Call for Cases of Data Reuse: Still Seeking Answers

May 17, 2016

Guest post by Stephanie Wykstra, Innovations for Poverty Action

As advocates for open data, my colleagues and I often point to re-use of data for further research as a major benefit of data-sharing. In fact there *are* many cases in which shared data was clearly very useful for further research. Take the Sloan Digital Sky Survey (SDSS) data, which researchers have used for nearly 6,000 papers. Or take Genbank, within bioinformatics, which is a widely used database of nucleotide and protein sequence data. Within social science, large-scale surveys such as the Demographic and Health Survey (DHS) are used by many, many researchers as well as policy-makers.

Research data re-use: where are the cases?

In spite of the obviousness of the value of data-sharing *in general*, I realized that we didn't have many cases of re-use of research data. By "research data" here, I have in mind data which were collected by an individual researcher or research team for their own project (e.g. from a field experiment), and then shared along with the publication. This differs from the databases like SDSS, Genbank and DHS in a few different ways:

- The data are often much smaller scale than DHS or SDSS; they are often studies of a few hundred to a few thousand subjects.
- They are not part of a unified data-gathering effort using common measures (as are SDSS and DHS), but rather use their own instruments, often with their own non-standardized measures.
- While it's fairly clear that researchers can use SDSS data for their own research, and bioinformaticists can use Genbank data, it's less clear how social scientists would re-use data that other researchers collected for the purpose of their own study. In general, they could use data for secondary analysis or meta-analysis; however, we haven't seen numerous examples.

A call for cases studies of data re-use

After a brainstorming session with Stephanie Wright, a colleague at Mozilla Science Lab, we decided to put out a call for cases of data re-use. For this project, we were particularly interested in cases of re-use within economics or political science. Since we support data-sharing among researchers and research staff, we want to be able to point to cases of real world re-use, and to delve into what made the data particularly useful. We wrote a post on our project, along with a survey on data re-use, and shared in venues such as IASSIST, Polmeth, Berkeley Initiative for Transparency in the Social Sciences' blog, Open Science

Collaboration's discussion board, Mozilla Science Lab's blog and various data librarian email lists.

What did we find through our call?

We received 14 responses to our call, including 10 responses to our survey and 4 emailed responses. While the number and quality of responses isn't sufficient for us to learn a great deal, we want to share what we found in any case, for two reasons: (1) This call and response could be informative to those who are considering putting out a similar survey and (2) we think our findings do provide some evidence which confirms our initial feeling, which is that this is an area which warrants further work and research.

Our 10 survey respondents are in a variety of fields: one in political science, two in psychology, one in education and most of the rest in biochemistry. While all respondents did mention some data that were re-used, only three gave examples of the kind that we had requested e.g. data that had been collected by other researchers for their study, and then re-used for further research. The three cases included:

- Re-use of data from a collaboration of Psychology instructors, which collected data on emerging adulthood and politics. The data were not initially used for a publication, as intended, but were archived and were used for nine published articles later on.
- A researcher in political science gave several of his own research re-use cases in which publicly available data were used for a) a replication to "illustrate the usefulness of a new fit assessment technique for binary DV [dependent variable] models," b) for pedagogical purposes in a book on causal inference and c) to test a new theory.
- Researchers in psychology used data from two large-scale studies on the benefits and transfer effects of a cognitive training for older adults. The data were used to test whether a subset of one test (the Useful Field of View test) were able to predict scores on another test (the Instrumental Activities of Daily Living test).

Beyond the cases above, we heard about re-use of protein sequence data and genomics data from databases such as ArrayExpress and Protein Data Bank, as well as government data from Open Data Toronto and Statistics Canada. See our spreadsheet for further details (we asked for permission to share responses).

In addition to the cases gathered through our survey, we received four emails with tips about where to find additional cases. One of the suggestions mentioned the Global Biodiversity Information Facility (GBIF), a database on global biodiversity, as well as International Polar Year (IPY), a coordination of research on the Polar regions. A second suggestion from a political scientist pointed to several sites, Uppsala Conflict Data Program and the Correlates of War Program. Both sites offer data which are widely used by scholars within international relations, and include variables which are constructed by scholars for their own research, and then submitted to the databases for others to re-use.

Finally, we received several suggestions from fellow open data advocates, of places to look for cases of re-use. The first source, Dissemination Information Packages for Information Re-use (DIPIR) is a study of data re-use in three communities (quantitative social scientists, archaeologists, and zoologists). The second is ICPSR's bibliography of data-related literature, which is a searchable database of "over 70,000 citations of published and unpublished works resulting from analyses of data held in the ICPSR archive." The third is UK Data Archive's list of case studies of data re-use.

Data re-use: key for rewarding data-sharing

The data-sharing movement is gaining steam. From funders requiring data-sharing to new guidelines for journals (TOP guidelines) and journal requirements, to the rise of many data repositories, there is plenty of effort going into requiring and supporting data-sharing. Yet there are huge issues to confront, as we move forward. One of the biggest is how to change from a culture in which data-sharing is not a norm among researchers (as is still the case in many scientific fields) to one in which it is.

Researchers are rewarded for publishing, not for sharing data, and many researchers cite barriers to sharing data such as lack of time and lack of support (Tenopir et al. 2011). How will we shift to rewarding researchers for sharing their data, so that they have professional incentives to take the time to prepare and share data? One of the most-discussed ways is to develop good data-citation norms, and then to reward researchers (via tenure committee decisions) when others re-use and cite their data.

So, the question of how to promote and encourage data re-use is of clear importance. Yet, as practitioners in the open science movement, we have many questions. When it comes to re-using data from colleagues' studies, particularly in the social sciences, what factors make datasets particularly helpful to researchers? What challenges arise in re-using data? As data curators and open data advocates, what could we do better to facilitate re-use? Is there something we can do to encourage others to look at and reuse existing data when they are considering new research projects? How can we increase opportunities for re-using data and decrease barriers?

Next steps

Particularly when it comes to data shared by researchers in the social sciences, we still need more examples of re-use. We also need much more investigation into what would make researchers more likely to re-use data from colleagues for their own research. We can think of a couple of interesting projects that we could undertake:

- Delving into the archives from ICPSR and UK Data Archive, as well as others mentioned above, and attempting to glean lessons from specific cases of re-use found there.
- Contacting researchers that have downloaded data from archives such as IPA's data repository (we track data users and ask them permission to contact them, when they download data). We could gather more detailed information about what was or wasn't helpful for re-use about the data and other materials as presented in the repository. We could also try to gather more information on whether data were re-used for further research (and if so, what made them particularly attractive for re-use).

We're certainly open to further suggestions, so please get in touch if you have ideas!

Stephanie Wykstra directs the Research Transparency Initiative at Innovations for Poverty Action, and also works as an independent research consultant. She may be contacted at stephanie.wykstra@gmail.com

2016 Leamer-Rosenthal Prizes for Open Social Science

BITSS Bits: Repro at UC Davis and the World Bank, Job Openings

One thought on "Call for Cases of Data Reuse: Still Seeking Answers"

1. Pingback: Definitions of open science - Open-Science BlogOpen-Science Blog

Leave a Reply

University of California, Berkeley
207 Giannini Hall
Berkeley, California 94720-3310

Center for Effective Global Action