Comments on Public Access to Federally-Supported Research and Development Data

The Data Preservation Alliance for the Social Sciences (Data-PASS) welcomes the February 22, 2013, White House Memorandum on "Increasing Access to the Results of Federally Funded Scientific Research." Data-PASS (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. Collectively, the founding partners have over 200 years of combined experience in social science data sharing

Data sharing needs to be built into the research and publication workflow — and not treated as a supplemental activity to be performed after the research project has been largely completed. Furthermore, ensuring long-term access to data requires a multi-institutional approach. As many threats to long-term access can be effectively ameliorated only when collections are replicated, geographically distributed, and audited by independent institutions. [Rosenthal 2005]

Data-PASS, as stewards of established non-profit data repositories, and as stakeholders, respectfully offers the following recommendations:

1. In 1985, the NRC [Fienberg, et. al 1985] issued recommendations for access to research data. The core recommendations of this report should guide the development of policies requiring data management plans and the creation of individual data management plans. In particular: (a) Sharing data should be a regular practice. (b) Investigators should share their data by the time of publication of initial major results of analyses of the data except in compelling circumstances. (c) Data relevant to public policy should be shared as quickly and widely as possible. (d) Plans for data sharing should be an integral part of a research plan whenever data sharing is feasible.

2. Any data that is essential for the full understanding of a published work should be recognized as an essential part of the scholarly record [Altman 2013]. Such data should (a) be included for public distribution in a data management plan; and (b) cited in any publications which rely on that data.

3. Robust infrastructure is now available for data citation. [Brase 2012] Data citation should include at least the following elements: author (or authoring entity), title (possibly a generic title), a date (or formal database version, if available), a persistent identifier (such as a DOI), and some form of fixity information (that can be used to validate data retrieved later). [Altman & King 2007]

4. At each stage of a research lifecycle, from project design through data collection, analysis and publication, knowledge about the research and data is created. When information about instruments, methods, context, and meaning from across the stages of research are shared, data are more trustworthy and linkages among disparate data can be formed. Standards for capturing metadata ("data about data") should be supported and encouraged.

5. Inconsistent and simplistic treatment of information confidentiality and security are a barrier to efficient access to and reuse of research data. A series of reports by the National Research Council [2005, 2009], have reinforced that data produced or funded by government agencies 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, mediated access to confidential data through interactive systems, and open access to data altered to maintain confidentiality.

6. Like treatment of other risks to subjects, treatment of data privacy risks should be based on scientifically informed analysis that includes the likelihood of p risks being realized, the extent and type of the harms that would result from realization of those risks, the availability and efficacy of technical, computational/statistical, and legal methods to mitigate risks. [Vadhan, et al. 2010]

7. There exists diversity in approaches for data management within various scientific communities, which is healthy. In cases where communities have resources for data management, it is worthwhile to build upon existing infrastructure. However reliability cannot be assumed. In support of the stipulation (OSTP memo, section 2.c) that agencies develop "strateg[ies] for measuring and, as necessary, enforcing compliance with its plan." We recommend that (a) providers of infrastructure for access to research data regularly demonstrate rather than simply assert capability for long term stewardship [NDSA, 2011]; (b) the effectiveness of agency data availability policies be regularly assessed; (c) individual data management plans be systematically evaluated for compliance.

8. We strongly support section 4.c of the OSTP memo which allows the inclusion of appropriate costs for data management, preservation and access in proposals for Federal funding for scientific research. The costs for these activities and their infrastructure over time are non-trivial.

We believe that the development of consistent federal agency policies to ensure access to data will accelerate scientific discovery, improve education, and empower entrepreneurs to translate research into commercial ventures and jobs.

Our organizations, unlike commercial entities, have a primary and enduring mission to generate new knowledge, to preserve it, and to share it. We are uniquely positioned to support the goals of the Memorandum. We commend the OSTP on the Memorandum, and we stand ready to provide additional input at any stage in the evolution of the implementation plans.

Respectfully submitted on behalf of the Data-PASS by it's steering committee. George Alter, ICPSR; Micah Altman, MIT; Mark Abrahamson, Roper Center; Merce Crosas, Harvard U. Jon Crabtree, Odum Institute; Gary King, Harvard; William LeFurgy, Library of Congress; Amy Pienta, ICPSR; Libbie Stephenson, UCLA

References:

Altman, M., & King, G. (2007). "A Proposed Standard for the Scholarly Citation of Quantitative Data". *DLib Magazine*, 13(3/4),

Altman, M. 2012. "Data Citation in the Dataverse Network", in *For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Worksho*p, National Academies Press.

Brase, Jan, 2012, The DataCite Consortirum in *Developing Data Attribution and Citation Practices and Standards: Summary of an International Worksho*p, National Academies Press.

Fienberg, et al. (eds). 1985. *Sharing Research data.* Washington, DC: The National Academies Press.

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.

NDSA 2011. "Response to Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research". Available from: http://digitalpreservation.gov/documents/NDSA_ResponseToOSTP.pdf

Vadhan, S. , et al. 2010. “Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections”. Available from: http://dataprivacylab.org/projects/irb/Vadhan.pdf

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)