Data Sharing in Social Science Repositories: Facilitating Reproducible Computational Research

Victoria Stodden
Department of Statistics
Columbia University
New York, NY 10027
vcs@stanford.edu

Abstract

From new types of data to new computational methodologies, computation is engendering a revolution in social science research and with this comes the issue of facilitating data and code sharing to encourage collaboration and reproducibility in scientific publishing. A repository designed for this purpose at Harvard University, The Dataverse Network, permits authors to upload data and code with their own terms of use. This paper examines these terms of use for $30,090$ uploads to discover barrier issues to sharing in the social sciences and compares them to those found in a survey of NIPS registrants. We find that the additionally specified terms of use in The Dataverse Network primarily address issues of maintaining subject confidentiality, preventing further sharing, making specific citation a condition of use, restricting access by commercial or profit-making entities, and time embargoes, which differs to those elucidated among NIPS participants. Using these findings we suggest a sharing framework for social science data to expand engagement of the larger social science community and encourage verification of research findings.

1 Introduction

Computation is emerging as central to the scientific enterprise (see e.g. [1], [2], [3]), and issues of code and data sharing in scientific publication have consequently become of deep importance to scientific integrity [4], [5], [6]. This phenomena of sharing is now engendering new and important questions regarding collaboration and social computing in research settings. The accommodation of a centuries old aspect of the scientific method – independent reproducibility of published results – in the age of digital science poses new challenges regarding understanding the conditions that best facilitate code and data sharing. Producing reproducible research in the new deeply computational context often requires the additional step of sharing the underlying data and code [7]. A movement is underway to increase transparency in computational science and ensure the replicability of findings [8], [9].

In 1995 Gary King defined the *Replication Standard*, advocating the sharing of research information beyond that included in the traditional publication: "The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author." [10]

As repositories are created to house scientific code and data and increase replicability (see `http://mloss.org` for an example in the Machine Learning community, or `http://thedata.org` for a social science example), a natural experiment can emerge regarding author-imposed sharing and re-use conditions for uploaded code and data. These datasets may provide a natural setting for the application of natural language processing techniques for identifying the author concerns, and

allow us to understand barriers to data and code sharing in the computational science community. At the heart of this paper are questions of facilitating data and code sharing by examining the conditions placed on shared data and code. To answer this inquiry we undertake an examination of the $30,090$ uploads to Harvard's Dataverse Network, each with potential terms of use and user-contributed restrictions and conditions.

2 Rationale, Data, and Analysis

Our approach builds on recent survey work of the NIPS community that sought to understand factors that facilitate and inhibit release of research data and code on the web [11]. We extend that effort to the setting of a collaboratively-built repository, analyzing responses of researchers who deposit code or data and seek to validate results from the NIPS survey. The NIPS survey work found the top four reasons not to share data to be (in order), 1) the time it takes to clean up and document, 2) fear of omitted citation in data re-use, 3) legal barriers, 4) the verification of privacy or administrative concerns. The issue of reproducibility in computational science has garnered wide attention recently, with efforts to create repositories that facilitate deposit and replicable science [5]. At Harvard University, the Institute for Quantitative Social Science (IQSS), a repository called The Dataverse Network project (http://thedata.org) is maintained with the stated purpose [12]:

> To enable data archiving and preservation through re-formatting, standards and exchange protocols.
>
> To provide control and recognition for data owners through data management and persistent citations.

A researcher is free to upload papers, data, and/or code, and a web interface is provided to share research data and increase scholarly recognition. From its webpage, "The Dataverse Network is an application to publish, share, reference, extract and analyze research data. It facilitates making data available to others, and allows others to replicate work. Researchers and data authors get credit, publishers and distributors get credit, affiliated institutions get credit. ... A Dataverse Network hosts multiple dataverses. Each dataverse contains studies or collections of studies, and each study contains cataloging information that describes the data plus the actual data and complementary files." [13]

When uploading to the repository, authors are able to list additional terms of use through an html text box, creating a dataset of authors' concerns regarding the sharing of their data and code. As of May 2010, there were $30,090$ Dataverses in the network, each with its own terms of use captured in four author entered fields: "special permissions," "citation requirements," "conditions of use," "restrictions on use." We hypothesize that the same concerns regarding data sharing as expressed in the NIPS survey will be at the forefront of those expressed by the depositors in The Dataverse Network.

The R software package was used to extract the unique entries in all four fields. The majority of entries in The Dataverse originate with a small group of uploaders, for example the National Archives and Records Administration, Harvard's Inter-University Consortium for Political and Social Research (ICPSR), the Roper Center for Public Opinion, the Henry A. Murray Research Archive, the Harvard Geospatial Library, or The University of North Carolina at Chapel Hill's Odum Institute for Research in Social Science, and many conditions of use were repeated within these groups. To show the reduction in analyzable observations, Table 1 gives the number of unique entries by field.

Table 1: Unique entries by Dataverse field

FIELD LABEL	COUNT
special permissions	15
citation requirements	33
conditions of use	36
restrictions on use	63

Table 1 indicates an enormous amount of data reduction due to duplicate records. Table 2 lists researcher concerns that emerged in the remaining unique entries.

Table 2: Data sharing concerns added by researcher

No further sharing; no reselling
Request form submission and approval required
No identification of individuals in dataset; report if this happens
When sharing in collaboration all files must be shared together
All resulting papers must be remitted to original researchers
No follow-up on dataset permitted; or follow-up permitted only through the original researcher
On-site use only
Home institute use only
Non-commercial use only
Not to be use for any profit-making purpose
Time embargo on data
Use by scholars only
Social science and behavioral research only
Original researcher must be informed about mistakes in the data
No linking the data to GPS data; or any datasets that would permit individuals to be identified
Derived works that include the original data must ensure the data are unrecoverable to a third party
College affiliated use only
Data are to be used only for health statistics reporting and analysis
The metadata file must accompany all dataset transfers
Educational non-commercial use only

These concerns fall into several broad categories: maintaining subject confidentiality, preventing further sharing, making a specific citation form a condition of use, restricting access by commercial or profit-making entities, and restricting use to a specific community, such as that of the researcher's home institution. The most frequently cited concern was that reuse of the data was to be limited to the academic sphere: for example scholarly use, college use, educational use, home institute use only. The main concerns of the NIPS survey respondents are quite different and reflect mainly the work involved in readying data for release and garnering appropriate citation. These differences may be due to the largely institutional nature of the datasets sharing through The Dataverse Network. The majority of the data are from a small number of institutional repositories such as Harvards Murray Research Archive and the University of Michigans Inter-University Consortium for Political and Social Research (ICPSR). Further, this analysis does not include barriers to sharing for unshared data, as the NIPS research does.

3 Conclusions and future work

In repository design, the concerns of the researcher must be balanced against the concern for progress of science in general. In the social sciences setting, these data give a clear indication of concern for individual identification, along with citation, and control over re-use. Perhaps surprisingly many researchers sought to prevent the deposited data from being used in a money making endeavor. Embedding these controls at the repository level could encourage deposit, and reproducibility and collaboration, by assuaging concerns of the researcher.

The dimensionality of this dataset would benefit from being increased through linking with metadata from other repositories. A larger dataset would provide a setting for the application of machine learning techniques to identify and assess researcher concerns, especially in this case where there are a fairly small number of underlying factors. The Dataverse dataset could provide a test case for machine learning methodologies, to see whether the underlying factors chosen by the algorithm match those emerging from the plain text evaluation.

Research on factors that underly data and code sharing concerns serves to permit the design of sharing frameworks compatible with researcher interests and incentives, as well as the scientific method. Focusing on the datasets emerging from repositories permits the greater understanding of sharing behavior among scientists. Identifying these underlying factors better permits for greater data generation, more widely usable data, and published findings that can be verified by others.

Acknowledgments

The author would like to thank Gary King, Ellen Kraffmiller, and Merce Crosas at The Dataverse Network both for encouragement and for kindly providing the data.

Appendix: Overview of Legal Barriers to Reproducible Computational Science

A motivating factor behind the examination of research sharing barriers is the hope that concern could be mitigated through the careful design of research sharing modalities. One pervasive bar to reproducible research is Intellectual Property law, particularly through copyright and patents.

Copyright and the Scientist: Copyright adheres by default to original expressions of ideas, with some exceptions and limitations. In the context of scientific research this means that code and articles are copyright to their authors, and any original selection and arrangement of data may be. Copyright acts counter to longstanding scientific norms by creating a barrier to the copying of, say, code and to its modification and re-use unless permission is granted by the copyright holder. Open licensing for science, such as that recommended by the Reproducible Research Standard [14] provides a means to realign the legal environment for scientific works with scientific norms.

Patents and Science: Since the passage of the Bayh-Dole Act in 1980, universities have encouraged the patenting of subject inventions developed in the academic setting, such as certain computer codes. This creates a licensing barrier to reproducible research, as well as encouraging deliberate opacity of the research while the patent is pending.

References

[1] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Life in the network: the coming age of computational social science." Science, 323(5915), 2009.

[2] P. Torrens. "Geography and computational social science," GeoJournal, 75(2), 2010.

[3] M. Szella, R. Lambiotte, and S. Thurner, "Multirelational organization of large-scale social networks in an online world," PNAS, 107(31), 2010.

[4] J. Buckheit and D. Donoho. "Wavelab and reproducible research," Technical report, Stanford University, 1995.

[5] The Yale Roundtable on Data and Code Sharing, 2009. `http://www.stanford.edu/~vcs/Conferences/RoundtableNov212009/References.html`. See especially references on reproducibility in computational science.

[6] Z. Tracer, "Nevins requests retraction of key genomics research paper," The Chronicle, Duke University, Nov 22, 2010. See `http://dukechronicle.com/article/nevins-requests-retraction-key-genomics-research-paper`

[7] D. Donoho, A. Maleki, I. Ur Rahman, M. Shahram, and V. Stodden, "Reproducible Research in Computational Harmonic Analysis," Computing in Science and Engineering Magazine, 11(1), 2009. `http://www.computer.org/portal/web/csdl/doi/10.1109/MCSE.2009.15`

[8] See e. g. `http://gking.harvard.edu/projects/repl.shtml`

[9] D. Donoho, " An invitation to reproducible computational research," Biostatisics, 11(3), 2010. `http://biostatistics.oxfordjournals.org/content/11/3/385.extract`

[10] G. King, "Replication, Replication," PS: Political Science and Politics, 28(3), 1995.

[11] V. Stodden. The scientific method in practice: reproducibility in the computational sciences. `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1550193`

[12] http://thedata.org/home

[13] From http://thedata.org/book/learn-about-project

[14] V. Stodden, "Enabling Reproducible Research: Licensing for Scientific Innovation," International Journal of Communication Law and Policy, Issue 13, 2008. http://www.ijclp.net/issue_13.html