Yale Institution for Social and Policy Studies

A Repository on a Mission: A Small Research Community Gets Serious About Reproducibility

Limor Peer

Yale University, Institution for Social and Policy Studies

The ISPS Data Archive

The ISPS Data Archive is a digital repository for social science research produced by scholars affiliated with the Institution for Social and Policy Studies at Yale University, with special focus on experimental design and methods.

The Archive was launched in September 2010 as a pilot for Yale's Office of Digital Assets and Infrastructure (ODAI) to find solutions relating to storage, persistent linking, long-term preservation, and integration with a developing institutional repository.

Objectives

1. Provide data and code to enable replication of research results
2. Easy for Researchers
3. High quality data and metadata
4. Preservation and stewardship within Yale environment
5. Fully integrated with ISPS website

"A journal article describing the results of scientific work is typically a distillation of experimental data aimed at a wider audience than the immediate peers of the authors. Generally, inferences are made only from the most pertinent results, which are reported in a summary format, and journal publication is detached from the production of the experimental data. This renders replication or reuse of the data impossible and results in severe information loss." (Coles et al, 2007)

ISPS website

Research area on website

The Process

To make research reproducible, data require certain enhancements, which add value to the data. Before data publication, staff process data and code files, including: adding metadata, verifying replication, creating a DDI-XML codebook, and converting to CSV and R formats. To date, the ISPS Data Archive has published over 750 files for about 45 studies.

The Bottom Line(s)

1. Labor-intensive process to meet all objectives; currently, there is no single streamlined workflow that addresses all the objectives.
2. Offering open access to data for the purpose of replication requires hands-on attention at every step (looking to automate technical steps).

ISPS Data Archive Workflow

PI

- Transfer files
- Sign deposit agreement
- Fill out study- and file-level metadata

Archive Manager

- Record in status doc
- Start process record
- Start metadata doc (study-, file - level)
- Record in master metadata doc
- Create shared folder with files + process record + article
- Copy in local hard drive
- Copy in ODAI server (private)

RA

- Check variables and labels
- Check for sensitive information
- Check for missing files
- Rename files
- Create XML doc; variable-level
- Check replication
- Convert to ASCII, PDF, R...
- Update process record
- Upload to shared folder

Archive Manager

- Communicate with PI as needed
- Update metadata record

ODAI

- Assign handles to each file
- Archive public files on ODAI server (public)

Archive Manager

- Update status document
- Update process record
- Update metadata record
- Update master metadata document
- Create web entry
- Notify PI

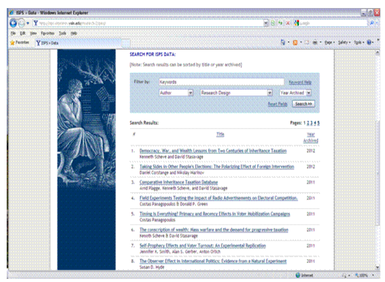
Focus on Reproducibility

Reproducibility, or replication, is the sharing of data and programs by researchers so that "another researcher using the same methods should be able to reach the same results" (Anderson et al, 2008). Advances in digital computing make it possible, and imperative, for researchers to allow access to their data, parameters, and programs (Peng, 2009, 2011; Stodden, 2009, 2011).

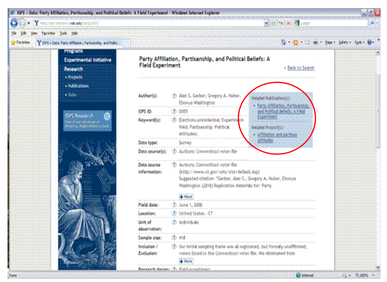
Replication can be a key motivator for researchers to share data.

Replication presents data sharing (and preservation) not only with a concrete purpose, it prescribes that steps be taken to ensure that (a) the right materials are shared and that, (b) they can be properly used (see diagram).

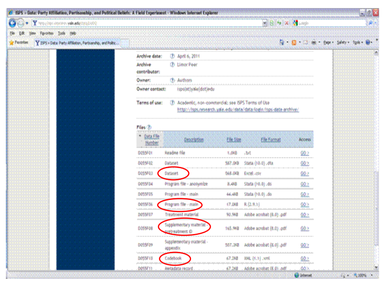
These steps should be taken by the entity that assumes responsibility over the data (e.g., a repository, a journal, funder website, etc.), and they are an essential part of data curation.



Gateway to data archive



Linking projects, publications, and data



Adding value

References

Anderson, R.G., Greene, W.H., McCullough, and Vinod, H.D. (2008). The Role of Data/Code Archives in the Future of Economic Research. *Journal of Economic Methodology*, 15(1): 99-119.

Coles, S., Carr, L., & Frey, J. (2007). R4L: The repository for the laboratory. *R4L Final Report*, v0.1. Retrieved from http://r4l.eprints.org/publications/docs/R4Lfinalreport.pdf

Peng, R.D. (2009). Reproducible research and Biostatistics (with discussion). *Biostatistics 10*(3).

Peng, R.D. (2011). Reproducible research in Computational Science. *Science 334*(6060). Retrieved from http://www.sciencemag.org/content/334/6060/1226

Stodden, V. (2009). Enabling reproducible research: Licensing for scientific innovation. *International Journal of Communications Law and Policy 13*. Retrieved from http://www.ijclp.net/files/ijclp_web-doc_1-13-2009.pdf

Stodden, V. (2011). Trust your science? Open your data and code! *Amstat News*. Retrieved from http://magazine.amstat.org/blog/2011/07/01/trust-your-science/

This work is supported by the Institution for Social and Policy Studies and the Office for Digital Assets and Infrastructure at Yale University. Thanks also go to Yale's StatLab and ISPS Data Archive staff.