Response to:
Request for Information: Input into the Deliberations of the Advisory Committee to the NIH Director Working Group on Data and Informatics (NOT-OD-12-032)

Dr. Micah Altman
Director of Research -- MIT Libraries, Massachusetts Institute of Technologies
Head/Scientist, Program for Information Science
Non-Resident Senior Fellow, The Brookings Institution

Libbie Stephenson
Distinguished Librarian
Director, UCLA Social Science Data Archive

Writing on behalf of the Data Preservation Alliance for Social Sciences (http://data-pass.org)

INTRODUCTION

Thank you for the opportunity to submit comments for input into the deliberation of the committee.

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 partners collaborate to acquire data at risk of being lost to the research community; to develop preservation and data sharing practices; and to create open infrastructure for collaborative data indexing, sharing, and preservation.

Collectively, the founding partners have over 200 years of combined experience in social science data sharing. 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 Section in the National Archives and Records Administration's Research Services – Archival Operations, Washington, DC (RD-DC), 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.

EMERGING STANDARDS AND PRACTICES FOR SHARING AND MANAGING DATA

The task force may wish to take note of broad-based and thoughtful commentary on data sharing that has emerged from the research community, including the following:

- The National Science Board's draft report on Digital Research Data Sharing and Management [NSB 2011], which emphasizes the value of open access to data, and identifies key challenges to promoting wide access.
- The NRC's recently report on C*ommunicating Science and Engineering in the Information Age*, which develops a number of recommendations, that although directed at NCSES are readily applicable to research data management, publication, and dissemination in general. Specifically, recommendations 3-1, 3-2, 3-3, and 3-4 together represent general good practice for data management and publication: Published results are more reliable when the underlying data is available; and when management of the that data incorporates versioning, open formats and protocols, machine-actionable metadata, and systematic tracking of information provenance and modification from initial data collection through subequent publications. [NRC 2011]
- Numerous responses to the recent ANPRM on proposed changes to the common rule noted how a overreaching and unsophisticated approaches to confidentiality can drastically erode the value of data sharing. Notably, two responses by data privacy and computer science researchers provide a roadmap for simultaneously increasing data sharing and privacy protections by leveraging advances in theoretical computer science, and by establishing mechanisms for accountability and transparency. [Sweeney, et al., 2010; Vadhan, et al. 2010]
- Numerous responses to the recent *OSTP Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research* [OSTP 2011], which comment on the benefits of data access, and draw attention to needs, protocols, and standards for open data access and interoperability. Notably, the responses of the National Digital Stewardship Alliance, the Data-Preservation Alliance for the Social Sciences, Carnegie Mellon University, the University of California Libraries, and the International Consortium for Political and Social Research, reference the need for and successful exemplars of community-based standards for open data dissemination, discovery, and preservation.

DIRECTED RESPONSES

RESEARCH INFORMATION LIFECYCLE

In his work on research data lifecycles, Charles Humphrey [2004] has provided an overview of this process applicable to a variety of research disciplines. “Life cycle models are shaping the way we study digital information processes. These models represent the life course of a larger system, such as the research process, through a series of sequentially related stages or phases in which information is produced or manipulated.“ Further, “well organised, well documented, preserved and shared data are invaluable to advance scientific inquiry and to increase opportunities for learning and innovation.”

At each stage of a research lifecycle, from when the project is first designed, through to data collection, analysis and publication, knowledge about the research and data is created. When the data can be shared for re-use and re-purposing, the relationships among the stages enables linkages among disparate data points to come to new understandings, new conclusions, and new ways of visualizing data relationships. When considering the management, integration, and analysis of large biomedical datasets, the roles and responsibilities of researchers and data management organizations need to be determined. Ideally, this should focus on documenting the stages in the research lifecycle, including:

- Design of a research project
- Data collection processes and instruments
- Data organization in digital format
- Documentation of data analysis process
- Publication or sharing of results
- Dissemination, sharing, and reuse
- Preservation, long term conservation, and long term access

Bernstein, et *al*. [2011] suggest that any approach to documenting research data life cycle must take into account gaps in the transfer of information about the data. An overlapping approach that looks toward the short and long term future should incorporate methods for catching information gaps; "There is no simple way to reach the goal of maintaining high volumes of important data in heterogeneous environments over many decades. Not only must a cross-generational communication system reach from the present to the future half a century from now, but from many points in time in the near future to many points in time in the distant future. No single current medium, no single current file format specification is likely to be sufficiently robust and adaptable to survive without major changes over half a century."

The best approach is to consider overlapping or collaborative data management solutions. Examples of this approach are reflected in the work of the MaDAM project recently concluded in the UK [Poschen 2010] . This project aimed to develop tools and an infrastructure for life-cycle management of biomedical data. Shahand [2011] and colleagues have addressed research life cycles in bioinformatics and have carefully examined roles and responsibilities "to support a wide spectrum of user profiles, with different expertise and requirements." Their work demonstrates that a "service-oriented architecture" will best support each of the phases of the research lifecycle and will enable the use and reuse of data by users with varying backgrounds.

CHALLENGES/ISSUES FACED BY THE EXTRAMURAL COMMUNITY -- A SOCIAL SCIENCE PERSPECTIVE

Social scientists are increasingly using biomedical data in research. Hauser, [2010] et al. have described "a growing tendency for social scientists to collect biological specimens such as blood, urine, and saliva as part of large-scale household surveys." By combining social and behavioral measures with biological measures, researchers are able to answer questions and make new connections in their field of inquiry. For example, "it becomes possible, for example, to estimate the distribution of a particular genetic variant within a representative sample of the general population and to correlate genetic variations with differences in human phenotypes." [Hauser, 2010]

A number of government surveys, such as the National health and Nutrition Examination Survey, collect biological samples and the measures are recorded in resulting public-use statistical files. However, conducting surveys in which biological specimens are gathered has new challenges for the individual social scientist which are financial, legal or ethical in nature, but also have to do with archiving and sharing data. Some areas have had a fair amount of attention, built on the experiences gained by non-governmental larger scale research projects, such as the Study of Women's Health Across the Nation (SWAN). In terms of policies and suggested procedures, issues on gaining access to, collecting, storing and using biomedical specimens have been addressed somewhat. However, the challenges in protecting privacy and confidentiality, informed consent, and data sharing are complex and will benefit from

collaborative efforts to find solutions to problems faced by all researchers regardless of methodological approach or discipline.

Some best practices have emerged through experience. There are numerous documents describing how specimens should be handled in laboratories, and university offices for protection of human subjects have detailed required protocols. Professional societies have also developed recommendations regarding how to organize and manage bio-data repositories, such as the ISBER 2008 document Best Practices for Repositories: Collection, Storage, Retrieval and Distribution of Biological Materials for Research. However, as Hauser [2010] states while the best practices so far developed are helpful, "they do not address questions that are most closely related to the design of the research itself, such as choice of biospecimen (e.g., blood, urine, saliva), choice of biomarker, and choice of assay". Further, "the data archive (i.e., the collection of data derived from the specimens, as well as the data from the survey) is likely to be maintained separately from the specimens themselves, while documentation about the specimen collection and survey protocols may be archived in yet another location." Because there are these multiple storage and access sites, social scientists face challenges in ensuring that all parts of a research project are equally preserved, secure and managed for the long term.

STANDARDS/PRACTICES FOR ACKNOWLEDGMENT OF THE USE OF DATA

Any information that is essential for full understanding of a published work should be recognized as an essential part of the scholarly record. Where such information is not directly incorporated as an integral part of the publication itself, this integral material should be cited as evidence. And all information that is cited, should be accessible to the scientific community.

As Altman & King [2007] point out, omitting data citation, or using ad-hoc footnotes, local id numbers, or other schemes, threatens the integrity of the scientific record:

> The data cited [in an ad-hoc way, lacking minimal standards] may no longer exist, may not be available publicly, or may have never been held by anyone but the investigator. Data listed as available from the author are unlikely to be available for long and will not be available after the author retires or dies. Sometimes URLs are given, but they often do not persist. In recent years, a major archive renumbered all its acquisitions, rendering all citations to data it held invalid; identical data was distributed in different archives with different identifiers; data sets have been expanded or corrected and the old data, on which prior literature is based, was destroyed or renumbered and so is inaccessible; and modified versions of data are routinely distributed under the same name, without any standard for versioning. Copyeditors have no fixed rules, and often no rules whatsoever. Data are sometimes listed in the bibliography, sometimes in the text, sometimes not at all, and rarely with enough information to guarantee future access to the identical data set. Replicating published tables and figures even without having to rerun the original experiment, is often difficult or impossible.

Science provides a succinct statement of this evidential principle in its General Information for Authors: "Citations to unpublished data and personal communications cannot be used to support claims in a published paper.", and "All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science." [Science 2011] Citation standards for data have been recently adopted by the American Sociological Association, the OECD, and over 20 institutional members of the DataCite coalition, and are emerging as a best practice in publishing [Alter 2012; NDSA 2012]

As Alter [2012] notes:

> Scientists who create digital data have a right to expect their contributions to be recognized through citations in publications based on those data. Citation has been the standard way of recognizing original scholarship for hundreds of years. As we noted above, academic careers are measured by citations, and proper citation of data would credit data producers for the impact of their work on science. Citations can also be linked to funding sources (e.g., grant numbers) in ways that can be captured to measure the impact of Federal investments on scientific productivity.

And also:

> Assigning proper citations and persistent identifiers to data resources is critical to enabling reuse and verification of data, understanding and tracking the impact of research data, and creating a structure that recognizes and rewards data producers for their contributions to the scientific record. Many data archives and repositories now provide citations that should be used in publications based on the data, and many are also registering persistent identifiers for the data they manage. Data citations permit data to be integrated into the system of scholarly communications and to be picked up by the electronic citation services so that data usage can be tracked.

Citation themselves need not be complex. Altman-King [2007] provide a set of minimal citation elements, most of which have been incorporated in subsequent data citation approaches, and which ensure the reliability of the citation. A citation should include 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).

And publication of the citations is also straightforward. Citations to data should be treated as first-class references, and treated in the same manner as citation to other publications. Authors should acknowledge the use of data by including citations to the data in the references section of their papers (where citation to other publications are also recorded). Treating data citations as first-class references provides attribution and recognition of the importance of data as an intellectual product. Journals and other publishers should require data citation as a perquisite for publication, just as they require proper citation of other publications referenced as evidence. And cited data should be held to similar standards of durability and accessibility as other cited works integral to the understanding and reproduction of the published results.

Indeed, since science is not merely about behaving scientifically, but also requires a community of scholars competing and cooperating to pursue common goals, scholarly citation of data can be viewed as an instantiation of a central feature of the whole enterprise.

WAYS TO IMPROVE THE EFFICIENCY OF DATA ACCESS REQUESTS

To improve the efficiency of responses to data access requests, 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. Over two decades ago, the path-breaking report *Sharing Research Data,* by the Committee on National Statistics [1985], identified this and related requirements

in its core recommendations:

> Recommendation 1: Sharing data should be a regular practice.
> Recommendation 2: Investigators should share their data by the time of publication of initial major results of analyses of the data except in compelling circumstances.
>
> Recommendation 3: Data relevant to public policy should be shared as quickly and widely as possible.
>
> Recommendation 4: Plans for data sharing should be an integral part of a research plan whenever data sharing is feasible.

Increasingly, these recommendations have been recognized by data management requirements and policies. However, all too often these recommendation are not followed in practice. Strengthening the implementation of recommendation two, which requires simultaneous publication of data along with derived research results, so that policies include reporting and auditing, would likely yield substantial improvements in practice. The data citation standards and practices described above could readily be utilized to improve systematic reporting and tracking in this area.

In addition, inconsistent and unsophisticated treatment of information confidentiality and security have become a major stumbling block to efficient access to and use of research data. A series of reports by the National Research Council [2005, 2007, 2009, 2010] , have reinforced the following key points:

- One size does not fit all -- multiple modes of access are needed to confidential data:"Recommendation 2: Data produced or funded by government agencies should continue to be made available for research through a variety of modes, including various modes of restricted access to confidential data and unrestricted access to public use data altered in a variety of ways to maintain confidentiality." [NRC 2005]
- The complexity, detail, richness, and temporal extents of new forms of scientific data -- such as geospatial traces (linked social-spatial data), long term longitudinal studies, social networks, and rich genomic data -- create significant uncertainties about the ability for traditional 'anonymization' methods and standards to protect the confidentiality of research participants.
- The approach taken by the HIPAA Privacy Rule, which emphasizes 'deidentification' through suppression of data values is a poor fit for much data, even within health research. The HIPAA approach is neither necessary nor sufficient to protect confidentiality. Moreover, the data suppression techniques used by the approach can severely impair the utility of the data, and lead to biased research results.

Furthermore, numerous responses to the responses to the recent ANPRM on proposed changes to the common rule, including extensively researched responses from Harvard [Barnes 2011], from twenty two leading research organizations [COSSA 2011] and from leading computer scientists and privacy scientists [Sweeney 2011; Vadhan 2011] noted how a overreaching and unsophisticated approaches to confidentiality can drastically erode the value of data sharing and reuse.

These responses to the recent ANPRM emphasized the points made in the NRC reports above:

- Multiple modes of access to confidential information should be provided including access to the original data through data enclaves/under restricted use agreements
- The HIPAA which emphasizes approach, which deidentification through removal of enumerated categories of information, is a poor fit for social science and behavioral research.

These responses emphasize that treatment of privacy risks require a nuanced approach. Like treatment of other risks to subjects, treatment of privacy risks should be based on a scientifically informed analysis that includes the likelihood of such 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 methods to mitigate risks, and the availability of legal remedies. The responses by data privacy and computer science researchers [Sweeney, et al., 2010; Vadhan, et al. 2010] provide a roadmap for simultaneously increasing data sharing and privacy protections by leveraging advances in theoretical computer science; creating legal mechanisms for accountability and transparency; and establishing a task force of privacy experts that would develop and update safe-harbor rules that would apply to emerging forms of data and disclosure limitation methods.

REFERENCES

Alter, G. 2012. Response to RFI: "Public Access to Digital Data Resulting From Federally Funded Scientific Research" Office of Science and Technology Policy" (response on behalf of the Inter-University Consortium for political and Social Research." Available from: http://www.data-pass.org/sites/default/files/ICPSR%20Response%20to%20RFI%20Public%20Access%20to%20data.pdf

Altman, M., & King, G. 2007. "A Proposed Standard for the Scholarly Citation of Quantitative Data". *D-Lib Magazine*, 13(3/4), Available from: http://www.dlib.org/dlib/march07/altman/03altman.html

Barnes 2011. Re: Human Subject Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay and Ambiguity for Investigators, Federal Register Vol 76, No. 143, July 26, 2011 (response on behalf of Harvard University) . Available from: http://dataprivacylab.org/projects/irb/HarvardUniversity.pdf

Bernstein, H. J., Folk, M. J., Benger, W., Dougherty, M. T., Eliceiri, K. W. and Schnetter, E. (2011). **Communicating Scientific Data from the Present to the Future**. Dowling College position paper. Temporary URL: http://www.columbia.edu/~rb2568/rdlm/Bernstein_Dowling_RDLM2011.pdf

Cossa 2011. Social and Behavioral Science White Paper onAdvanced Notice for Proposed Rulemaking (ANPRM) Federal Register 44512-531 (July 26, 2011); ID Docket HHS-OPHS-2011-0005. Available from: http://dataprivacylab.org/projects/irb/COSSA.pdf

Humphrey, C. & Hamilton, E. 2004. "Is it Working? Assessing the Value of the Canadian Data Liberation Initiative." **Bottom Line**, Vol. 17 (4), pp. 137-146.Available from: http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

Data-PASS 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://www.data-pass.org/sites/default/files/datapass-otsp-rfi-response.pdf

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

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

National Research Council. 2007. *Putting people on the map: Protecting confidentiality with linked social-spatial data.* 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.

National Research Council. 2010. *Conducting biosocial surveys: Collecting, storing, accessing, and protecting biospecimens and biodata.* Washington, DC: The National Academies Press.

NRC. 2011. *Communicating Science and Engineering Data in the Information Age*. National Academies Press. Available from: http://www.nap.edu/catalog.php?record_id=13282

NSB 2011. *Digital Research Data Sharing and Management*. (Draft) NSB-11-17. Available from: http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf

Science staff. 2011. "General Information for authors." Available from: http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml

OTSP. 2011. "Public Access to Digital Data: Public Comments". Available from: http://www.whitehouse.gov/administration/eop/ostp/library/digitaldata

Poschen, M., et *al*. 2010. User-Driven Development of a Pilot Data Management Infrastructure for Biomedical Researchers." Available from: https://www.escholar.manchester.ac.uk/api/datastream?publicationPid=uk-ac-man-scw:117518&datastreamId=FULL-TEXT.PDF

Shahand, s. et *al.* 2011. "Front-ends to Biomedical Data Analysis on Grids." Available from: http://www.bioinformaticslaboratory.nl/twiki/pub/EBioScience/EBioinfraGateUserDoc/ebioinfragate.pdf

Sweeney, L. , et *al*. 2010. "Comments from Data Privacy Researchers". Available from: http://dataprivacylab.org/projects/irb/DataPrivacyResearchers.pdf

UK Data Archive. 2012. "Research Data Life-cycle." http://www.data-archive.ac.uk/create-manage/life-cycle

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