Response to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research

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

January 12, 2012

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 records wInstitute 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 had 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]

The preservation of quantitative data has a more extensive history and more well-established practices than in most other disciplines. Social science continues to rely heavily on data in its traditional forms, such as opinion polls, voting records, surveys, and government statistics and indices. On the other hand, although most large data sets are in public archives, most data produced by and used in social science research is neither publicly available nor preserved by an archival organization. And digital content is evolving into more forms than can be preserved readily. Changes in technology and society are greatly affecting the types and quantities of potential data available for social-scientific analysis. Any data describing human activity may be a subject of social science research. Taken as a whole, the evidence base of social science is shifting [King 2011], and consequently, approaches to curating this evidence, or data, is shifting as well

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.

Supporting Long-Term Access to the Scientific Evidence Base

The values of the Alliance are highly relevant to establishing approaches for ensuring long-term stewardship and encouraging broad public access to digital data that result from federally-funded scientific research. When applied, these values support the practical collaboration of private, public, and governmental memory institutions to support long term access to research data.

Institutional Collaboration

The Data-PASS partnership is based on institutional collaboration, in which multiple organizations and virtual organizations adopt joint stewardship of collections. Partnership, is of course, an ancient approach, but the revolution in communication technology has lowered the barrier to widely distributed partnerships, and the ease of replication of digital content is enabling partnerships to take on a far more direct role in the stewardship of content.

Many threats to long-term access can be effectively ameliorated only when collections are replicated, geographically distributed, and audited by independent institutions. Independent replication and auditing reduce the risks of loss from software, hardware, and physical failure. Moreover replication across a diverse set of institutions that use a variety of business models and operate under different legal regimes insures against many forms of institutional risks, such as curatorial error, institutional mission change, and loss of funding. [Rosenthal, et al. 2005]

Building Mutually Reinforcing Infrastructure and Archival Practice

Institutional collaboration for long-term access requires institutions to establish mutual trust. And it is critical that there be a sound basis for this trust. Data-PASS is committed to good archival practice based on criteria for trustworthy organizations and partnerships that provide solid evidence. One widely recognized example of good archival practice is TRAC, the Trustworthy Repositories Audit & Certification Criteria and Checklist [CRL 2007], which is now in process of being reformulated as an ISO-standard.

In support of good archival practice, Data-PASS has developed open source infrastructure, SafeArchive [Altman & Crabtree 2011], that automates archival replication and auditing polices. The SafeArchive system provides a way to ensure that replicated collections are both institutionally and geographically distributed and to allow for the development of increasingly measurable and auditable trusted repository requirements. Designed as a virtual overlay network on LOCKSS [Rosenthal, et al. 2005], the system provides the auditability and reliability of a top-down replication system with the resilience of a peer-to-peer model. This enables any library, museum, or archive to audit that its content is being replicated across an existing LOCKSS network in conformance with documented archival policies; and to allow groups of collaborating institutions to automatically and verifiably replicate each others' content consistent with a set of expressed commitments. The result is that archives can more easily collaborate to preserve content through geographically and institutionally replication; which mitigates against technical and organizational threats to preservation.

SafeArchive and LOCKSS are exemplars of open community infrastructure that is designed explicitly to support long-term preservation and access. Another exemplar, developed by Partnership members, is the DataVerse Network. An exemplar of collaborative community efforts is the Dataverse Network project [King 2007] recently described by the National Research Council of the National Academies as the "State of the Practice in Data Sharing." [National Research Council 2011] The DVN is a data preservation and dissemination system that is based on open standards, supports open protocols, and integrates with systems such as LOCKSS to enable institutionally distributed replication and stewardship.

Interoperability through Open Standards

Long-term access to data, and effective collaboration to ensure it require the development of open protocols and information standards. For example, the Data-PASS shared catalog is based on open standards for metadata and metadata exchange. In the social sciences, many data producers and data archives have converged on the Data Documentation Initiative (DDI) metadata standard. [see http://www.ddialliance.org] The DDI specification provides a mechanism to document data in a structured, machine-actionable way. Combining information standards such as DDI with with open protocols such as the Open Archives Initiatives metadata harvesting protocol [Lagoze, et 2002], enables institutions to more effectively collaborate to manage and provide access to data.

Standards are equally important, and most often absent, for the use of data in scientific publications.. Data-PASS actively promotes citation standards for research data. Accurate citation of data will promote more and better science. It will make data easier to find, to replicate, and to manage for the long term. Moreover it will make it much easier to trace the influence of data on social science. We advocating a simple baseline standard: title, author, data, and a persistent identifier (of any widely recognized type, such as URN's, handles, DOI's). [See, for an example, Altman & King 2007]

Summary of Major Recommendations

- Support institutional collaboration to provide short and long-term access and stewardship of data, and to develop related standards and infrastructure

- Establish policy that requires auditable replication of data across multiple institutions that have demonstrated capacity and commitment to long-term access.

- Leverage substantial national and international efforts to develop archiving, metadata, and citations standards.

Additional Responses on Selected Questions

The principles and recommendations above apply broadly to the set of questions posed by the RFI.

In addition we fully endorse the recommendations of the National Digital Stewardship Alliance of which Data-PASS is a founding member. These recommendations may be found here (and are also attached with this submission):

http://digitalpreservation.gov/documents/NDSA_ResponseToOSTP.pdf

Finally, institutional members of Data-PASS have submitted responses on behalf of their institutions. (Copies of these may be found here: http://www.data-pass.org/node/95) We support the principles of embodied in these responses and recommend that they be carefully considered.

References

Altman, M., Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. (2009). Digital Preservation Through Archival Collaboration: The Data Preservation Alliance for the Social Sciences. *American Archivist*, 72(1).

Altman, Micah and Jonathan Crabtree. 2011 "Using the SafeArchive System: TRAC-Based Auditing of LOCKSS," Archiving 2011 Final Program and Proceedings, May 16–19, 2011, Salt Lake City, Utah: 165–170. Society for Imaging Science and Technology: http://bit.ly/tLzUmr

Gutmann, M., 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).

Hedstrom, Margaret, Jinfang Niu, Kaye Marz, (2008). "Incentives for Data Producers to Create "Archive/Ready" Data: Implications for Archives and Records Management", *Proceedings of the Society of American Archivists Research Forum*.

King, G., (2007). An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research*, 36(2), 173-199

King, Gary. "The Changing Evidence Base of Social Science Research." In The Future of Political Science: 100 Perspectives, edited by Gary King, Kay Schlozman and Norman Nie. New York: Routledge Press, 2009.

Carl Lagoze, Herbert Van de Sompel,M. Nelson, M., & S. Warner, "The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0.", (2002). http://www.openarchives.org/OAI/openarchivesprotocol.html

National Research Council. 2011. *Communicating Science and Engineering Data in the Information Age: Panel on Communicating National Science Foundation Science and Engineering Information to Data Users*. Preprint. Washington, D.C.: National Academies Press: http://bit.ly/NCSES

David S. Rosenthal, Thomas Robertson, Tom Lipkis, Vicky Reich, Seth Morabito. "Requirements for Digital Preservation: A Bottom-Up Approach", D-Lib Magazine 11 no. 11 (2005)