasAssociatedWith>

<prov:wasGeneratedBy>
    <prov:entity prov:ref="RePEc:eee-econom-v-161-y-2011-i-1-p-82-99"/>
```

Figure 2: Authorship with research activity




```
    <prov:activity prov:ref="act:research12345"/>
</prov:wasGeneratedBy>
<prov:wasGeneratedBy>
    <prov:entity prov:ref="RePEc:cen-wpaper-10-11"/>
    <prov:activity prov:ref="act:research12345"/>
</prov:wasGeneratedBy>

<prov:wasAttributedTo>
    <prov:entity prov:ref="RePEc:eee-econom-v-161-y-2011-i-1-p-82-99"/>
    <prov:agent prov:ref="repeca:pab175"/>
    <prov:type>act:author</prov:type>
</prov:wasAttributedTo>

<prov:wasAttributedTo>
    <prov:entity prov:ref="RePEc:eee-econom-v-161-y-2011-i-1-p-82-99"/>
    <prov:agent prov:ref="repeca:pvi26"/>
    <prov:type>act:author</prov:type>
</prov:wasAttributedTo>
```

The straightforward research activity as traditionally focusing on papers would look like Figure 2.

4 Combining the subgraphs

Describe how we would link to other provenance chains.

Figure 3: Provenance back to data

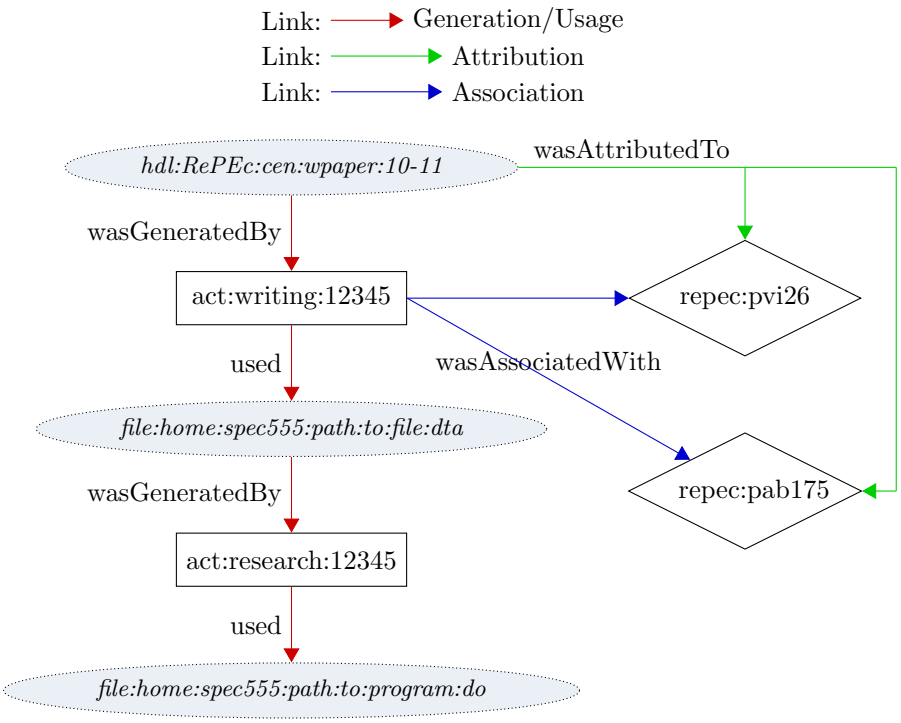
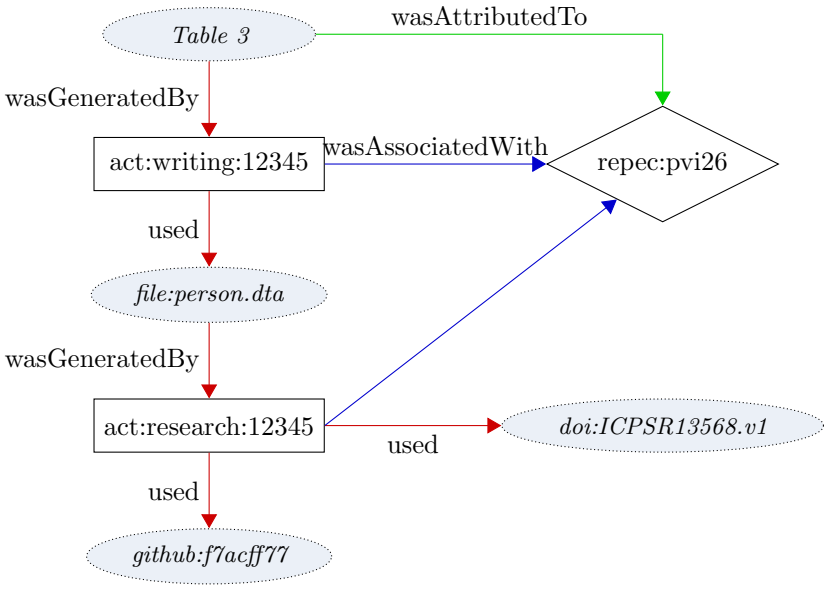


Figure 4: Sample research activity with full provenance



5 A simple example

To provide an example, we read in a portion of the U.S. Census Bureau's Census of Population and Housing Public Use Microdata Sample (Decennial PUMS). Specifically, we used the Alaska subset (file `DS2` or `DS0002`, depending on where it is referenced) [1]. The data file and a draft program were obtained from the Inter-university Consortium for Political and Social Research (ICPSR). After editing[1], the program `01_stata.do` (see Appendix A) was run to obtain a table with the distribution of those identifying with a particular Alaska Native tribe in the population, and as a fraction of those with some Alaska Native mention. To do so, we tabulate both `RACE2`, which lists a variety of groupings, but also lists 4 Alaska Native categories (`31-34`: Alaskan Athabascan, Aleut, Eskimo, Tlingit-Haida alone), and `RACE1`, which allows for either "4 - Alaska Native alone" or "5 - American Indian and Alaska Native tribes specifies, and no other races". We use `pweight` to construct the relevant table. The resulting table is Table 1.

Table 1: **Identifying with one of the four tribes**

Item	Number	Per cent
Not identified	554204.00	88.50
Identified with one of the four tribes	71,983.00	11.50
Total	626187.00	100.00

Source: person.dta

Expressed as PROV, the resulting provenance graph is represented by Figure 4. We choose to describe the entity "file:person.dta" (a dataset), rather than "Table 1", because the latter is a summary (a simple specialization) of the former. [need to express this in the PROV]. This allows us to simplify the graph somewhat more, without loss of generality, to Figure 5.

[1]The program as provided does not work, for two reasons: it is meant to be edited in its structure, and the commands included are sometimes erroneous.

Figure 5: Sample research activity with simplified provenance

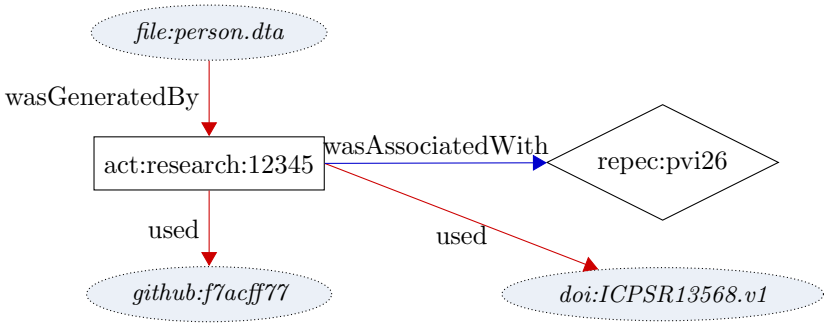
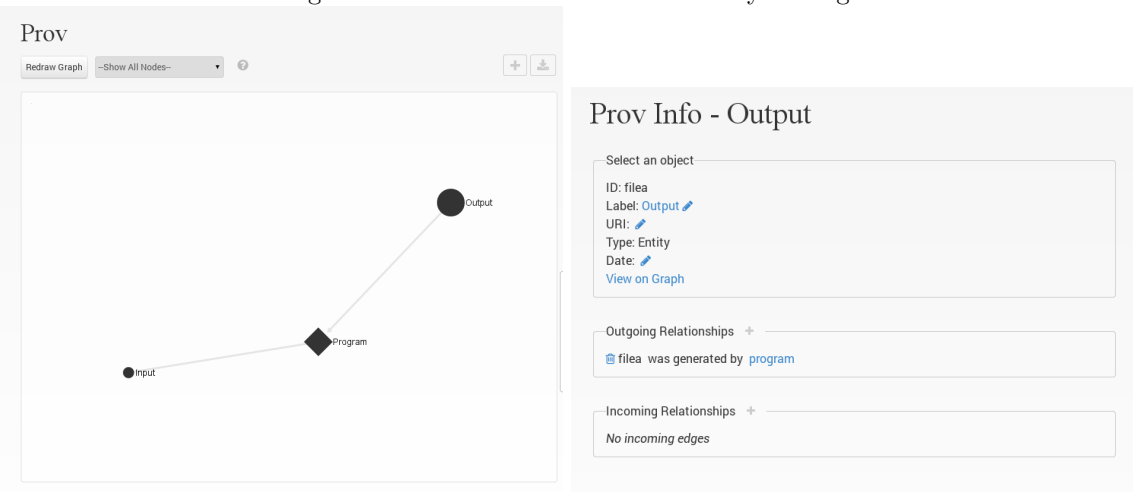


Figure 6: Initial user interface and entity editing



The full PROV associated with this graph is depicted in Appendix B.

6 An graphical interface

The graphical workflow tool elicits all the information needed in Figure 3 from the user in a flexible manner, starting with an initial blank template (Figure 6). The user can then edit the information on entities (Figure 6). Information on outside datasets can be prompted using selectors, for known or user-provided repositories, dynamically loading in information using supported protocols (OAI-PMH, DDI-Disco) (Figure 7). Information on internal entities (programs, unpublished datasets) are elicited from the user, using a file browser, Figure 7 (caching previous runs of the application will also be explored). . Once selected, the user is encouraged to provide a full documentation for any datasets through the integrated DDI editor. Additional entities can be added to make more complex workflows (Figure)The PROV serves both as a guide to the workflow tool, as well as a standards-compliant documentation of the provenance of the research article.

I don't know how to ge to that selec tor from a existing node

Figure 7: Selector for external and local entities

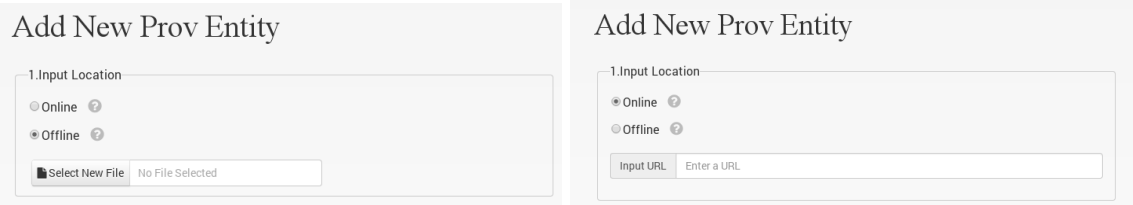
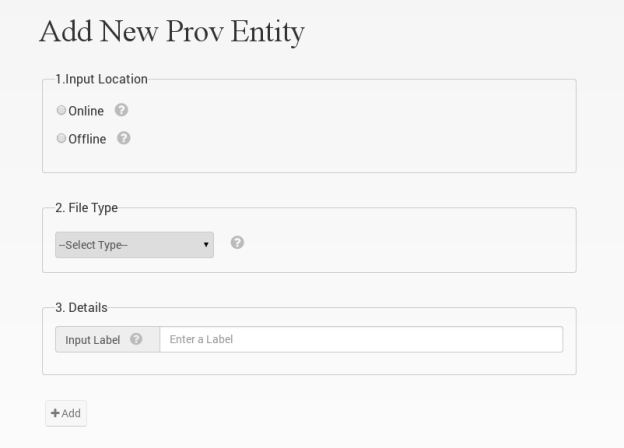


Figure 8: Adding entities



7 Future work

Already present in the RePEc network are citation links (linking papers among themselves). This work links in with CED^{2}AR work (citations) and other efforts for linking papers and articles to the data used for (empirical) papers. Establishing such links can be represented by a tripartite graph:

References

[1] U.S. Dept. of Commerce, Bureau of the Census. *CENSUS OF POPULATION AND HOUSING, 2000 [UNITED STATES]: PUBLIC USE MICRODATA SAMPLE: 5-PERCENT SAMPLE. ICPSR release.* Electronic data file. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. Washington, DC, 2003. DOI: 10.3886/ICPSR13568.v1. URL: http://doi.org/10.3886/ICPSR13568.v1.

A Stata program for readin of PUMS as used

The following program was run to obtain Table 1. The program can be run in batch mode (`stata -b do 01_stata.do`), and will generate the table included above dynamically. It was derived from the (non-functioning) program and layout provided at http://doi.org/10.3886/ICPSR13568.v1 (and archived at https://github.com/ncrncornell/workflow/tree/master/samples/ICPSR-PUMS/ICPSR_13568).

```
/* $Id: 01_stata.do 1259 2014-12-10 14:37:31Z lv39 $ */
/* This file reads in Alaska PUMS data */
/* SRC: http://doi.org/10.3886/ICPSR13568.v1 */
/* Source program:  "ICPSR_13568/13568-Setup.do" was used
   as a template, but that program cannot
   be run as-is */
/* Author: Lars Vilhuber */

  /* Define local macros, filenames and locations */
  local datpums "13568-0002-Data.txt"    /* PUMS Data */
  local datpath "ICPSR_13568/DS0002"  /* local relative path */
  local dtahu "housing.dta"  /* Stata Housing Unit data */
  local dcthu "housing.dct"  /* Stata Housing Unit dictionary */
  local dtap  "person.dta"   /* Stata Person data */
  local dctp  "person.dct"   /* Stata Person dictionary */
  local dtam  "pumsak.dta"  /* Stata PUMS merged data */

  capture log close
  log using pums, replace
  set more 1
  capture qui net install latab, from(http://fmwww.bc.edu/RePEc/bocode/l/)
  clear
  infile using `dcthu' if rectype=="H", using ("`datpath'/`datpums'")
  sort serialno   /* sort data by Serial Number */
  save `dtahu', replace  /* save housing unit data */

  clear
  infile using `dctp' if rectype=="P", using ("`datpath'/`datpums'")
  sort serialno   /* sort data by Serial Number */
  save `dtap', replace /* save person data */

  merge serialno using `dtahu' /* merge person and housing unit data */
  drop _merge
  /* keep only relevant information */
  keep pweight race2 race1 numrace
  /* code a dummy to the four tribes */
  gen specific_ak=(race2 == "31" | race2 == "32" | race2 == "33" | race2 == "34")
```

```
  /* convert weights */
  destring pweight, gen(pweight_num)
  /* label variables */
  label variable specific_ak "Identifying with one of the four tribes"
  label variable pweight_num "Person weight"
  /* table with appropriate weights */
  tab specific_ak [fweight=pweight_num]
  /* output the table to latex */
  label define spec_0 "Not identified" 1 "Identified with one of the four tribes"
  label value specific_ak spec
  latab specific_ak [fweight=pweight_num], tf("freq_specific_ak") replace dec(2)
  saveold 'dtam', replace  /* save merged data */
  log close
```

B PROV for simple example of Section 5

The following omits the first document declaration for clarity. The complete file is attached.

```
<!--   ENTITIES     -->
<prov:entity
  prov:id="github:f7acff773673289301c19a46789f25cb89d7b569">
  <dc:title>Program to readin and subset PUMS data</dc:title>
  <dc:date>2014</dc:date>
</prov:entity>

<prov:entity prov:id="file:person.dta">
  <dc:title>My temporary dataset</dc:title>
  <dc:date>2014</dc:date>
</prov:entity>
<prov:entity prov:id="doi:ICPSR13568.v1">
  <dc:title>CENSUS OF POPULATION AND HOUSING, 2000 {[UNITED STATES]}
     : {P}UBLIC USE MICRODATA SAMPLE: 5-PERCENT SAMPLE.</dc:title>
  <dc:date>2014</dc:date>
</prov:entity>
<!--  AGENTS  -->
<prov:agent prov:id="repeca:pvi26">
  <prov:type>prov:Person</prov:type>
  <foaf:givenName>Lars Vilhuber</foaf:givenName>
  <foaf:workInfoHomepage>https://ideas.repec.org/e/pvi26.html</
     foaf:workInfoHomepage>
</prov:agent>
<!--    ACTIVITIES  -->
<prov:activity prov:id="act:research"/>
<prov:activity prov:id="act:research-12345"/>
<!--    LINKS    -->
<prov:wasAssociatedWith>
  <prov:activity prov:ref="act:research12345"/>
  <prov:agent prov:ref="repeca:pvi26"/>
  <prov:role>act:author</prov:role>
</prov:wasAssociatedWith>
<prov:wasGeneratedBy>
  <prov:entity prov:ref="file:person.dta"/>
  <prov:activity prov:ref="act:research12345"/>
```

```
    </prov:wasGeneratedBy>
    <prov:used>
      <prov:activity prov:ref="act:research12345"/>
      <prov:entity
        prov:ref="github:f7acff773673289301c19a46789f25cb89d7b569"
      />
    </prov:used>
    <prov:used>
      <prov:activity prov:ref="act:research12345"/>
      <prov:entity prov:ref="doi:ICPSR13568.v1"/>
    </prov:used>
  </prov:document>
```

C Additional processing

For the purpose of this paper, we used OpenDataForge Sledgehammer to convert the Stata file produced by the program in Appendix A to DDI 2.5. In order to use the freeware version of Sledgehammer, we subset the file to 4999 observations.

```
 __  _           _
/ _\| | ___  __| | __ _  ___  /\  /\__ _ _ __ ___  _ __ ___   ___ _ __
\ \ | |/ _ \/ _` |/ _` |/ _ \/ /_/ / _` | '_ ` _ \| '_ ` _ \ / _ \ '__|
_\ \| |  __/ (_| | (_| |  __/ __  / (_| | | | | | | | | | | |  __/ |
\__/|_|\___|\__,_|\__, |\___\/ /_/ \__,_|_| |_| |_|_| |_| |_|\___|_|
                  |___/
OpenDataForge/SledgeHammer v2014.06.05.
Copyright Metadata Technology North America, 2012
Contact/Support dataforge@mtna.us | http://www.openmetadata.org/dataforge

FREEWARE EDITION
Issued:Fri May 23 16:13:24 EDT 2014
Expires:Sat Jun 06 20:00:00 EDT 2015
Max. #Variables:500
Max. #Observations:5000
```

Acronyms used

ICPSR Inter-university Consortium for Political and Social Research