Response to National Institutes of Health Request for Information: Input on Development of a NIH Data Catalog

Submitted by the Data Preservation Alliance for Social Science (Data-PASS)

June 25, 2013

Introduction to Data-PASS

The Data Preservation Alliance for the Social Sciences (http://Data-PASS.org) is a broad-based voluntary partnership of data archives dedicated to acquiring, cataloging, and preserving social science data, and to developing and advocating best practices in digital preservation. The Data-PASS partners collaborate to acquire data at risk of being lost to the research community; to develop preservation practices; and to create open infrastructure for collaborative cataloging and preservation.

Collectively, the founding partners have over 200 years of combined experience in social science data archiving. These partners include the Inter-university Consortium for Political and Social Research, The Roper Center for Public Opinion Research, The Howard W. Odum Institute for Research in Social Science, the Electronic and Special Media Records Service Division, National Archives and Records Administration, the Institute for Quantitative Social Sciences at Harvard University (which contains both the Harvard-MIT Data Center and the Henry A. Murray Archive), and the Social Science Data Archive at the University of California, Los Angeles (UCLA).

Thus far, the partnership has identified thousands of at-risk research studies (collections of data) and acquired many of these for permanent preservation. These range from data collections created under NSF (National Science Foundation) and NIH (National Institutes of Health) grants, to surveys conducted by private research organizations, to state-level polling data, to data records created by governmental research or administrative programs. [Gutmann, et *al*, 2009]

A National Digital Stewardship Alliance Founding Member, the Data-PASS partnership works to archive social science data collections at-risk of being lost; to catalog and promote access to data collections; to establish verifiable multi-institutional collaborative replication and stewardship of data; and to develop and advocate best practices in digital preservation.

Data-PASS Experience with Data Citation

Over the last decade the Data-PASS partners have been early adopters of data citation. The Virtual Data Center [Altman, et al 2001], created by one of the partners in 1999, was the first digital library system to systematically support permanent data citation. Data-PASS developed a proposed data citation standard [Altman & King 2007]; and a set of data citation principles emerging from an expert workshops [Altman 2012] Currently, the partners, enabled by open infrastructure such as the Dataverse Network [King 2007], uniformly provide persistent, verifiable data citations for all content [See] http://data-pass.org/citations.html.

The Data-PASS partners are now involved in a number of projects advancing data citation. These include a Sloan-funded project run by ICPSR to work with journals, funders and repositories to build community engagement in data citation and open access to data [See http://openscholar.web.itd.umich.edu/] and a project run by IQSS to integrate open source software for open access journal publication to support open data and data citation. [See http://projects.iq.harvard.edu/ojs-dvn]

Common Data Citation Principles and Practices

While there are a number of different communities of practice around data citation, a number of common principles and practices can be identified.

Two high-level principles for the use of data citations have emerged. The editorial policy *Science* [see http://www.sciencemag.org/site/feature/contribinfo/prep/] is an exemplar of two principles for data citations: First, that published claims should cite the evidence and methods upon which they rely, and second, that things cited should be available for examination by the scientific community. These principles have been recognized across a set of communities and expert reports, and are increasingly being adopted by number of other leading journals. [See Altman 2012; CODATA-ICSTI Task Group, 2013; and http://www.force11.org/AmsterdamManifesto]

Implementation Considerations

Previous policies aiming to facilitate open access to research data have often failed to achieve their promise in implementation. Effective implementation requires standardizing core practices, aligning stakeholder incentives, reducing barriers to long-term access, and building in evaluation mechanisms.

A set of core recognized good practices have emerged that span fields. Good practice includes separating the elements of citation from the presentation; including in the elements identifier, title, author, and date information, and where at all possible version and fixity information; and listing data citations in the same place as citation to other works – typically in the references section. [See Altman-King 2006; Altman 2012; CODATA-ICTI Task Group 2013; http://schema.datacite.org/ ; http://data-pass.org/citations.html]

Although the incentives related to data citation and access are complex, there are a number of simple points of leverage. First, journals can both create positive incentives for sharing data by requiring that data be properly cited. Second, funders can require that only those outputs of research that comply with access and citation policies can be claimed as results from prior research.

To facilitate evaluation, the metadata underlying any NIH data catalog should be available open-access license and through an open API. Further, data citations produced under NIH policy be should include persistent identifiers that are compatible with and indexable by emerging cross-disciplinary catalogs, such as DataCite [http://datacite.org], the Dataverse Network [http://thedata.org], and the Data Citation Index [http://wokinfo.com/products_tools/multidisciplinary/dci/]

To support appropriate granularity of citation, citation standards should support the finest grained description necessary to identify the data. [Codata Task Group 2013] This can be done with simple extensions to the citation or citation metadata. Note that rule does not require that separate persistent identifiers (such as DOIs) be created for fine grained citation – merely that there be some way of unambiguously identifying the portion used within the data cited.

Data citation lacks much of its value unless there are mechanisms to provide long-term access to cited material. [See Altman & King 2007] Data cited should be made available through institutions with demonstrated capability to provide long-term access. And where full access is not possible because of externally-imposed data restrictions (such as confidentiality restrictions) should continue to be made available for research through a variety of modes, including full access to original data under appropriate license and security restrictions, and open access to data altered to maintain confidentiality. [See NRC 2005,2009; Vadhan 2011]

Finally, to ensure that access is fully meaningful, access to data should include access to the information necessary to support both "production transparency" and "analysis transparency" -- to enable, where possible reproduction of the data collection and of the analysis that support published conclusion.

Summary of Recommendations

To maximize the utility of an NIH data catalog, and to align incentives for researchers to register data the NIH should require that:

- Articles published under NIH funding provide data citations to any evidence required to verify or understand the claims published. And these citations should at minimum contain a persistent identifier, title, author, and date.
- The metadata associated with NIH funded data and data citations be available under an open-access license and through an open API, and the identifiers compatible with and indexable by emerging cross-disciplinary catalogs.
- Proposals for future NIH funded research be able to list any relevant data produced by the Investigators, under the condition that such data follows NIH citation and access policies.
- Data cited should be made available through institutions with demonstrated capability to

provide long-term access. And where full access is not possible because of externally-imposed data restrictions (such as confidentiality restrictions) should continue to be made available for research through a variety of modes, including full access to original data under appropriate license and security restrictions, and open access to data altered to maintain confidentiality.

Respectfully submitted on behalf of the DataPASS by it's steering committee.

George Alter, ICPSR; Micah Altman, MIT; Mark Abrahamson, Roper Center; Merce Crosas, Harvard U. Jon Crabtree, Odum Institute; Ted Hull, NARA ; Gary King, Harvard; William LeFurgy, Library of Congress; Amy Pienta, ICPSR; Libbie Stephenson, UCLA

[Submitted via email to data-catalog@mail.nih.gov , on June 25, 2013. Under heading: Response to NOT-HG--13-011]

References

M. Altman and G. King. 2007. "A Proposed Standard for the Scholarly Citation of Quantitative Data", D-Lib, 13, 3/4 (March/April).

M. Altman 2012. Data Citation in The Dataverse Network ®,. In Developing Data Attribution and Citation Practices and Standards: Report from an International Workshop.

CODATA-ICSTI Task Group on Data Citation Standards and Practices, (Forthcoming 2013), *Citation of Data: The Current State of Practice, Policy, and Technology,* CODATA.

M. Gutmann, Abrahamson, M., Adams, M., Altman, M., Arms, C., Bollen, K., Carlson, M., Crabtree, J., Donakowski, D., King, G., Lyle, J., Maynard, M., Pienta, A., Rockwell, R., Timms-Ferrara, L., & Young, C. (2009). From Preserving the Past to Preserving the Future: The Data-PASS Project and the Challenges of Preserving Digital Social Science Data. *Library Trends*, 57(3).

G. King. 2007. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociological Methods and Research 36: 173–199NSB

National Research Council. 2005. Expanding access to research data: Reconciling risks and opportunities. Washington, DC: The National Academies Press.

National Research Council. 2009. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. Washington, DC: The National Academies Press

S. Vadhan , , et al. 2011. "Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections". Available from: http://dataprivacylab.org/projects/irb/Vadhan.pdf