# Public Access to NSF-Funded Research Data for the Social, Behavioral, and Economic Sciences

## Workshop Report

May 17, 2016

*This workshop was sponsored by the Directorate for the Social, Behavioral and Economic Sciences at the National Science Foundation through a supplement to award #0940818. Any opinions, findings, conclusions, or recommendations presented in this material are only those of the authors and do not necessarily reflect the views of the National Science Foundation.*

# Contents

# Overview

In the fall of 2014, with support from the National Science Foundation's Directorate for Social, Behavioral, and Economic Sciences (SBE), Dr. Steven Ruggles formed a working group to study the revamping of Data Management Plan Guidelines for proposals submitted to SBE. The group's goal was to envision more detailed guidance regarding data management plans so that norms for reviewers and preparers are articulated and data are accessible and interoperable and to investigate additional measures that could help broaden public access to research data. The working group represented several program areas within SBE:

- Steven Ruggles (History and Population Studies, University of Minnesota)
- Karen Adolph (Psychology and Neural Science, New York University)
- Robert Chen (Data Science and Geography, Columbia University)
- Barbara Entwisle (Sociology and Geography, University of North Carolina, Chapel Hill)
- Janet Gornick (Political Science and Sociology, City University of New York)
- Myron Gutmann (History and Demography, University of Colorado, Boulder)

Together with project coordinators Gina Rumore and Catherine Fitch, this group organized a workshop on Public Access to NSF-Funded Research Data for the Social, Behavioral, and Economic Sciences held at the University of North Carolina at Chapel Hill on January 27-28, 2016.

In consultation with SBE Program Officers, the Working Group selected workshop participants to represent the breadth and diversity of scientific disciplines funded by SBE—from cultural anthropology and history and philosophy of science to neuroscience and economics—and also the wide variety of data types used in SBE research—from field notes and oral histories to functional MRIs and corporate data. In addition to subject matter experts, the Working Group also invited participants who have expertise in data management in the social, behavioral, and economic sciences. (A complete list of participants, their disciplines, and their home institutions can be found in Appendix B; a list of types of data used in SBE research, in Appendix E.) Although the workshop participants identified many challenges with making data publicly available and raised many important and practical concerns, the consensus of workshop participants was overwhelmingly that all data resulting from NSF-funded research should be stored in reputable repositories and made publicly available to the greatest extent possible. Further, many participants challenged the Working Group to be "bold" in their recommendations to NSF, suggesting that not only should data sharing be required but that there should be strong incentives to do so.

In the following sections we present the background to the current call for increased public access to federally-funded research data, the questions the Working Group posed to workshop participants, a discussion of their answers, and, finally, our recommendations to NSF. Our Working Group endorses, without reservation, that all data produced by NSF-funded research should be made publicly available, with appropriate protections when needed, and we have specific recommendations for principal investigators, review panels, program staff, SBE and NSF. In Appendix A of this report we provide an edited version of the "Data Management for NSF SBE Directorate Proposals and Awards"[1] to show the changes that we recommend.

[1] http://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf

# Background

For the past three decades, leaders of the scientific community have called for sharing access to scientific data (e.g., Fienberg, Martin, and Straf 1985; National Research Council 1995, 1997, 2009; Esanu and Uhlir 2004).

Borgman (2012) summarized the main rationales for data sharing:

- To reproduce or to verify research
- To make the results of publicly funded research available to the public
- To enable others to ask new questions of extant data
- To advance the state of research and innovation

Increasingly, funding agencies around the world are requiring investigators to share data created with public support (OECD 2007; Belmont Forum 2015; Castro and Korte 2015).

The NSF has encouraged data sharing for decades. In 1988 the Grant Policy Manual observed that "some NSF grants support the accumulation of a large body of machine-readable data … the data bank may be so large and comprehensive that it would probably be useful to others for other purposes ... NSF encourages and in some cases may require that such materials be distributed or made available" (NSF 1988: VII-13). When the Manual was revised in 1995, the language was strengthened:

> Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing (NSF 1995: 11).

After much discussion and debate, in 2010 NSF began requiring proposals to include a two-page data management plan (DMP) to address such issues as data and metadata format and content; policies for access, sharing, reuse, and redistribution; and plans for archiving and preservation (NSF 2011). This is currently the most comprehensive requirement of any federal agency; unlike NIH and other agencies, the NSF DMP requirement applies to all grants, and the plans are subject to peer review.

In February of 2013, the White House Office of Science and Technical Policy (OSTP) released a memorandum entitled *Increasing Access to the Results of Federally Funded Scientific Research*. In this memorandum the OSTP directed "each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government." The memorandum further asserts that "digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze."

In response to the OSTP memorandum, in March of 2015, NSF released its Public Access Plan, *Today's Data, Tomorrow's Discoveries.* This document outlined a clear policy and method for the public sharing of all publications resulting from research funded by NSF. It left the second issue under consideration, the sharing of data resulting from NSF-funded research, largely up to individual directorates and communities of researchers.

The NSF Public Access Plan has contributed to a sense of urgency about public access to research data. Because the required DMPs are peer-reviewed, the investigator's description of what the data are and how they will be shared plays a direct role in the funding process. Nevertheless, there remain real questions about the scope and effectiveness of the current policy and the parameters of any new policy to be implemented.

## Workshop Structure

To address the current NSF Data Management Plans and data sharing policy and to formulate recommendations for improvements, we held a workshop at the University of North Carolina, Chapel Hill on January 27-28 2016, in which we asked our participants to consider the following questions:

- *What is data and why share it?*
- *What is the proper scope of a data management plan? What kinds of data should be covered?*
- *What are the ethical issues that must be considered in sharing sensitive data?*
- *What are the logistics to enable data sharing? Once you have a plan, who implements it, pays for it, enforces it, etc.?*

In advance of the workshop, we asked participants to write and circulate to the larger group two-page briefs addressing these questions. We then structured the first day of the workshop around these questions, allowing roughly 90 minutes of moderated discussion on each (more details on each question is provided below). At the end of the first day of the workshop, the working group drew on the feedback we received from our participants in writing and in person to formulate a first draft of recommendations for what the SBE directorate of NSF should require in all DMPs and how these plans should be enforced. At the start of the second day, the Working Group presented our draft recommendations, followed by nearly four hours of discussion about the details.

The following sections provide an expanded list of the questions we asked, a synopsis of the discussion around each of these questions, and the conclusions drawn by the Working Group at the end of the two-day workshop. In the final section of the report we offer our recommendations to the Directorate of Social, Behavioral, and Economic Science at NSF as to how Data Management Plan instructions should be amended and how the Plans should be enforced.

## Questions

### Question 1: What is Data and Why Share It?

- What does your research community consider to be "data" when it comes to writing a Data Management Plan? (e.g., Does your community distinguish between source data and processed data used for analyses? Are raw source data useable or interpretable? How can data provenance and workflow be characterized adequately?)

- What is the overall goal of data sharing from the perspective of your community? What are the incentives and disincentives for data sharing from the point of view of individual researchers?
- How can data be more widely used? Should data use be restricted to research and educational/informational purposes? Should data sharing also involve use for commercial purposes (both NSF and NIH have commercial entities)? Are there ways to expand usage of data while maintaining the integrity of the research process?

## Question 2: What is the proper scope of a data management plan? What kinds of data does it apply to?

- How can data management plans better capture the opportunities and challenges of providing access to data to researchers for your areas of research?
- When is a complementary policy on non-digitized data (e.g., biological specimens) applicable/necessary?
- How should data management plans address qualitative (e.g., raw research video, audio files, transcripts) data?
- How should a Data Management Plan address linkages to data not funded by NSF?  Such data may be proprietary.  They may also be governed by other entities and rules.
- Are there new or emerging sources of data or methods in your field that are or likely to be constrained by existing data policies, norms, or practices? What needs to be changed? Do these new sources and methods raise important ethical issues that need to be addressed?

## Question 3: What are the ethical issues that must be considered in sharing sensitive data?

- Who should be allowed to access data? Is it global? Do you need ethics training? Does it depend on the nature of consent given by participants? What kind of institution is the applicant in?
- How can participants' rights be respected? If participant permission is required, what should participant permissions include? Are any data so sensitive that no form of access by other investigators is possible?

## Question 4: What are the logistics to enable data sharing? Once you have a plan, who implements it, pays for it, enforces it, etc.?

- Does your community have repositories (digital libraries) for curating, storing, and serving data? Do you think these repositories should meet community criteria for trustworthiness? Which repositories are trusted? Should repositories be mandated or sanctioned by NSF? If so, what are the criteria for NSF-sanctioned repositories? Who should pay for data curation and storage? How can long-term access be financed, especially for data that are expensive to maintain or manage? Relatedly, how can file formats be kept up to date long-term?

- How can NSF ensure that the promises of data management plans are actually carried out?
- Should NSF allow data to be embargoed, and if so, under what conditions and for how long? How should NSF deal with longitudinal data collection that may require months or years before the original researchers can analyze or publish the data?

# Discussion

## What is data and why share it?

Scientists in the fields funded by the SBE Directorate of NSF rely on extremely diverse types of data in their research. Workshop participants from multiple disciplines use administrative and survey data; developmental and cognitive scientists use brain imagery, video recordings, transcripts, and text-based flat files; and economists and others are increasingly using private, commercial data. Other fields of SBE, including branches of sociology, geography, anthropology, and history and philosophy of science, rely predominantly on qualitative data, including interviews, field notes, and archival materials. Appendix E provides an extensive but not exhaustive list of types of SBE data identified by workshop participants in their pre-workshop briefs and discussed during the workshop.

SBE researchers frequently employ data that involve human subjects—from video recordings of infants learning to crawl to biomarker collection for major social science surveys and experimental data involving human subjects. Further, research funded by SBE often involves linking data, bringing data from two or more sources that may have different funding agencies, access policies, disciplinary norms and practices, and ethical considerations. NSF may fund a study that creates a new linked dataset from existing datasets, and these data could be subject to many restrictions (all of the restrictions of each of the source datasets, plus more, if linking the data creates new risks and concerns).

Matthew Woollard, Director of the UK Data Archive, opened our session on "What is data and why share it?" by putting forth the definition used by the UK Economic and Social Research Council Research Data Policy (ESRC): "Research data are defined for the purpose of this document as information relevant to, or of interest to researchers, either as inputs into or outputs from research. They are research materials resulting from primary data collection or generation, or derived from existing sources intended to be analysed in the course of a research project."[2]

In response to this definition, participants raised many concerns:

- Is the definition of data disciplinarily and personally dependent?
- Who owns and controls access to the data?
- What needs to be shared to make data usable?
- What is data versus what do we expect people to share?
- What does it mean for data to be usable?

[2] http://www.esrc.ac.uk/files/about-us/policies-and-standards/esrc-research-data-policy/

Based on feedback from workshop participants in writing and in discussion, the Working Group proposes the following definition of data for the purposes of the NSF Data Management Plan:

> **Data** are defined for the purpose of this document as information relevant to, or of interest to social, behavioral, and economic researchers, either as inputs into or outputs from research. They are research materials resulting from primary data collection or creation, or derived from existing sources.

This definition is adapted from ESRC Research Data Policy and is a broader definition than the 2010 guidelines, as we include physical specimens and commercial data, for example.

Participants not only discussed what data are, but why they should or should not be shared. Motivations for sharing data ranged from the mostly self-interested—the goodwill of the granting agencies—to the utilitarian—the moral obligation to promote the greater good. More immediate reasons to share included the need for replicability/reproducibility of results in science, transparency, longitudinal studies, and the potential of reuse for new purposes. Disincentives for sharing included, most prominently, the fear that scientists have of "being scooped" and the time and money often involved in making data readily available to the research community, with appropriate protections where needed. On the other hand, Margaret Levenstein, Director of the Michigan Research Data Center, argued that if data citations accrued to the researchers whose data is used by others, in the same way publication citations accrue, this would compel purely self-interested researchers to share their data more openly.

Further discussion among workshop participants revealed a real hope that, if the funding agencies do indeed make data sharing a requirement, community norms may evolve over time to make data sharing the norm among scientists and, in turn, the tenure and promotion process will, too, come to recognize and reward the creation and sharing of data.

Despite the disparate types of data used by SBE scientists in the disciplines represented by our workshop participants and despite the potential disincentives to sharing, all participants agreed that the data generated by NSF-funded research should be made freely available to the research community. While our workshop and report use the term "public access" in their titles, most, if not all, of our workshop participants agreed that the general spirit of data sharing is to make the data available to other researchers, as may be defined by institutional affiliation or, in some cases, Institutional Review Board (IRB) approval.

## What is the proper scope of a data management plan (DMP)? What kinds of data does it apply to?

The discussion of scope quickly became tied to purpose. DMPs do not currently mandate data sharing. If the purpose of our recommendations to NSF is to increase data sharing, participants agreed that the DMP instructions and the Plans themselves should specifically require data sharing plans.

Participants identified several key issues in defining the scope of data to be shared and what kinds of data must be shared: (1) the need for creating and sharing metadata as well as data; (2) the need to decide what level of data should be shared (e.g., raw data in the forms of functional MRI scans would not be very useful for other researchers); (3) how to appropriately protect confidential data or data protected by law (e.g., Health Insurance Portability and Accountability Act, Family Educational Rights and Privacy Act); (4) the need to identify responsible repositories acceptable for data sharing (i.e., data stored on a researcher's desktop computer is not being publicly shared even if the researcher has agreed to share if asked); (5) how to handle proprietary data (e.g., data subject to copyright, license agreements, or private contracts); and (6) how to manage data that are created by linking to or utilizing a dataset not owned/controlled by the NSF-funded researcher(s).

To address the many concerns raised about the appropriate scope of a DMP, the Working Group recommends that every SBE review panel should have a data management expert on it. Further, the Working Group recommends (see below) specific disciplinary or interdisciplinary norms that should be adhered to in all DMPs and that deviations from these norms must be convincingly justified by the PI/Co-PIs in their DMP.

## What are the ethical issues that must be considered in sharing sensitive data?

The conversation on ethical issues steered predominantly toward the important role played by reputable repositories in creating open access to sensitive data. For example, a repository like The Inter-university Consortium for Political and Social Research (ICPSR) might create two or more versions of a dataset to facilitate data sharing: one might be a public, open access dataset, and another, with more personal identifiers, might be a restricted-use dataset that would require confirmation that all those with access to the data meet certain requirements and agree to certain protocols. In cases where data need even more protection, a secure data facility could be considered.

Workshop participants also greatly supported creating consent forms that build more open data sharing directly into the consent process. Based on their experiences working with human subjects, workshop participants did not feel that broader consent would dissuade participation in research. Kathleen Mullan Harris, the Director and Principal Investigator of the National Longitudinal Study of Adolescent to Adult Health (Add Health), pointed to the Add Health model that has two levels of consent: one for current research and one for archiving data for future uses. Data that are considered to be more sensitive and private at the time of sharing are likely to become less sensitive and private some years later (although in some cases, data could become more sensitive—e.g., if a person later becomes famous). Accordingly, as time passes it is appropriate to re-evaluate risks to participants.

Among participants working directly with human subjects there was a general frustration with lack of clarity and consistency in the institutional review board (IRB) review process that these researchers feel inhibits their ability to share data. For example, Brian MacWhinney, Professor of Psychology at Carnegie Mellon University, discussed the difficulty of having an IRB decide that voice recordings constitute identifiable data, like a finger print, when this simply isn't the case. There is no national database of voiceprints against which a voice can be compared and the technology for doing so is not up to the task. The participants who work with audio and video data expressed a great deal of frustration with the IRB approval process and felt that it unnecessarily hinders attempts at data sharing. General consensus among the larger group was a hope that the reformulation of the Common Rule would lead to more consistent understanding among IRB officers and review committees.

Although workshop participants strongly felt that there are very few exceptions of data that are too sensitive to share, they did raise some important concerns that need to be addressed by researchers in their DMPs. Rachel Croson, an economist, voiced a concern about personal data being used for unintended purposes, such as for-profit use by commercial companies. Anthropologist Lisa Cliggett added that some anthropological data could pose a potential threat to study participants, giving the examples of land claim disputes in rural communities or study participants who are engaged in illegal activities. Finally, several workshop participants discussed the need for proper training in how to handle data involving human subjects. This could logically be addressed in Responsible Conduct of Research (RCR) training.

In our recommendations to NSF, the Working Group advises that these concerns could all be addressed within the scope of the DMP. First and foremost, DMP instructions should include depositing the data in a reputable data archive that can provide limits to access as required. Further, the DMP should clearly identify and justify any data that cannot be shared as well as any data that will require restricted access. Information needed for interoperability, such as personal identifiers or detailed geography, should be included unless there is a well-justified need to suppress it.

The Working Group further recommends including a data management expert on every review panel who should be able to competently evaluate if data are, indeed, too sensitive to be shared openly and do require special restrictions to access. The Working Group strongly recommends to NSF that the vast majority of sensitive data are not too sensitive to share, given appropriate protections, and should not be excluded from public access requirements.

## What are the logistics to enable data sharing? Once you have a plan, who implements it, pays for it, enforces it, etc.?

Workshop participants almost unanimously agreed that all data produced by NSF-funded research should be deposited in and shared through a responsible data repository. Suggestions included having NSF and/or NIH review and approve repositories or require repositories to meet existing standards, e.g., DataPASS, Data Seal of Approval, ICSU World Data System, Nestor (Network of Expertise in Long-term Storage of Digital Repositories), or ISO 16363. Responsible repositories usually assign a digital object identifier (DOI) to each dataset.

Some participants expressed concern that domain repositories could limit linking data and interdisciplinary research, but the several data archivists among the participants assured the group that domain repositories typically facilitate cross-disciplinary work, noting that successfully linking data across disciplines generally depends on the types of detailed metadata that repositories provide.

Based on feedback from workshop participants in writing and in discussion, the Working Group proposes the following definition of a responsible digital repository for the purposes of the NSF Data Management Plan:

> **Responsible digital repository** is a digital data repository or archive that takes responsibility for data assets according to the “FAIR” data principles—data that are findable, accessible, interoperable, and reusable (except for physical specimens that may not be reusable).

One concern shared among many workshop participants is how researchers should be expected to pay for data preservation. Given that existing repositories in the social, behavioral, and economic sciences are largely grant funded, the group felt that a more secure funding model is necessary. One suggestion was that NSF should require applicants to build expenses for data archiving into their budgets. It was pointed out that NASA has a very successful model of providing separate funds for research and data archiving.

To add teeth to the data preservation and data sharing policy, workshop participants felt that DMPs should be an important part of the proposal review process and that there needed to be repercussions for not meeting the promises made in a DMP. Thus, the Working Group has two specific recommendations to NSF:

(1) Data Management Plans should be reviewed as part of the Broader Impact criteria. Each panel should include at least one reviewer with data management expertise, and DMPs should be rated as “adequate” or “inadequate” (the working group further felt that “awesome” or “lame” rating categories might better convey the message to both PIs and reviewers that data sharing is to become a norm in the research community, but we defer to the judgment of NSF officials on this matter). Plans judged to be “inadequate” (or “lame”) should not be eligible for the “Highly Competitive” category until sufficiently revised.

(2) Principal Investigators should not be eligible for further NSF funding until they have met the promises of previous DMPs or have justified deviation from them. This should be addressed by PIs in their annual reports, final reports, and in the Results of Prior NSF research section of new or competing continuation proposals.

Finally, most workshop participants felt that there needs to be some sort of allowed embargo period before data are required to be deposited and made publicly available. Henry Brady, Dean and Professor of Political Science and Public Policy at UC Berkeley, argued that as the tenure process does not currently give faculty credit for creating data, we must therefore allow PIs the first shot at getting publications from their research data. Participants suggested several alternative embargo periods that could be stipulated by NSF, from until the first publication from the project up to three years from the end of the funding period. Our final recommendation to SBE and NSF is to leave it up to the PIs/co-PIs to define an appropriate embargo period, recognizing that the proposed embargo period will be subject to peer review and, if not reasonable, could negatively impact the evaluation of the grant. Further, even if an embargo period is stipulated in the Data Management Plan, the data should be deposited as soon as possible in a responsible repository, and the embargo period requested by the researcher will be administered by the repository.

## Recommendations

Based on the briefs written by our participants and our discussion during the workshop, the Working Group makes the following recommendations to the SBE Directorate of NSF. To make the implementation of these recommendations as straightforward as possible, we present them in four categories: Instructions to PIs, Instructions to Review Panels, Instructions to Program Staff, and Recommendations to SBE and NSF.

### Instructions to Principal Investigators:

(1) Data sharing is a key aspect of the evaluation of Broader Impacts for all SBE proposals. This includes the Broader Impacts section under Results of Prior NSF Support.

(2) The following norms are expectations; Data Management Plans that deviate from these norms must provide justification:

- Data created with NSF support should be shared with the research community.
- Data should be thoroughly documented with full information about data collection procedures, sources, access procedures, and usage guidelines.
- Data should be deposited in a responsible digital repository that can ensure discovery, long-run preservation, and appropriate access. Data should be deposited even if limits are placed on their use. Data should be deposited as soon as they are available to the PI, even if there is an embargo period.
- Data should be made available as soon as possible. If a PI proposes a limited period of exclusive use of the data collected, it must be explained and justified.

(3) Data Management Plans should be specific about the following:

- Which data will be preserved and shared (and if any data will not be preserved and shared, which data and why they cannot be shared)
- How, when, and where will the data be preserved and shared?
- Will there be any restrictions on data access? What are they and why? How will they be implemented?
- How and for how long will access be maintained after the life of the project?
- The outcome of previous Data Management Plans, including but not limited to those prepared for NSF-funded research.
- Prior experience in data sharing or management.
- Data sharing/data management resources available to PI/Co-PIs.

### Instructions to Panels:

Panels should discuss data management, including data sharing, as a part of every discussion of Broader Impacts.

(1) Panels should take into account publications and data sharing from prior NSF-funded research as well as other sponsored research.

(2) Panels should evaluate a DMP as adequate or inadequate and provide feedback. (See note about data management experts on panels.) Supporting text should explain the reasons for the evaluation.

(3) Panels should not recommend proposals with inadequate DMPs for the most competitive funding category

### Instructions to Program Staff:

(1) Program staff should ensure that panels have appropriate expertise to evaluate DMPs and that DMPs are discussed in the meeting and that feedback is included in the review.

(2) Program staff should ensure compliance with the DMP with an explicit check-off on the final report.

### Recommendations to SBE and NSF:

(1) SBE should fund research about data management plans, what they have contained, and how PIs have met promises (or not). This research should also investigate and report on equitable access to repositories (by discipline or by institution). It should also investigate whether data have been re-used, and if so, the nature of the re-use (analysis of the same dataset, combining with other datasets).

(2) SBE should support training in data management and the development of tools and facilities for sustainable data sharing, including removing barriers to sharing confidential/ sensitive data.

(3) NSF and SBE should work with major data repositories to ensure that PIs can be compliant with NSF Public Access Policy.

(4) NSF should make Data Management Plans of funded projects public on the NSF website along with their project summaries.

(5) NSF should update annual and final reports so that there is a place to report Digital Object Identifiers (DOIs) for data alongside other outputs resulting from each grant.

(6) NSF should reach out to other federal agencies supporting social, behavioral, and economic research to coordinate and communicate expectations about data sharing, management, and protection of human subjects.

# References

Belmont Forum. 2015. *Data Policy and Principles*. http://www.bfe-inf.org/sites/default/files/doc-repository/BelmontDataPolicyandPrinciples.pdf

Borgman, C.L. 2012. The Conundrum of sharing research data. *Journal of the Association for Information Science and Technology* 63: 1059-1078

Castro, D. and Korte, T. 2015. *Open Data in the G8*. Washington, DC: Center for Data Innovation. http://www2.datainnovation.org/2015-open-data-g8.pdf

Esanu, J.M. and Uhlir, P F. (Eds.). 2004. *Open access and the public domain in digital data and information for science: Proceedings of an International Symposium*. Washington, DC: The National Academies Press.

Fienberg, S.E., Martin, M.E., and Straf, M.L. (Eds.). 1985. *Sharing research data*. Washington, DC: National Academies Press.

Holdren, John P. Memo to the heads of executive departments and agencies. 22 February 2013. "Increasing Access to the Results of Federally Funded Scientific

King, G. 1995. Replication, Replication. *PS: Political Science and Politics* 28: 444-452

National Research Council. 1995. *Preserving scientific data on our physical universe: A new strategy for archiving the nation's scientific information resources*. Washington, DC: The National Academies Press.

National Research Council. 1997. *Bits of power: Issues in global access to scientific data*. Washington, DC: The National Academies Press.

National Research Council. 2009. *Ensuring the integrity, accessibility, and stewardship of research data in the digital age*. Washington, DC: National Academies Press.

National Science Foundation. 1988. *Grant Policy Manual*. NSF 88-47.

National Science Foundation. 1995. *Grant Policy Manual*. NSF 95-26. http://www.nsf.gov/pubs/stis1995/nsf9526/nsf9526.txt.

National Science Foundation. 2007. Sustainable Digital Data Preservation and Access Network Partners (DataNet). Program Solicitation NSF 07-601. http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm.

National Science Foundation. 2011. Grant Proposal Guide. Accessed at http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp.

National Science Foundation. 2015. *Today's Data, Tomorrow's Discoveries*. http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

OECD (Organization for Economic Co-operation and Development). 2007. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. https://www.oecd.org/sti/sci-tech/38500813.pdf

# Appendix A: Recommended Revision of "Data Management Plan for NSF SBE Directorate Proposals and Awards"

### Executive Summary

Since January of 2011, all National Science Foundation full proposals have required a data management plan (DMP) describing how data generated from NSF-supported research will be managed.

The plans are short, no more than two pages, are submitted as a supplementary document, and do not count toward the 15-page limit for proposals. The plans need to address three main topics:

*What data are generated by your research?*

*What is your plan for managing the data?*

*How will you make the data publicly available?*

> **Data** are defined for the purpose of this document as information relevant to, or of interest to social, behavioral, and economic researchers, either as inputs into or outputs from research. They are research materials resulting from primary data collection or creation or derived from existing sources.

It is acknowledged that there are many variables governing what constitutes "data," the management of data, and the sharing of data and that each area of science has its own culture in this regard. The data management plan will be evaluated as part of every proposal. Proposals must include sufficient information that peer reviewers can assess both the data management plan, the data sharing plan, and past performance regarding data sharing. The plan should reflect best practices in the investigators' area of research and should be appropriate to the data they generate. This document is meant to provide guidance for investigators within the Social, Behavioral, and Economic Sciences as they develop their Data Management Plans.

### The Requirement: Include a Data Management Plan (DMP) in Proposals

An appropriate data management plan is required (maximum of two pages) for all full research proposals submitted. This plan is to be included in the Supplementary Documents section of the proposal. It is not part of the 15-page limit for the Project Description. The NSF will not accept any full proposal submitted that is lacking a DMP. Even if no data are to be produced, e.g. the research is purely theoretical or is in support of a workshop, a DMP is required. In this case, the DMP can simply state that no data will be produced.

The plan should describe how the PIs will manage and share data generated by the project. The DMP will be considered by NSF and its reviewers during the proposal review process in consideration of broader impacts. Proposals with DMPs rated at "inadequate" during the review process will not be recommended for the most competitive funding category. Strategies and eventual compliance with the proposed DMP will be evaluated not only by proposal peer review but also through project monitoring by NSF program officers, by Committees of Visitors, and by the National Science Board.

NSF is aware of the need to provide flexibility in assessment of data management plans. In developing a plan, researchers may want to consult with university officials as many universities have explicit data management policies. Some professional organizations also have recommended data management practices (e.g., The American Economic Association at http://www.aeaweb.org/aer/data.php). A resource on preparing a data management plan can be found at ICPSR at http://www.icpsr.umich.edu/icpsrweb/ICPSR/dmp/index.jsp, including some useful examples. Additionally, organizations that offer to store data may also focus on specific types of data. For instance, Open Context (http://opencontext.org/) and the Digital Archaeological Record (http://www.tdar.org/) provide data storage services for the archaeological community. NSF does not require or endorse the use of any specific repository.

**Contents of the Data Management Plan**

The DMP should clearly articulate how the management and sharing of primary data are to be implemented during and after the funding period. It should outline the rights and obligations of all parties as to their roles and responsibilities in the management and retention of research data. It should also consider changes to roles and responsibilities that will occur should a principal investigator or co-PI leave the institution or project. Any costs should be explained in the Budget Justification pages. Specific components are listed below.

*Expected data*. The DMP should describe the types of data, samples, physical collections, software, curriculum materials, or other materials to be produced in the course of the project. It should then describe the expected types of data to be retained and disposed of.

PIs must address the following issues in the DMP:

- The types of data that their project will generate and how they will be protected, preserved and shared;
- Other types of information that should be maintained and shared regarding data, e.g., the way they were generated, analytical and procedural information, and the metadata;
- When and where the data will be preserved and shared;
- Will there be any restrictions on data access? What are they and why?
- How will access be assured and for how long after the life of the project?
- Will linkable identifiers be preserved to facilitate interoperability?
- The outcome of previous Data Management Plans, including but not limited to those prepared for NSF-funded research;
- Any prior experience in data sharing or management;
- Institutional or domain repositories and other resources available to PI/Co-PIs for data preservation and sharing.

*Period of data retention.* SBE is committed to timely and rapid data sharing. However, it recognizes that types of data can vary widely and that acceptable norms also vary by scientific discipline or interdisciplinary research area. SBE is strongly committed, however, to the underlying principle of timely access, and applicants should address how this will be met in their DMP, referencing the applicable community norms.

*Data formats and dissemination.* The DMP should describe data formats, media, interoperability standards, and dissemination approaches that will be used to make data and metadata readily available to others, including such best practices as use of Digital Object Identifiers (DOIs), investigator identifiers (e.g., ORCIDs), and machine-readable open access licenses. Policies for public access and sharing should be described, including provisions for expanding open access as well as appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements.

*Data storage and preservation of access.* The DMP should describe physical and cyber resources and facilities that will be used for the effective preservation and storage of research data and metadata. These can include third party facilities and repositories.

*Additional possible data management requirements.* More stringent data management requirements may be specified in particular NSF solicitations or result from local policies and best practices at the PI's home institution. Additional requirements will be specified in the program solicitation and award conditions. Principal Investigators to be supported by such programs must discuss how they will meet these additional requirements in their Data Management Plans.

### Post-Award Monitoring

After an award is made, execution of the data management plan will be monitored primarily through the normal Annual and Final Report process and through evaluation of subsequent proposals.

*Annual Reports.* Annual reports, required for all multi-year NSF awards, must provide information on the progress on data collection and management and progress on plans for preserving, protecting, and ultimately sharing of the research products. This information could include citations of relevant publications, conference proceedings, and descriptions of other types of data sharing and dissemination of results as well as work with relevant repositories on compliance, assignment of DOIs, and preparation of metadata and other documentation.

*Final Project Reports.* Final Project Reports are required for all NSF awards. The Final Project Report must discuss execution and any updating or revisions of the original DMP. This discussion should describe:

- The types of data produced during the award, how they are being protected, preserved, and shared;
- Other types of information that have been maintained and shared regarding data, e.g., the way they were generated, analytical and procedural information, and the metadata;
- Where are the being preserved and shared and are they or when will they be made publicly available?
- Verification that data will be available for sharing as indicated in the DMP or reasons for exceptions or changes;

- Are there any restrictions on data access? What are they and why?
- How is access being maintained and what are the long-term plans for data preservation and access?

*Subsequent proposals.* Data management outcomes must be reported in subsequent proposals by the PI and Co-PIs under “Results of prior NSF support.” If the PI or Co-PIs have failed to meet or fail to report on the promises of the Data Management Plan from previously funded projects, the subsequent proposal(s) will not be eligible for NSF funding.

## Resources

The American Economic Association http://www.aeaweb.org/aer/data.php

Data FIARport: http://datafairport.org/

Data Preservation Alliance for the Social Sciences (Data-PASS) http://www.datapass.org/

Economic and Social Research Council of the UK

- Research Data Policy: http://www.esrc.ac.uk/files/about-us/policies-and-standards/esrc-research-data-policy/
- Guidance to Peer Reviewers: http://www.esrc.ac.uk/files/funding/guidance-for-peer-reviewers/data-management-plan-guidance-for-peer-reviewers/

Digital Archaeological Record http://www.tdar.org/

ICSU World Data System https://www.icsu-wds.org/

Inter-university Consortium for Political and Social Rresearch (ICPSR) https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/tools.html

The Odum Institute Data Archive. http://www.odum.unc.edu/odum/contentPrimary.jsp?nodeid=7

Open Context http://opencontext.org/

Open Science Framework: https://osf.io/

UK Data Service

- General guidance on data management: https://www.ukdataservice.ac.uk/manage-data/plan
- Guidance for PIs on Creating Data Management Plans: https://www.ukdataservice.ac.uk/manage-data/plan/dmp-esrc
- Guidance on costing data management: https://www.ukdataservice.ac.uk/manage-data/plan/costing

## References

Council on Governmental Relations, Access to and Retention of Research Data: Rights and Responsibilities, March 2006. http://206.151.87.67/docs/CompleteDRBooklet.htm

Holdren, John P. Memo to the heads of executive departments and agencies. 22 February 2013.

"Increasing Access to the Results of Federally Funded Scientific Research." https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

National Science Foundation. 1988. *Grant Policy Manual*. NSF 88-47.

National Research Council. 1995. *Preserving scientific data on our physical universe: A new strategy for archiving the nation's scientific information resources*. Washington, DC: The National Academies Press.

National Science Foundation. 2011. Grant Proposal Guide. Accessed at http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp.

National Science Foundation. 2015. *Today's Data, Tomorrow's Discoveries*. http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

National Science Foundation, Proposal and Award Policies and Procedures Guide, January 2016. http://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg_index.jsp

Office of Management and Budget, Circular A-110, September 30, 1999. White House Website, OMB Home. http://www.whitehouse.gov/omb/circulars/a110/a110.html

## Appendix B: Workshop Participants

***Working Group***

Steven Ruggles, Chair
Director, Minnesota Population Center
Regents Professor of History
University of Minnesota

Karen Adolph
Director, Databrary
Professor of Psychology and Neural Science
New York University

Robert S. Chen
Director, Center for International Earth
Science Information Network
Columbia University

Barbara Entwisle
Vice Chancellor for Research
Kenan Distinguished Professor of
Sociology
University of North Carolina

Janet Gornick
Director, Luxembourg Income Study
Professor of Political Science and
Sociology
The Graduate Center, City University of
New York

Myron Gutmann
Director, Institute of Behavioral Science
Professor of History
University of Colorado, Boulder

***Workshop Coordinators***

Catherine Fitch
Associate Director and Research Scientist
Minnesota Population Center
University of Minnesota

Gina Rumore
Program Development Director
Minnesota Population Center
University of Minnesota

***Workshop Recorder***

Cate Sturtevant
Carolina Population Center
University of North Carolina

***Participants***

George Alter
Director, Inter-university Consortium for
Political and Social Research
University of Michigan

Henry Brady
Dean, Goldman School of Public Policy
University of California, Berkeley

Tom Carsey
Director, Odum Institute
Thomas J. Pearsall Distinguished Professor
in the Department of Political Science

Margaret Levenstein
Directory, Michigan Research Data Center
Research Scientist, Survey Research
Center
University of Michigan

Brian MacWhinney
Professor of Psychology
Carnegie Mellon University

Steven Manson
Professor of Geography
University of Minnesota

Lisa Cliggett
Professor of Anthropology
University of Kentucky

Rachel Croson
Dean, College of Business
University of Texas, Arlington

William A. Darity
Samuel DuBois Cook Professor of Public Policy
Duke University

Rick Gilmore
Associate Professor of Psychology
The Pennsylvania State University

Philip Kasinitz
Presidential Professor of Sociology
Graduate Center, City University of New York

Joanna Morris
Associate Professor of Cognitive Science
Hampshire College

Kathleen Mullan Harris
James Haar Distinguished Professor of Sociology
University of North Carolina

Seth Sanders
Director, Duke Center for Population Research
Professor of Public Policy in the Sanford School
Duke University

Ashley Sorgi
Dissemination Coordinator, Add Health
University of North Carolina

Matthew Woollard
Director, UK Data Archive and the UK Data Service
University of Essex

# Appendix C: Workshop Agenda

***Thursday, 1/28***

9:00 AM Introduction by Steve Ruggles: Charge to Committee and Organization of the Workshop

9:10 AM Introductions: participants will get one minute to state their names, affiliation, their relationship with data, and the one most important point they would like to see in our report to NSF

9:45 AM *What is data and why share it?* [Moderated by Barbara Entwisle]

10:30 AM Coffee Break
*Coffee, tea, and snacks provided*

10:40 AM *Scope of data management policy: What kinds of data does it apply to?* [Moderated by Janet Gornick]

12:15 PM Lunch Service
*Lunch will be provided in the workshop room*

1:00 PM *Ethical issues and data sharing* [moderated by Karen Adolph]

2:30 PM Coffee Break

2:40 PM *Logistical issues and data sharing* [moderated by Robert Chen]

4:00 PM Personal Time [Working Group meets to analyze discussion and briefs and to prepare draft of Recommendations to Directorate for Social, Behavioral and Economic Sciences]

7:00 PM Dinner at Mediterranean Deli
*Restaurant is 10 minute walk from hotel; hotel shuttle is available*
*(Address: 410 W Franklin St, Chapel Hill, NC 27516)*

***Friday, 1/29***

9:00 AM *Recommendations to SBE* [moderated by Myron Gutmann]

| | |
|---|---|
| 10:30 AM | Coffee Break<br>*Coffee, tea, and snacks provided* |
| 10:40 AM | Discussion continues on NSF recommendations |
| 12:15 PM | Lunch Service<br>*Lunch will be provided in the workshop room* |
| 1:30 PM | Workshop Closes |
| 2:00 PM | Working Group reconvenes to outline final Recommendations to Directorate for Social, Behavioral and Economic Sciences |

# Appendix D: Directorate of Social, Behavioral, and Economic Sciences

### Behavioral and Cognitive Sciences (BCS)

Anthropological Sciences

- Archaeology and Archaeometry
- Biological Anthropology
- Cultural Anthropology

Geography and Environmental Sciences

- Dynamics of Coupled Natural and Human Systems
- Geography and Spatial Sciences Program
- Long-Term Ecological Research

Psychological and Language Science

- Cognitive Neuroscience
- Developmental and Learning Sciences
- Documenting Endangered Languages
- Linguistics
- Perception, Action, and Cognition
- Social Psychology

### National Center for Science and Engineering Statistics (NCSES)

### Social and Economic Sciences (SES)

- Decision, Risk and Management Sciences
- Economics
- Law and Social Sciences
- Methodology, Measurement, and Statistics
- Political Science
- Science of Organizations
- Science, Technology, and Society
- Sociology

## Appendix E: Example Types of Data Used in the Social, Behavioral, and Economic Sciences

### *Anthropology*

- Recorded and transcribed interviews
- Written field notes
- Photos
- Maps: hand drawn or digital
- Recordings: audio or visual
- Quantitative data
- Secondary sources
  - media
  - interactive web-based text
- Data that has been analyzed using qualitative data analysis software

### *Cognitive Science*

- Outcome measures
  - response times
  - error rates
- Audio or video recordings of verbal or motor responses
- Online measures
  - eye tracking
  - mouse tracking
  - event related potentials
- Functional brain imaging

### *Developmental Science*

- Video or audio recordings
- Transcripts of verbal and behavioral exchanges
- Questionnaires
- Computer based data
  - touch screen
  - eye tracking
  - tutors
- Text-based flat files for statistical analysis

### *Economics*

- Experimental
- Observational
- Survey
- Linked datasets
- Federal (e.g., census microdata)
- Corporate data (shared only with a nondisclosure agreement)

### *Geography*

- Models
- Geographic Information System (GIS) data
- Global Positioning System (GPS) data
- Field notes

- Remote sensing data and imagery
- Administrative data

***Psychology***

- Audio recordings
- Video recordings
- Transcripts
- Surveys
- Standardized tests
- Experimental data

***Public Policy***

- Survey
- Interviews
    - audio
    - video
- Methods used for data generation

***Sociology***

- Survey data
- Contextual files constructed and maintained by survey administrators
- Administrative data
- In depth interviews
    - recordings
    - transcripts
- Ethnographic field notes
- Biomarker data collected from physical measurements
- Biological specimens from respondents

# Appendix F: Pre-Workshop Data Briefs

## Sharing Identifiable Data in the Developmental and Learning Sciences

January 11, 2016

Karen E. Adolph
New York University
Databrary.org

Rick O. Gilmore
Penn State University
Databrary.org

### What is data?

*- What does your research community consider to be "data" when it comes to writing a Data Management Plan? (e.g., Does your community distinguish between source data and processed data used for analyses? Are raw source data useable or interpretable? How can data provenance and workflow be characterized adequately?)*

Most developmental scientists consider data to be video or audio recordings, transcripts of verbal and behavioral exchanges, questionnaires, computer-based data (touch screen, eye tracking, tutors, etc.), and text-based flat-files used for statistical analyses. Video or audio recordings are raw source data, which are evaluated by human coders to produce processed data for analyses. The raw recordings are, in most cases, useable and interpretable by others without extensive metadata beyond the characteristics of the people recorded and the setting. Workflows for video/audio data are idiosyncratic, the coding tools are diverse and largely incompatible with one another, and data provenance is rarely documented well if at all. Similarly, workflows for questionnaire and computer-based data are idiosyncratic and data provenance is rarely documented.

### Why share data?

*- What is the overall goal of data sharing from the perspective of your community? What are the incentives and disincentives for data sharing from the point of view of individual researchers?*

The overarching goal of data sharing in the developmental and learning sciences is and should be to discover more, faster. At present, there are more disincentives than incentives. Incentives include a desire to contribute to the good of the research community and to garner goodwill from granting agencies. The disincentives include the additional time and money required to prepare data for sharing, confusion and lack of knowledge about what to share, how and where to share data, and the lack of a research culture that values data sharing, data publication, or data reuse.

*- How can data be more widely used? Should data use be restricted to research and educational/informational purposes? Should data sharing also involve use for commercial purposes (both NSF and NIH have commercial entities)? Are there ways to expand usage of data while maintaining the integrity of the research process?*

Data use (and reuse) must abide by the permissions granted by participants and their parents or guardians. Developing and implementing consistent standards for seeking and communicating permissions granted about data sharing will help to expand more widespread use. Commercial uses may be contemplated or encouraged, if appropriate permission has been given, but should not be the highest priority.

**Scope of data management policy: What kinds of data does it apply to?**

*- How can data management plans better capture the opportunities and challenges of providing access to data to researchers for your areas of research?*

NSF could suggest and endorse a set of data repositories that researchers could choose among in developing their Data Management Plans.

*- How should data management plans address qualitative (e.g., raw research video, audio files, transcripts) data?*

In many respects, qualitative data are more valuable for reuse by others and more easily documented than are quantitative data. These types of data are extraordinarily rich and allow researchers to address questions outside the scope of the original study. NSF should strongly encourage researchers to find suitable outlets for these sorts of data.

*- Are there new or emerging sources of data or methods in your field that are or likely to be constrained by existing data policies, norms, or practices? What needs to be changed? Do these new sources and methods raise important ethical issues that need to be addressed?*

Many of the most interesting and exciting questions in the developmental, psychological, and neural sciences involve the integration and analysis of multiple data sources across levels of analysis. Maximizing the potential for discovery will require new technologies for storing, linking, and analyzing diverse datasets, and careful consideration of the risks of participant reidentification posed by data linking.

**Sensitive data: ethical issues and data sharing**

*- Who should be allowed to access data? Is it global? Do you need ethics training? Does it depend on the nature of consent given by participants? What kind of institution is the applicant in?*

Because Databrary stores and shares identifiable video and audio recordings, it limits access to researchers who are specifically authorized by an institution. The institution takes responsibility for ensuring that its researchers have ethics training and seek appropriate approvals to conduct research on shared data. This access model works well except where researchers are not affiliated with an institution.

*- How can participants' rights be respected? If participant permission is required, what should participant permissions include? Are any data so sensitive that no form of access by other investigators is possible?*

The well-established principles of informed consent should apply. Participants should be asked for their permission before data are shared. Databrary has developed template language for sharing recordings (https://databrary.org/access/policies/release-template.html) that researchers may adopt. Participant permissions for sharing should *not* enumerate the kinds of future reuses of the data. Some data may be so sensitive that data sharing is ill-advised.

**Logistics: if you have a plan, who implements it, pays for it, enforces it, etc.?**

*- Does your community have repositories (digital libraries) for curating, storing, and serving data? Do you think these repositories should meet community criteria for trustworthiness? Which repositories are trusted? Should repositories be mandated or sanctioned by NSF? If so, what are the criteria for NSF-sanctioned repositories? Who should pay for data curation and*

*storage? How can long-term access be financed, especially for data that are expensive to maintain or manage? Relatedly, how can file formats be kept up to date long-term?*

The developmental science community has an emerging array of repository options. These repositories should meet community criteria for trustworthiness and other criteria. To our knowledge, there are little data about which repositories are "trusted" by the community, but several repositories (ICPSR, TalkBank/CHILDES, NDAR) have long-standing histories and significant numbers of datasets. Newer repositories (Databrary, OSF, Dataverse) are growing rapidly.

Repositories should be reviewed and sanctioned by NSF. The criteria should include security practices; support for search across and within datasets by participant characteristics, settings, and tasks or measures; ease-of-use/user support; institutional support, and other criteria as determined by NSF and the research community. Most of the existing data repositories for developmental science are sustained through continual grant funding. This is not a practical long-term plan. Institutional contracts cannot support a large repository. The NSF and NIH should finance the costs of long-term access. Specific set-asides within project grant budgets could be made based on the volume or complexity of data to be collected and shared. A portion of those funds would go to the NSF-sanctioned repositories to support data curation and storage services.

*- How can NSF ensure that the promises of data management plans are actually carried out?*

Proposals from applicants who have received NSF funding in the past should report on how data from prior awards were shared, pursuant to the data management plan(s) in those prior proposals.  This information should enter into the evaluation process for new proposals. NSF could set aside a pool of funds to provide top-offs to grantees that have track records of data sharing or particularly well-thought-out data management/sharing plans. Those funds would be used to support curation and storage services. Since the implementation of data management plans often depends on institutional resources, not just individual investigator capabilities, NSF could evaluate data management/sharing practices across an institution's NSF portfolio and provide incentives to those institutions that have a history of providing investigators exceptional support for data management and sharing.

*- Should NSF allow data to be embargoed, and if so, under what conditions and for how long? How should NSF deal with longitudinal data collection that may require months or years before the original researchers can analyze or publish the data?*

NSF should allow data to be embargoed, but with clear limits. For example, NSF might permit data to be embargoed for up to a year after the end of an award. Ultimately, questions must be addressed about who "owns" the data a researcher collects under an NSF award and what "rights" over the data a researcher can legitimately maintain.

## Comments for NSF Data Access Workshop

George Alter
ICPSR, University of Michigan
January 21, 2016

In addition to the responses below to the questions posed by the organizers, I would like to point out an important feature of the "Notice of Proposed Rulemaking for the Federal Policy for the Protection of Human Subjects for the Common Rule" issued on September 8, 2015. In some circumstances the proposed rules impose more requirements on data that will be shared than on data used only by the original investigator. For example, secondary research on data that include identifiers or sensitive information may require a written consent [NPRM §ll.104(f)(1))], even though the same data used by the primary investigator may not [NPRM §ll.104(e)(2))]. This difference reflects the view that subjects should consent to all the ways that their data might be used. NSF policy should insist on data sharing, even when the investigator must take additional steps to meet the requirements for secondary analysis. NSF should also be actively involved in the development of templates for "broad consent" to be developed by HHS under the proposed rules.

- *How can data provenance and workflow be characterized adequately?)*

Provenance of social science data is generally described as a narrative. These narratives are usually incomplete, and they are not machine actionable. There is great potential for tools and workflows that would automatically collect provenance metadata, which would reduce costs for data creators and improve the documentation received by data users.

- *How should a Data Management Plan address linkages to data not funded by NSF? Such data may be proprietary. They may also be governed by other entities and rules.*

Journals in Economics have already developed policies for authors who cannot share proprietary data from commercial sources. Authors are expected to explain how other researchers can apply for access to the same data, and they are required to make available the program code that was used to produce their results. NSF should adopt similar policies. NSF could go beyond these policies by encouraging entities that share proprietary data for research to allow future researchers to use the same data under the same terms.

NSF should also issue guidance on the agreements that researchers make with commercial entities for use of proprietary data. With funding from the Sloan Foundation, ICPSR commissioned a study of these agreements by two experienced university lawyers, Alex Kanous and Elaine Brock. They found that many of these agreements are poorly conceived, lack important information (such as a description of the data covered by the agreement), and include unnecessary restrictions on researchers. Kanous and Brock also wrote a model data use agreement with annotations explaining key aspects of the agreement. These documents are available at:

Kanous, Alex; Brock, Elaine. Contractual Limitations on Data Sharing. Report prepared for ICPSR as part of the "Building Community Engagement for Open Access to Data" project. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2015. DOI: 10.3886/Contractual Limitations Data Sharing

Kanous, Alex; Brock, Elaine. Model Data Sharing Agreement. Customizable model created as part of the "Building Community Engagement for Open Access to Data" project. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2015. DOI: 10.3886/Model Data Sharing Agreement

NSF is welcome to refer researchers to these documents for guidance in obtaining proprietary data.

- *How can participants' rights be respected? If participant permission is required, what should participant permissions include? Are any data so sensitive that no form of access by other investigators is possible?*

The NPRM promises that HHS will issue guidelines on “broad consent” that investigators can use to enable a wide range of secondary research. Research projects like ANES, GSS, and PSID already have experience obtaining broad consent from participants.

We now have many years of experience sharing sensitive data in data enclaves and other secure environments, and we have been successful sharing very sensitive data.

- *Does your community have repositories (digital libraries) for curating, storing, and serving data? Do you think these repositories should meet community criteria for trustworthiness? Which repositories are trusted? Should repositories be mandated or sanctioned by NSF? If so, what are the criteria for NSF- sanctioned repositories?*

The social sciences are well served by data repositories. The Data Preservation Alliance for the Social Sciences (DataPASS) is a partnership of eight U.S. social science data repositories. DataPASS maintains a common catalog and shares best practices for data curation and archiving. DataPASS partners sign a succession agreement that assures access to data will continue if any partner becomes insolvent.

There are currently four international systems for certifying “trusted digital repositories”: the Data Seal of Approval, the ICSU World Data System, Nestor (Network of Expertise in Long-term Storage of Digital Resources), and ISO 16363 Audit and certification of trustworthy digital repositories. Data Seal of Approval is a self-audit system with 16 guidelines. ISO 16363 involves an independent auditor evaluating more than 100 criteria.

NSF should require repositories to meet community standards. Since standards for trusted repositories are already being administered by national (DataPASS) and international bodies (Data Seal of Approval, ICSU, ISO), NSF does not need to develop its own criteria for sanctioning repositories. A policy that requires repositories to seek certification under an established would be sufficient.

- *Who should pay for data curation and storage? How can long-term access be financed, especially for data that are expensive to maintain or manage? Relatedly, how can file formats be kept up to date long-term?*

The issue of funding long-term access to data is extensively discussed in

Ember, Carol; Hanisch, Robert. Sustaining Domain Repositories for Digital Data: A White Paper. Output of the workshop, "Sustaining Domain Repositories for Digital Data," Ann Arbor, MI, June 24-25, 2013. Ann Arbor, MI: Inter- university Consortium for Political and Social Research,
2013. DOI: 10.3886/Sustaining Domain Repositories Digital Data

Ember and Hanisch compares a number of alternative funding models, including membership dues, submission fees, institutional support, sponsored projects, commercial services, user fees, overhead, and infrastructure. Most funding models have drawbacks, and models that rely on fees are especially inequitable to researchers at under-funded institutions. Our European competitors have decided that data repositories are essential scientific infrastructure, and they have developed a legal mechanism, the European Research Infrastructure Consortium (ERIC), to finance long term projects like data repositories. The Council of European Social Science Data Archives (CESSDA) is one of the first ERIC projects, and it is funded directly by government research councils in 15 countries. NSF has been reluctant to fund data repositories, but other government agencies in the U.S. provide funding to data repositories directly or through grants and contracts (e.g. NIH, NOAA, NASA).

Maintaining file formats for the long term is a problem that has already been solved in a number of disciplines. The key is storing data and metadata in a standard non-proprietary format that is easy to manage. Two formats of this kind are internationally accepted for most kinds of social science data. The Data Documentation Initiative (DDI) is a standard for describing microdata. Statistical Data and Metadata Exchange (SDMX) is used by national statistical agencies to describe aggregate data, like censuses. The organizations that manage DDI and SDMX have a cooperation agreement to develop links between the two standards. Both DDI and SDMX are implemented in XML, but they are also developing other implementations, like RDF.

- *How can NSF ensure that the promises of data management plans are actually carried out?*

The simplest and cheapest way for NSF to ensure that data management plans are carried out is to make them public. Currently DMPs are considered confidential information. NSF could designate DMPs as one of the parts of an application that will be published when an award is made. PIs would be subject to community pressure if they do not live up to their promises.

- *Should NSF allow data to be embargoed, and if so, under what conditions and for how long? How should NSF deal with longitudinal data collection that mayrequire months or years before the original researchers can analyze or publish the data?*

PIs who create data have a legitimate expectation to be the first to use those data. Some journals and professional associations allow postponing the release of data to

allow data creators time to analyze new data. For example, the American Political Science Association "Guide to Professional Ethics" allows data access to be delayed by one year after results have been published. However, NSF policy should set a limit on the delay in sharing funded data and discourage the piecemeal release of data over many years.
Journal replication requirements only apply to data used to produce published findings. This results in the release of incomplete subsets of a larger data collection, which hinders secondary research and new analyses.

NSF should insist on prompt release of costly data that have been funded for their value to the research community. We have many examples of large and complex data collection projects that make their data available without long delays, such as PSID, HRS, and AddHealth.

## NSF Data Management Plan Workshop

**Thomas M. Carsey, Professor of Political Science, UNC-Chapel Hill**
**Director, Odum Institute for Research in Social Science**

### What is data?

Data constitutes the measurable/observable attributes of actors, objects, and processes that are the subject of study. Data provides the link between the processes we study and our theories about those processes. Data should include raw and processed data, metadata, descriptions of data collection, cleaning, and analysis. It should also include any computer code used in the collection, cleaning, and analysis of data. Data constitutes a record of everything that was collected and done by researchers in order to produce the knowledge claims they make.

### Why share data?

Sharing data makes the research process more transparent and replicable. This protects the integrity of scholarly research. It allows reuse for new purposes, but also allows communities of scholars to build a body of knowledge. The major disincentives I have heard voiced generally focus on protecting a scholar's ability to publish out of data they collected and protecting confidentiality/privacy of research subjects.

Data should be shared as widely as is feasible. Better cyberinfrastructure needs to be developed that can protect sensitive data and the integrity of data while making it easier to discover and access. Federated systems like the Dataverse provide a good model for sharing, though we need better resources for curating sensitive data.

### Scope of data management policy: What kinds of data does it apply to?

I think stronger general principles will make it easier to apply data management policy to widely diverse data. The same standards for transparency and replicability should be set for all research – could one researcher reproduce the results of another given the data, metadata, etc. provided? Some research efforts could provide more certainty in answering that question than others, and that uncertainty should be stated explicitly as part of the metadata that characterizes the project.

It is a separate question to decide how much of the data and metadata can be shared publicly, shared only under an authorized data use agreement, or cannot be shared for privacy, ethical, commercial, or other reasons. However, just because data cannot be shared does not mean that the same care toward transparency should not be taken. The Odum Institute, which operates a trusted digital repository for the social sciences, provides for the independent execution and monitoring of data management plans. The Odum Institute also currently provides independent verification of data and computer code for quantitative papers published in the *American Journal of Political Science* and the journal *State Politics and Policy Quarterly*.

## Sensitive data: ethical issues and data share

Access to sensitive data will be case dependent, but could be thought of in term of categories. The Data Tags project (http://datatags.org/) developed by Latanya Sweeney at Harvard allows researchers to classify the sensitivity of their data into one of six categories. We have a proposal under review that would attach Data Tags to data that could be used to automatically control access, sharing, etc. by using a file management system used here at UNC called iRODS in combination with the Dataverse. Other options could include the production of synthetic data that could be shared. Researchers could develop models using synthetic data alone, or they might submit those models to run on real data behind a firewall and receive a measure of the model's performance on the real versus synthetic data.

New ethics training on consent and data sharing should be part of researcher training. It should also be mandatory for staff working with data at any data repository.

## Logistics: if you have a plan, who implements it, pays for it, enforces it, etc.?

Social scientists can use repositories like ICPSR, the Odum Institute, or now the Qualitative Data Archive at Syracuse. One problem is the lack of full integration of the software used to manage different archives. Odum uses a platform called the Dataverse, which is open source and the newest version is very API compatible. Dataverse developers at Harvard and in Odum have been working with the Open Journal System and the Open Science Center at UVA to further link such tools.

Dataverse converts files from many standard formats into a number of common formats, including simple CSV files, to promote the long-term preservation and access of data. It also produced a unique statistic that guarantees the integrity of the data across formats.

A new organization now provides a mechanism for archives to earn a Data Seal of Approval (http://datasealofapproval.org/en/). The Odum Institute was the first to receive this designation in the U.S. – there are now five. New standards are also emerging for auditing archives. These efforts should be strongly supported. NSF does not need to mandate standards, but should monitor standards and encourage researchers to use archives that meet approved standards in the relevant field of study. Standards will necessarily evolve over time and differ across domains – NSF should not try to develop a one-size-fits-all standard.

The best way to ensure compliance with data management plans is to: a) automate as much of it as is possible, and b) use third party groups rather than relying on individual PIs to self-monitor. The Odum Institute helped produce a software program called SafeArchive that audits the integrity of federated data systems. It can report on the status of data files and any back-up copies. Our next proposal is to leverage the Dataverse and iRODS to write software that can translate requirements written into data management plans into rules that can be attached to data files and subsequently audited for compliance.

NSF could invest more in the development of certification standards and software tools like these to increase not just compliance with data management plans but the development of better data management plans in the first place.

Embargoes of data to give researchers some time to produce papers seem reasonable. Another alternative would be to work to change professional standards such that data collection and sharing was more effectively rewarded. Standards for citation to datasets (and software and other products of research) have emerged – the Dataverse generates a formal citation for every dataset deposited. We need to promote the citation of datasets and reward the production and distribution datasets. Before tools like Google Scholar made it easy, nobody measured citation counts, but now they do. If the production of an impactful dataset was valued by promotion and tenure committees as much as the production of an impactful article, scholars would share more data.

NSF could support proposals to develop tools that made data citation and sharing easier. NSF could also incentivize the development of better tools and datasets themselves if more calls for proposals were issued that focused on these two goals. Historically NSF has been less supportive of general infrastructure development. New programs like Data Infrastructure Building Blocks (DIBBs) are a good start, but even these programs can sometimes demand motivation by a specific scientific question rather than a more general assertion of serving a community of scholars.

Paying for all of this is challenging. NSF could certainly provide more support to the development of tools and standards of best practice. NSF could also fund data curation through each grant it awards, potentially requiring a budget line for data curation in every proposal. Repositories could charge researchers to deposit data or researchers who use the data from a repository, but such charges would create disincentives for both practices. NSF could provide support similar to center grants for data centers. There should be scores of them – maybe hundreds – rather than a single repository, but they could be designed using a federated model to permit easy search and discovery. If universities are going to be given the task, NSF might allow for increased F&A rates if a formula is provided that directs the additional F&A to data management, sharing, and preservation.

## NSF Data Management Plan Workshop
## Lisa Cliggett (Professor, U. Kentucky) – Anthropology

*What is data?*

- For Anthropologists the most common types of raw “data” include: recorded and transcribed interviews, written field notes, photos, maps (hand drawn or digital), recordings (audio and visual), quantitative data, secondary sources (media, interactive web based text, etc). In writing the DMP, most anthropologists would speak about this material. Data that has been analyzed using qualitative data analysis software should also be included in the DMP (in my opinion), but frequently, scholars will discuss the actual analysis as “data management,” not as data to be managed. This is a result of confusion over the larger goal of the DMP.
- A key challenge in data sharing in anthropology centers on the deep contextual knowledge researchers have as a result of being the primary data collector (anthropologists tend to do their own data collection, including deep immersion within the research communities). Thus, a common refrain among anthropologists when asked about data sharing is: “no one will understand my data” because without the context knowledge held as “head notes,” that data is meaningless.
- The other major challenge in sharing anthropological is the issue of confidentiality combined with data linkage across file types and time. In general, data linkage (across file types, time frames and projects) is extremely important for anthropologists. We return to our field sites year after year; we work with families over multiple generations; although we may have different research questions, data from an earlier project can inform the new project in the same place, etc. The linkage question is very much a technological / software question.

*Why share data?*

- Most anthropologists have little interest in sharing their data. One reason is the above point about data being useless without the context of “head notes” necessary for interpretation. Other reasons include: concern for privacy and confidentiality of research communities and individuals, notions about data ownership (a sense that qualitative data is owned by the researcher), concerns about being “outed” (Margaret Mead; Darkness in Eldorado) or scooped.
- If the barriers to qualitative researchers archiving and sharing data were resolved, a few of the most obvious ways that data could be reused include: restudies and longitudinal studies; new studies considering new questions, but drawing on detailed raw data; study communities accessing and using data for community advocacy, development, historical purposes; other disciplines asking different questions of the data (political science, religion, history of science, etc).

*Scope of data management policy: What kinds of data does it apply to?*

- How should data management plans address qualitative (e.g., raw research video, audio files, transcripts) data?
    - At the least, qualitative data needs to be preserved – even if researchers claim their data cannot be shared (due to confidentiality or other reasons). Creating a preservation plan, and following it, ensures the possibility of sharing at a later date.

    - To help the current cohort of scholars imagine and draft a DMP, funding agencies could provide examples of what managing, preserving and sharing qualitative data looks like. Ideally, methods classes will help current and future cohorts of students and scholars embrace the DMP.
- How should a Data Management Plan address linkages to data not funded by NSF?
    - I think this will be very difficult. Many anthropologists conduct research outside of any funding agency. In those cases the sense of individual data ownership makes the likelihood of adhering to any mandate quite low.
    - Once anthropologists understand the value of preserving data, they may come to the realization that all their data needs “management” for preservation and sharing, meaning that linkages _should_ be included in data management.
    - For my broader point on data linkage, see bullet 3 under “what is data?”

*Sensitive data: ethical issues and data sharing*

- Who should be allowed to access data?
    - This is another central concern for anthropologists.
    - I think the question is more “who can access what kinds of data?” Each piece of qualitative data could have different access protocols. Anonymized, de-linked data could, perhaps, be global access with no oversight. Identifiable data may require submitting a new IRB by the secondary user.
    - The point about “consent” of research participants speaks specifically to the challenge of IRB. IRBs (in my experience) still do not know how to handle data sharing – especially of qualitative research. IRBs need training in HOW they can approve projects, particularly qualitative projects, which intend to share data. Researchers and IRBs need to develop an appropriate informed consent that applies well to qualitative data.
- How can participants’ rights be respected? Are any data so sensitive that no form of access by other investigators is possible?
    - See point above about IRB.
    - Yes, sometimes data is too sensitive to share. Ie: illegal activities in a known study community.

*Logistics: if you have a plan, who implements it, pays for it, enforces it, etc.?*

- Does your community have repositories (digital libraries) for curating, storing, and serving data? Do you think these repositories should meet community criteria for trustworthiness? Which repositories are trusted? Should repositories be mandated or sanctioned by NSF? If so, what are the criteria for NSF-sanctioned repositories? Who should pay for data curation and storage? How can long-term access be financed, especially for data that are expensive to maintain or manage? Relatedly, how can file formats be kept up to date long-term?
    - The repository system is not consistent across institutions. My university’s repository only “preserves” data; they do not provide access for data sharing. A different university web based portal can be used for access, but they do not serve as a long term preservation repository. Alternatively, ICPSR (Michigan) is a one-stop repository with access (and management of different levels of access).
    - It would be helpful if NSF identified (and funded perhaps) some of the best repositories for different kinds of data. Because of inconsistencies across institutions, researchers may not have good options at their home institution.

    - If researchers are educated about the steps needed to save data in durable formats, there is somewhat less burden on institutions to do detailed curating. The cost of storage and maintaining infrastructure is different. Universities will want to put the cost on researchers, by adding an additional % to the overhead on each grant. What will NSF do when that happens?

- How can NSF ensure that the promises of data management plans are actually carried out?
    - In principle, once data is archived, it should have a searchable digital ID number. Researchers could create the actual data ID (as preparation for depositing data, if not actually having deposited data) as part of the project final report. NSF could follow up periodically to see if data has actually been submitted. Future grants could be contingent on having archived data.
- Should NSF allow data to be embargoed, and if so, under what conditions and for how long?
    - For qualitative researchers, embargoing must be allowed. Until there is a proven system of anonymizing, yet maintaining linkage/ context, anthropologists will want the option to embargo (for the life of the subject, until memory of particular events have faded into a blur, until decedents of key subjects have died, etc).
    - Students in particular must be able to embargo. Their career depends on publishing from their data; they cannot risk having their data used before the can complete the publication plan.

## Broad Themes and Questions to Consider

**Rachel Croson**
**Dean, College of Business**
**University of Texas at Arlington**

*What is data?*

- What does your research community consider to be "data" when it comes to writing a Data Management Plan? (e.g., Does your community distinguish between source data and processed data used for analyses? Are raw source data useable or interpretable? How can data provenance and workflow be characterized adequately?)

  **Croson: Within the subareas of economics, data has multiple meanings. Some data is created or collected by the researcher (experimental, observational or survey). That data is then processed (cleaned, transformed), analyses are run and output is reported in publications. It is typical for a young researcher to invest 2-3 years in developing a unique dataset, and they are expected to publish 5 or more papers from it. Sometimes this contribution involves linking (by hand) existing datasets. Other data is federally protected (e.g. microdata from Census) and cannot be shared. Still other data is confidentially given to the researcher or collected by the researcher (e.g. from a firm, with explicit nondisclosure agreements). Within business research, confidential data from firms is even more common.**

  **Metadata can include sampling strategies, experimental instructions and information on any transformations that occurred. Sharing metadata is not typically controversial, although see a recent heated debate on the creation of an instrumental variable: http://www.wsj.com/articles/SB113011672134577225.**

*Why share data?*

- What is the overall goal of data sharing from the perspective of your community? What are the incentives and disincentives for data sharing from the point of view of individual researchers?

  **Croson: Economists are actively concerned about replication, and sharing data in support of replication is likely to be a compelling argument for this audience. Concerns include the requirement to share data that has been obtained under conditions of confidentiality, and the potential drying up of those sources if data sharing became mandatory. Other concerns include the need to reap the rewards (via publication) from investment in data creation and acquisition. The solution which has emerged from the American Economic Review (AER) is that proprietary data need not be shared. Data which forms the basis of the analyses in the paper, as well as the econometric code will be shared. https://www.aeaweb.org/aer/data.php, but this may be a subset of the data actually collected.**

- How can data be more widely used? Should data use be restricted to research and educational/informational purposes? Should data sharing also involve use for commercial purposes (both NSF and NIH have commercial entities)? Are there ways to expand usage of data while maintaining the integrity of the research process?

  **Croson: I believe that any sharing policy needs to include flexibility for proprietary (or otherwise confidential) data. It needs to allow for faculty who invest in constructing unique datasets to benefit from that investment, thus a moratorium will be important. If the data was collected from sources who provided it explicitly for research purposes (e.g. under an IRB which stated this), then its use needs to be restricted to research purposes.**

*Scope of data management policy: What kinds of data does it apply to?*

- How can data management plans better capture the opportunities and challenges of providing access to data to researchers for your areas of research?

  **Croson: I believe that data management plans need to be flexible. Most of what I've seen has attempted to fit all pegs into one round hole, and given the diversity of the SBE fields, this is likely to cause more harm than benefit.**

- When is a complementary policy on non-digitized data (e.g., biological specimens) applicable/necessary?

  **Croson: This is a great question. The rise of neuro-economics and bio-economics (which uses biological specimens) creates significant problems for data management. This area is currently quite small, and most of this biological data is eventually digitized and could be shared. But I could imagine a world where this would grow.**

- How should data management plans address qualitative (e.g., raw research video, audio files, transcripts) data?

  **Croson: They should be consistent with the confidentiality and human subjects protections offered to the participants or data source.**

- How should a Data Management Plan address linkages to data not funded by NSF? Such data may be proprietary. They may also be governed by other entities and rules.

  **Croson: I believe that appropriate constraints on data sources need to be respected, including those funded by the NSF and those not funded by the NSF.**

- Are there new or emerging sources of data or methods in your field that are or likely to be constrained by existing data policies, norms, or practices? What needs to be changed? Do these new sources and methods raise important ethical issues that need to be addressed?

  **Croson: The emerging area of field experiments, especially in developing economies (sometimes called Randomized Control Trials (RCTs)), has raised some data challenges. First, the question of data confidentiality and sharing is highlighted in**

**this methodology, because collecting the data often requires significant political agreements and negotiations. Second, the question of metadata has been expanded in this area. In response to concerns that researchers were going to the field with one hypothesis and then testing and publishing a different one, the AER launched a repository to register the studies' hypotheses in advance. https://www.aeaweb.org/rct.php**

*Sensitive data: ethical issues and data sharing*

- Who should be allowed to access data? Is it global? Do you need ethics training? Does it depend on the nature of consent given by participants? What kind of institution is the applicant in?

  **Croson: The nature of participant consent should certainly determine the scope of the data sharing. I am less concerned with the kind of institution the applicant is in, and more concerned with the use to which the data will be put. Most data collected for research purposes explicitly say that the data will not be used commercially, and this should be a pivotal condition of data access. Ethics training may be required, but I am skeptical that it will be effective.**

- How can participants' rights be respected? If participant permission is required, what should participant permissions include? Are any data so sensitive that no form of access by other investigators is possible?

  **Croson: Yes, there certainly are data so sensitive that no form of access by other investigators is possible (e.g. Census microdata, proprietary data mentioned above, …). However, this should not prevent us from constructing data-sharing principles and policies.**

*Logistics: if you have a plan, who implements it, pays for it, enforces it, etc.?*

- Does your community have repositories (digital libraries) for curating, storing, and serving data? Do you think these repositories should meet community criteria for trustworthiness? Which repositories are trusted? Should repositories be mandated or sanctioned by NSF? If so, what are the criteria for NSF-sanctioned repositories? Who should pay for data curation and storage? How can long-term access be financed, especially for data that are expensive to maintain or manage? Relatedly, how can file formats be kept up to date long-term?

  **Croson: The experimental economics community has a strong norm for data sharing. Some researchers make data available directly on their websites (e.g. http://econlab.ucsd.edu/getdata/), others do so upon request and with some minimal vetting of the requestor. In Economics more broadly, journals host data (the AER is mentioned above), but there is no sanctioned repository. If NSF itself wanted to create and maintain a repository, that might be a solution but the budgetary issues**

**would be challenging. I am reluctant to have the NSF sanction particular repositories.**

- How can NSF ensure that the promises of data management plans are actually carried out?

  **Croson: This is a tricky one. Up until now, the only consequence that the NSF could impose involved withholding subsequent funding. This should certainly be continued, but I could envision a consequence for the institution receiving the grant if the PI does not share as promised.**

- Should NSF allow data to be embargoed, and if so, under what conditions and for how long? How should NSF deal with longitudinal data collection that may require months or years before the original researchers can analyze or publish the data?

  **Croson: Yes, see response above regarding confidential or proprietary data, and appropriate moratoriums. My rule of thumb is that once a paper is accepted for publication using data, the data (and metadata) included that paper should be shared (although it may only be a subset of all the data collected).**

**NSF Data Sharing Workshop: Broad Themes and Questions**

**William Darity Jr., Duke University**

Let me comment at the outset that these are very, very tough questions, and I have much less confidence than usual in answering them. I am looking forward to our exchange and conversation which I view as a real opportunity to move forward on these fronts.

My strongest personal concern is with the issue of sharing of survey, interview, audio, video, and experimental data. These are the types of data typically used by researchers in the social and behavioral sciences. The methods used for their generation should be shared as well. Data sharing is critical for all four of the reasons mentioned in the Background document, but I am especially concerned about reproducibility of the findings and the capacity to conduct new and innovative studies. On a number of occasions I have wanted to make use of data that has been produced by researchers using public funds several years down the road but have been denied access outright or confronted with obstacles and restrictions that have proven discouraging.

As a general principle primary data that was generated via the use of public funds should be made available as soon as feasible for public use, including other researchers who were not engaged in the data collection process. The NSF Grant Policy Manual says data gathered with NSF support should be made available "within a reasonable amount of time." I think we need to explore whether greater specificity should be given to the "reasonable amount of time."

One possible guideline is to make the data publicly available in three years after the completion of the grant period; I think it is important to fix a specific amount of time that is common to all NSF funded data collection. The data should be made available by application for use to NSF – where the data should be required to be stored – rather than by application to the project PIs. NSF should establish a uniform set of conditions that dictate the terms of successful application for use of the publicly funded data after the "patent" period ends. I find problematic the lack of uniformity (and arbitrariness) under existing arrangements. Anyone (university researchers, journalists, commercial researchers, individual citizens) should be able to apply for use of the data as long as they have learned how to and commit to maintaining confidentiality and privacy of the persons whose experiences and attitudes are represented in the data.

As co-PI on a foundation funded survey on race, ethnicity, and wealth that, thus far, has been executed in five cities (Boston, Miami, Los Angeles, Tulsa, and Washington DC), I am very aware of proprietary interests in maintaining exclusive control over data. The administration of the survey in Boston involved a combination of funding from the Ford Foundation and the Federal Reserve of Boston, partially a private and partially a public donor. This raises the question of what threshold of public support (Is it $1?) should dictate whether the data gathered is subject to NSF (or other public sector) regulations for sharing.

We are undertaking a new survey that will be conducted in Baltimore with resources from Ford again and from the Annie E. Casey Foundation. Since both of those are private foundations, I am working on the assumption that my research team and I will have full discretion on if and when our data is made available generally for any interested users, but it is unclear what are the guidelines for projects with mixed public and private support.

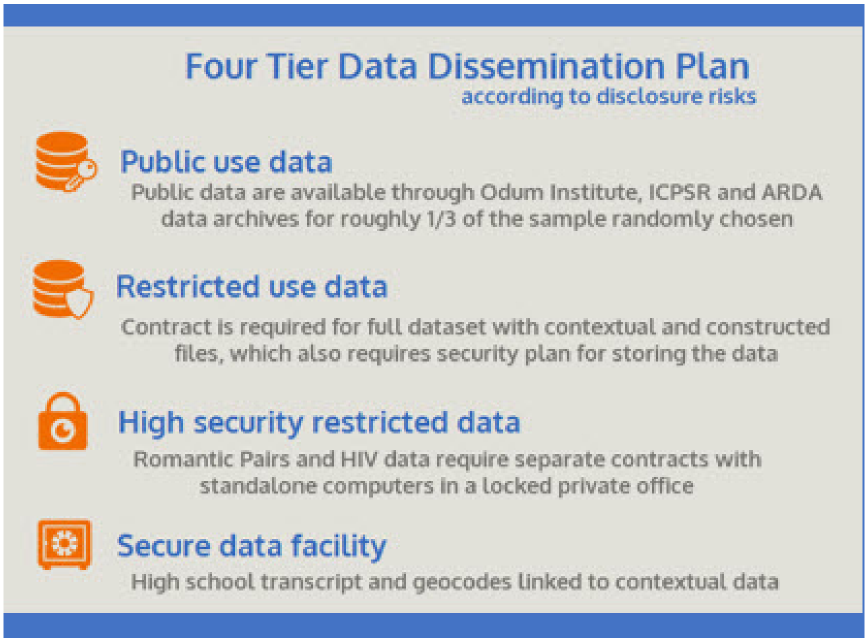
What about synthetic data bases that link NSF funded data with data that was collected with private support? Should the public rules or private rules apply? As a researcher whose most recent projects have been funded largely by private foundations, if linking my data set to one that has been publicly funded will reduce my discretion on releasing my survey information for general use, it creates a disincentive for me to agree to the linkage. On the other hand, if the discretionary rules governing privately funded data are applied to the linked data, it may create an incentive for researchers who have gathered publicly funded data to pursue linkages not motivated by scientific value.

# NSF Data Access Policy Workshop

## January 28 & 29th, Chapel Hill

Add Health
The National Longitudinal Study of Adolescent to Adult Health

**Kathleen Mullan Harris, PI and Director**



### Four Tier Data Dissemination Plan
according to disclosure risks

**Public use data**
Public data are available through Odum Institute, ICPSR and ARDA data archives for roughly 1/3 of the sample randomly chosen

**Restricted use data**
Contract is required for full dataset with contextual and constructed files, which also requires security plan for storing the data

**High security restricted data**
Romantic Pairs and HIV data require separate contracts with standalone computers in a locked private office

**Secure data facility**
High school transcript and geocodes linked to contextual data

## What are "data" for Add Health?

- Cleaned raw source data from Wave I- IV in-home surveys disseminated with contextual data files that have been constructed and maintained by staff at CPC.
- Biomarker and genetic data collected from physical measurements and biological specimens from consenting respondents.

## Why share data?

### Incentives/positives of data sharing

- Add Health was founded on the principles of open and broad sharing of its data beginning in the early 1990s.
- Add Health was a pioneer in the development of security protocols for sharing confidential data to a broad multidisciplinary research community.
- Funders see Add Health as a significant investment whose data should be shared widely within the research community.
- From inception, Add Health policy has no proprietary period for investigators, data becomes available as soon as it is processed and cleaned from the field.
- Design provided unprecedented and unique opportunities for research that no other study allowed.

- Omnibus study with comprehensive coverage of health and health behavior beginning in adolescence and into adulthood.
- Race, ethnic, immigrant, socioeconomic, and geographic diversity on a nationally representative sample.
- Longitudinal and genetic pairs design provide unique methodological solutions to statistical and causation inference issues.

### Disincentives/ potential issues of data sharing

- Deductive disclosure risk (*see sensitive data section below*) limits the ability to more widely share the full sample data without a restricted-use contract, which includes IRB approval and a well-executed data security plan.
- Dissemination is costly and demands full time staff for contract administration. The sensitivity of the data increases this cost and staff effort.
- Metadata needs for be updated for Waves I-IV, codebooks are currently in pdfs. New codebook explorer tool has been developed for in-home data for Waves I-IV. Interactive codebooks and DDI standard metadata needs to be developed to adhere to industry standards for data archiving.
- Data harmonization is particularly important for a longitudinal study, more work is needed to organize data codebooks across waves and incorporate contextual and genetic data files.
- Better dissemination methods are needed for restricted-use data, including the possibility of remote access data enclaves. More funding opportunities and collaboration are needed to develop these kinds of resources for the social science data community.

## Scope of data management policy: What kinds of data does it apply to?

- Policies apply to all Add Health data: survey, geographic, administrative, biomarker and biological samples.
- Genetic Wide Association Study (GWAS) data will be disseminated through NIH's database of Genotypes and Phenotypes (dbGaP) by early summer 2016.
- Add Health will still require a restricted-use data contract for linking to phenotype files.

## Sensitive data: ethical issues and data sharing

- Balancing the tension of sharing public resources funded by the government with continuing to fulfill the pledge of confidentiality to our human subjects.
- The problem of deductive disclosure of an individual respondent's identity was a major concern of Add Health when it was developing its dissemination plan in the early 1990s.
- Deductive disclosure is the discerning of an individual respondent's identity and responses through the use of known characteristics of that individual.
- This is not unique to Add Health—a person who is known to have participated in ANY survey may be identified by a combination of personal characteristics, allowing identification of that person's record. For example, in the Add Health in-school dataset of more than 90,000 cases, a cross-tabulation of five variables can distinguish an individual record.
- Given the large number of people who know someone who, they know, participated in Add Health, researchers who use the Add Health contractual dataset are obligated to protect respondents from deductive disclosure risk by taking extraordinary precautions to protect the data from unauthorized use.

## Logistics

- Add Health has partnered with data repositories for the dissemination of public-use data, which includes only a subset of the sample to limit deductive disclosure risk. Currently, the public-use data is hosted through the UNC Odum Institute, ICPSR, and the Association for Religion Data Archives (ARDA).
- These repositories also offer data discovery tools and online codebooks that we encourage our data users to utilize.
- More funding is needed for the data archiving submission process. Some repositories have fees associated with maintenance and file upload and others are free.
- Standardization of process and fee structure is needed across all research data management repositories.
- To ensure that data management plans are actually carried out, submission of data to repositories, including those like NIH's dbGaP, should be required and funding made contingent on data dissemination. Data submission requirements should still be guided by sensitive data security limitations.

**Philip Kasinitz**
**Presidential Professor of Sociology**
**Graduate Center and Hunter College**
**City University of New York**

*What is data?*

- What does your research community consider to be "data" when it comes to writing a Data Management Plan? (e.g., Does your community distinguish between source data and processed data used for analyses? Are raw source data useable or interpretable? How can data provenance and workflow be characterized adequately?)

*For most users of qualitative, in depth interviewers as well as ethnographic field workers, this can be a difficult question. Obviously in the case of in depth interviews—the narrative is data and should be shared, either in the form of audio recordings or, most commonly in transcribed form. This raises logistical issues of how to properly protect confidentiality in long, detailed in depth interviews in which much identifying information needs to be redacted, as well as how to condense long narrative interviews into a form which other researchers will find useful and accessible. In a large study, simply sharing long, uncoded transcripts of interviews may give the appearance of openness and accessibility but in fact it may not be very useful to later researchers.*

*However I am less clear on whether, when using qualitative coding programs (as is increasingly the case in most large qualitative studies) the coding is "data" –which should be shared--or is it analysis--- which many researchers are reluctant to share? Indeed popular programs such as ATLAS Ti blur the line between data organization and analysis. At their simplest coding schemes are in some, sense indexes, that help us sort through large amounts of narrative data. Yet as these programs have become more sophisticated they have encouraged use the coding process not just to categorize data, but to construct arguments about it. As such researchers may be justified in not wanted to share coded data.*

*Ethnographic field notes make this even more complicated. As "L'affair Goffman" has made clear, the need to protect informants imposes special obligations on ethnographers* <u>*not*</u> *to share some of their material. Further, at what point are field notes "sharable"? Surely some notes—full of short hands and quick memory jogs, are not meant to be shared. I am not sure ethnographers can do their jobs if they think everything they jot down might be seen by later researchers or that they need to do everything in a standardized form that will be accessible to others. On the other hand, sometimes "finished" field notes are important sources for later research, as we often see in anthropology. This is even more complicated in the case of team ethnographies.*

*Why share data?*

- What is the overall goal of data sharing from the perspective of your community? What are the incentives and disincentives for data sharing from the point of view of individual researchers?

*The goal of data sharing, apart from the obvious value of simply having the insights of various researchers with different views and talents look at the same data, is some notion of replicability. This presents qualitative researchers with some problems. Many qualitative interviewers and ethnographers do not share the premise of replicability—that different competent observers will understand and interpret data in the same way. And, again as the Goffman controversy shows, replicability may not be the gold standard in qualitative work.*

- How can data be more widely used? Should data use be restricted to research and educational/informational purposes? Should data sharing also involve use for commercial purposes (both NSF and NIH have commercial entities)? Are there ways to expand usage of data while maintaining the integrity of the research process?

*Call me old fashioned, but I am very nervous about making scientific data available for commercial purposes. Subjects give us their valuable time and sometimes their privacy on the grounds that we are advancing scholarly or scientific understanding. If they think we are gathering data for commercial purposes they may be more reluctant to do so. Or they may fell entitled to a "cut" of whatever profit it generated (and maybe they would be right).*

*Scope of data management policy: What kinds of data does it apply to?*

- How can data management plans better capture the opportunities and challenges of providing access to data to researchers for your areas of research?

  Good question.

- When is a complementary policy on non-digitized data (e.g., biological specimens) applicable/necessary?
- How should data management plans address qualitative (e.g., raw research video, audio files, transcripts) data?

*See above. I would welcome more technical help in archiving qualitative transcripts, redacting identifying information. There really are very few standards of archiving and retrieving unstructured interview transcripts and field notes (perhaps for good reason?).*

*I also feel that too often we act as if transcripts are the only proper format for keeping in depth interview data. In many cases I have found that audio files accompanied by extensive notes, which specify which topics are covered at which part of the file (and perhaps can also flag unusually useful quotes) along with interviewer created summaries of the interviews can actually be more useful than full transcripts (also a lot cheaper and easier to store). Should these also be shared?*

- How should a Data Management Plan address linkages to data not funded by NSF?  Such data may be proprietary.  They may also be governed by other entities and rules.

*Much more can be done in terms of linking data from various sources—particularly geographically coded data from governmental sources (educational data, crime data, locally conducted housing and vacancy surveys, public health data, etc.).*

- Are there new or emerging sources of data or methods in your field that are or likely to be constrained by existing data policies, norms, or practices? What needs to be changed? Do these new sources and methods raise important ethical issues that need to be addressed?

*Sensitive data: ethical issues and data sharing*

- Who should be allowed to access data? Is it global? Do you need ethics training? Does it depend on the nature of consent given by participants? What kind of institution is the applicant in?

*Good questions all. Current ethnics training and IRB requirements seem to me to be more designed to provide protection from litigation than to either  protect participants or facilitate research. I am not sure how we can restrict data to only "professionals" and certainly we can not restrict to only one type of institution. On the other hand subjects will not participate in qualitative studies if they feel that their data can be used in ways they do not think appropriate. I would like to hear what others have to say about this.*

- How can participants' rights be respected? If participant permission is required, what should participant permissions include? Are any data so sensitive that no form of access by other investigators is possible?

*There are! (Again think of the Goffman affair). Some work simply cannot be done if subjects don't feel confidentiality will be guarded. Still I think we over do this some time.  Sometimes the requirement to anonymize data can be create misleading impressions. I also think we need to make exceptions for public  figures and  public events, of the sort  journalists do.*

*Logistics: if you have a plan, who implements it, pays for it, enforces it, etc.?*

- Does your community have repositories (digital libraries) for curating, storing, and serving data? Do you think these repositories should meet community criteria for trustworthiness? Which repositories are trusted? Should repositories be mandated or sanctioned by NSF? If so, what are the criteria for NSF-sanctioned repositories? Who should pay for data curation and storage? How can long-term access be financed, especially for data that are expensive to maintain or manage? Relatedly, how can file formats be kept up to date long-term?

- How can NSF ensure that the promises of data management plans are actually carried out?

- Should NSF allow data to be embargoed, and if so, under what conditions and for how long? How should NSF deal with longitudinal data collection that may require months or years before the original researchers can analyze or publish the data?

*This is a major problem at my institution. We rarely plan for storage of qualitative data after the research period is up, or make plans for changes in technology. On the other hand we usually feel we can't just through it out. So file cabinets and hard drives fill up with stuff no one knows how to properly access or store but we don't feel we can get rid of. Trusted repositories would be a huge help.*

*As for embargoing data—I suspect there is no blanket rule. Different situations will call for different policies. Longitudinal data are a particularly complex case in point. The need to publish before data becomes public may lead researchers to put out material too early in the process—after the first wave of a multi-wave study—that ends up being misleading.*

**Margaret Levenstein**
**Executive Director, Michigan Research Data Center**
**Research Scientist, Survey Research Center**

### *What is data?*

My sense is that, when constructing a data management plan, most people address a fairly finished data product. That is, the management plan is about documenting, preserving, and sharing data that have been cleaned.

A somewhat different, but related point, that I have been trying to integrate into the language that we use to discuss administrative and other naturally occurring or organic "data," is to say that there is a lot of information out there in the world. The process of turning that into social science data, amenable for scientific research, requires documentation that includes provenance, so that we know what it is we are analyzing when we use data for research, and that facilitates sharing, so that we have the possibility of replication.

### *Why share data?*

1. Data sharing is valuable because it permits replication and advancement in scientific knowledge. Analysis of the same data holds a whole lot of things constant, so that we can learn by analyzing exactly the same data in somewhat different ways.
2. Data are a public good, in the sense that one person's use of them does not diminish their usefulness to someone else (like a park or a street). In fact, there may be positive externalities in the use of data, as one person's use may make them more valuable to others, both because there is a common language and basis for comparison and because data can be improved by usage. In this sense, it is simply more efficient to share data. The investment in data creation generates a higher return when they are used by more researchers.
3. There are enormous disincentives to academics, especially younger academics, in sharing data. Data creation is costly, and the rewards to the individuals making those investments are often very small. As is often the case with goods with positive externalities, there is underinvestment. The researchers who bear the costs (in terms of their time, even if a funding agency like NSF has supported it) are often not compensated, in terms of career progression and rewards, for the data products they create. Their rewards are measured in terms of publications from the analysis of the data, not the data itself, so it is in their individual interest to exclude others from access, so that they can publish as much as possible, and distinguish themselves as much as possible from others, based on analysis of the data.
4. I would be extremely cautious about commercial use of data produced using public funds. If the data are simply made public, without restriction, and someone can add value to them in a way that makes others willing to pay for them, that's fine. But I can imagine both researchers and data respondents/generators being concerned about commercial use of data produced by or about them. There is a tension in the "big data" world today created by commercial data producers who do not subscribe to scientific values of transparency, data sharing, and

replicability. We need to model alternatives and public funding should support those alternatives.

### *Scope of data management policy: What kinds of data does it apply to?*

My sense is that the biggest challenge is that proper data management requires resources. Asking researchers to write up a data management plan without additional resources to help researchers properly document and archive data is just a paperwork requirement that will not yield much in the way of valuable data. Most social science researchers do not even know how to manage and preserve data. As journals begin to require data sharing for publication, researchers have an incentive to acquire this knowledge. As graduate schools require data management training, a new generation of researchers is acquiring some of the necessary capabilities. But it is still a large hurdle to overcome, given the incentive structure in the academy and the general lack of institutions for developing researchers' capabilities in data management.

Changing incentives requires working with universities, publishers and journals, and scientific societies. These organizations need to change their requirements and expectations regarding both data sharing and data citations, and their expectations and rewards for data production and sharing. These and related organizations, perhaps including NSF, also need to provide both new and more experienced researchers with both individual capabilities and tools in data preservation. We might even think about working with software producers, such as STATA, to build tools into their products that facilitate data documentation, preservation, and sharing.

- When is a complementary policy on non-digitized data (e.g., biological specimens) applicable/necessary?
- How should data management plans address qualitative (e.g., raw research video, audio files, transcripts) data?

There are a couple of principles here that should apply. Resources go into collecting these qualitative resources (I'm not sure why you call one non-digitized data and the other qualitative data). As with digital, quantitative data, they are only useful to others if they are documented and archived in a way that allows others to make sense of them. This process takes resources. We think about this as a one-time investment for digitized, quantitative data. (Of course it's not, given technological change, but that's a closer approximation.) Once this investment is made, it has the characteristics of a public good in that many others can make use of it without diminishing its value. In the case of these qualitative resources, the costs to documenting them are higher, but more importantly there are much higher storage costs. And one person's use can easily impinge on others. The investment necessary to turn a biological sample into a public good is larger, and therefore makes sense when the data are valuable. Similarly with other archival files, we have to make decisions to discard information when we believe that the storage costs are higher than the potential benefits from future analyses. Part of the challenge is that the ability to quantify and digitize data from these "qualitative resources" is increasing, so one is tempted to hold on to more, knowing that it may be possible to do more with them in the future. But especially for biological samples, I just see lots of freezers sitting around full of undocumented samples, with no funding to support documentation and storage, let alone further analysis, and very little understanding of what the quality of future analyses may be, given our lack of knowledge of the physical rate of deterioration. I'm not saying that we shouldn't encourage, and where cost-effective, archiving these data. But it is not costless, and we shouldn't pretend that it is.

- How should a Data Management Plan address linkages to data not funded by NSF? Such data may be proprietary. They may also be governed by other entities and rules.
- Are there new or emerging sources of data or methods in your field that are or likely to be constrained by existing data policies, norms, or practices? What needs to be changed? Do these new sources and methods raise important ethical issues that need to be addressed?

These are very important issues as social scientists are increasingly using proprietary and commercial data whose use is governed by rules at variance with principles of transparency and replicability. It's very important to develop institutions and practices that establish new norms, cooperatively with data generators and the public more broadly, that facilitate appropriate use of data to protect both individuals represented in the data and the scientific process.

*Sensitive data: ethical issues and data sharing*

- Who should be allowed to access data?

I like the idea of a credentialing process that would reflect shared standards on data protection and ease researcher access, e.g., by a "qualified researcher" card accepted by multiple data custodians and archives that certify that a researcher has completed relevant training in confidentiality protection.

- How can participants' rights be respected? If participant permission is required, what should participant permissions include? Are any data so sensitive that no form of access by other investigators is possible?

Increasingly, we are using "naturally occurring" data rather than survey or experimental data. This is an enormous challenge for "participant's rights" because information about their activities is being used without any intention or consent on their part. We must not constrain this research by requiring consent that would be impossible to obtain, but then we are also obligated to protect the data and be sensitive to its appropriate use.

I am trying to think of cases where a study participant provided information that no access could be provided to any other researchers. I can't. I can imagine saying that there are legal reasons to protect information from non-researchers (e.g., enforcement authorities), and I can imagine a participant saying that their participation was contingent on no sharing until after their death, or the death of relevant people. But again, this seems only relevant when we are talking about identifiable information, more akin to the Irish interviews at BC than anything I would call data.

*Logistics: if you have a plan, who implements it, pays for it, enforces it, etc.?*

- Does your community have repositories (digital libraries) for curating, storing, and serving data?

There are repositories, such as ICPSR.  Many universities have also created their own depositories.  I think we need both standards for these depositories and ways to search across them, network them, etc., or data are lost by being put in a depository that nobody knows about.

- How can NSF ensure that the promises of data management plans are actually carried out?

There are carrots and sticks.  It's easiest to think of sticks, like future funding. But more credible and more effective is funding specifically for data sharing post-project, and funding for the support institutions that create the capabilities and incentives to implement the plans.

- Should NSF allow data to be embargoed, and if so, under what conditions and for how long? How should NSF deal with longitudinal data collection that may require months or years before the original researchers can analyze or publish the data?

I am opposed to embargoes or privileged access, in principle.  In practice, given the lack of incentive to create data that are public goods, some projects may only be feasible if the original researchers are given exclusive access for a period of time. We can think of it as akin to a patent that expires. But you only get the patent, that is, the period of exclusivity, if you make the data public (i.e., you share it in appropriate fashion).

**Brian MacWhinney**

My contribution to this discussion focuses on data-sharing for the study of spoken language interactions. This is the type of data which is being collected, curated, and disseminated through the TalkBank project which I direct. TalkBank (talkbank.org) is a system that includes data-sharing projects for three NIH-sponsored research areas, one NSF-sponsored area, and one NEH-sponsored area. The NIH-sponsored areas are CHILDES for child language data, PhonBank for phonological development, and AphasiaBank for aphasia studies. The NSF- sponsored area is HomeBank for day-long recordings in the home. The NEH-sponsored area is LangBank for data on adult language learning. In addition, we promote data sharing in a variety of non-funded areas such as conversation analysis, classroom discourse, fluency, etc. The unification of these areas under a common technology was supported from 2001-2006 by an NSF Infrastructure grant, but that program was discontinued after the first period.

For spoken language data, government (NSF and NIH) recommendations for data-sharing have not yet been effective in stimulating data-sharing. This is because these guidelines include no clear requirements for data-sharing and no methods for enforcement of requirements. People who decide to contribute their data for data-sharing usually do so not in response to pressure from government requirements, but as a result of a personal commitment to the importance of promoting scientific progress. However, these motivations are not always strong enough to overcome the various additional barriers to data-sharing. This means that, in effect, large amounts of data from both previous project and current work are not being shared. This is a great loss for scientific progress and the public funding of scientific research.

For many researchers, the decision not to share research data stems from concerns about providing other researchers with competitive advantages. Others are worried that their methods for data collection and analysis could be criticized. However, the largest number of researchers who refuse to contribute their data are motivated by a concern that data-sharing would be subject to censure from their IRB committees. In some cases, universities are also blocking data-sharing by invoking their right to protect grant-related IP (intellectual property). Although NSF and NIH mandate data-sharing, they provide no support for dealing with IRB and IP issues. As a result, investigators often decide that the safest course is to not share data. In effect, IRB restrictions are impeding and often completely defeating funding agencies' attempts to encourage data-sharing.

There are several steps that should be taken to correct this situation.

1. NSF (perhaps in collaboration with NIH) should establish methods to help investigators navigate through the complex and often conflicting requirements from IRB policies and university IP protection. This could be done through a special office for the promotion of data-sharing.
2. This office should formulate methods that block the invocation of university IP restrictions for observational data on human behavior.
3. This office should formulate guidelines that protect participants' rights and still permit data-sharing. These policies should focus on providing real protection to human

subjects, and the elimination of overlapping reviews on a procedural level. These guidelines should be promulgated in ways that maximize standardization across highly varying local IRB implementations and standards.

4. On a technical level, maintenance of data anonymity can be furthered through tools and methods for removal of identificatory information from transcripts, audio, and even video. However, the exact level of de-identification should correspond to the requirements requested by the participants, rather than additional requirements invoked from IRB review.
5. These new guidelines should be enforced by making future NIH or NSF funding contingent on evidence of completion of data-sharing, based on a fixed time limit on data-sharing completion.
6. The integrity and sustainability of data-archiving projects should be guaranteed by attainment of the Data Seal of Approval (DSA) by any government-sponsored repository.
7. In areas in which multiple repositories provide similar services, there should be methods for providing sustained access to data and tools when funding or management of a given repository ends. To maximize sustainability, data formats and access processes should be standardized.
8. TalkBank has outlined specific methods for data usage, human subjects training, consent forms, levels of data access, DSA approval, deidentification, etc. and these methods could serve as a guideline for this work.

## Sharing Data, Sharing Models

Steven Manson, U of Minnesota

Any conversation about data sharing must eventually extend to model sharing. Data and models are growing increasingly inseparable in many fields, and it will therefore be necessary to extend data policies to models that create or examine these data. Below are a few general issues along with some advantages and challenges in model sharing. These lists are meant to be more illustrative than exhaustive.

General issues

- The state of model archiving is generally worse than data archiving, although some fields are ahead of others.
- Some research areas, like climate modeling, have a suite of fairly commonly-held models that are shared in the sense of being readily accessible (if not necessarily understandable to someone without advanced degrees in computer science, mathematics, or physics). Others, like integrated assessment, have shorter histories of sharing a fairly small number of well-accepted models.
- Some research areas, like mathematical modeling of microbiology, have standards based on commonly held understanding of a specific research domain, but they are in turn tied to this domain.
- Other areas have de facto ‘sharing’ in that the models are simple or widespread. Regression-based models, for example, can be specified in a straightforward way as a mathematical equation or through recourse to a generic description of a well-understood model (e.g., OLS or logistic regression). That said, there are a growing number of cases where researchers have failed replicate seemingly straightforward analyses, even where both data and model formulation have been freely shared. Causes for this failure vary, but can range from obvious problems such as un-reported variable transformations to more subtle issues such as different statistical packages (or versions thereof) having slight variations on how they implement seemingly standard approaches.

Advantages of model sharing

- The four reasons for data sharing noted by Borgman (2012) all apply to modeling; in short, to reproduce or to verify research, make the results of publicly funded research available to the public; enable others to ask new questions of extant data; and advance the state of research and innovation.
- Some models create terabytes of results, and it may be more efficient or tractable to archive the model and its calibration/initialization conditions than to archive to data. Stochastic models and scenario models in particular may be more useful as test platforms that may be run repeatedly than as one-time generators of data.
- There many reasons why models are useful beyond the standard ones of seeking explanation or prediction, and these additional reasons speak to the importance of sharing models. Among these are that models often structure knowledge, in that understanding how a model is constituted is to gain insight into the patterns and processes at play, and can in turn be useful for education and policy-making.

Challenges

- Underlying model languages and systems change rapidly, ranging from shifts in underlying operating systems (e.g., there are geostatistical packages that run on DOS), language (e.g., the oft-used language Objective C is dying in the face of Apple spurring the use of Swift), and model-specific languages (e.g., there are a dozen agent-based modeling languages and they constantly evolve). There are countless other related challenges in how software is coded, maintained, and run.
- Beyond basic software issues lies the gnarly mess of ontology and knowledge representation more generally. Successful modeling sharing schemes like CellML for cellular biology work in part because the domain is highly specialized and the core concepts are broadly agreed on, unlike many other research fields. We have ways to quantitatively represent abstract notions such as trust and power in models of society but encoding those in a model and then expecting them to transfer to different contexts is extraordinarily difficult. Modeling well known diseases or conditions such as Malaria or Hypervitaminosis relies on capturing a broad array of social and environmental conditions and contexts that are difficult to represent.
- Model sharing has a lot to do with the larger culture. Some research cultures see model sharing as essential to scientific discovery, while others see models in proprietary terms, where data may be shared but models are protected for as long as possible.

Moving forward

- As noted above, there are several fields that offer examples of model archiving. For example, there are heartening efforts in agent-based modeling (ABM) and they are instructive. Marco Janssen (ASU) and others have pushed for polices such as requiring model archiving at openabm.org for any ABM-based paper (e.g., it is a requirement for submissions at *Ecology & Society*), a move that has done much to ensure authors share their models. ABM is also home to a modeling documentation format -- Overview, Design concepts and Details (ODD) (Grimm 2010) -- that is becoming a de facto standard for many journals, often at the request of reviewers seeking better model specification.
- Nonetheless, these ABM efforts also illustrate various needs for advancing model sharing. For example, ODD is still too abstract for many modelers, in that it can fall short in providing enough detail to facilitate model replication. Extensions are regularly proposed by researchers in sub-fields who feel that the generic ODD formulation is not specific enough (e.g., in how it handles decision making or networks). While this confers flexibility and specificity, it means that the single standard is at risk of fracturing into many sub-standards after only a few years after creation.

**Cited**

Grimm V, Berger U, DeAngelis D L, Polhill J G, Giske J and Railsback S F (2010). The ODD protocol: A review and first update. Ecological Modelling 221 (23), 2760-2768.

# NSF Data Management Workshop January 28-29, 2016

Joanna Morris, School of Cognitive Science, Hampshire College

May 7, 2016

## What does your research community consider to be data when it comes to writing a Data Management Plan? (e.g., Does your community distinguish between source data and processed data used for analyses?

The goal of research in the behavioral and cognitive sciences is to understand how the cognitive system pro- cesses information in real time. To do this researchers have used both online and outcome-based measures to assess participants' performance in a wide variety of psychological tasks. These typically involve participants being asked to make a judgment about a visual or auditory stimulus, and then indicating their judgment via a motor or verbal response. Dependent measures include behavioral responses such as response times, error rates, and mouse or eye tracking, as well as physiological responses such as changes in pupil dilation, or the amplitude of brain potentials. Researchers are also interested in understanding the functional architecture of the brain. This latter goal has been primarily pursued via noninvasive neuroimaging, including fMRI and PET.

Data collected in the behavioral and cognitive sciences are highly heterogeneous and complex. Outcome measures such as response times and error rates can result in simple tabular data with a single data point for each trial. However, in studies where verbal or motor responses are recorded for offline coding, data may take more complex forms such as audio and video recordings. Online measures such as eye tracking, mouse tracking and event related potentials can give rise to time-series data with a complex matrix of data points for every trial. Functional brain imaging studies result in a time series of images or maps of signal amplitude divided into voxels (volume elements that are typically a few millimeters across). Each subject may be scanned at two- or three-second intervals for many minutes, yielding a time-series for each of many thousands of voxels. In many cases, the datasets generated by a single study can run to gigabytes or even terabytes of storage [5].

Once the raw data have been acquired, they may be subjected to extensive post-acquisition processing prior to statistical analysis. This can range from discarding of outliers and response times associated with errors in response time data, to the extensive processing required for fMRI data where the raw output of the scanner must first be converted into three-dimensional space and then corrected for head movement, superimposed onto a standard anatomical frame of reference and smoothed (some of the raw activation level of a given voxel is spread to neighboring voxels) to increase the signal-to-noise ratio.

In all cases, once the data have been processed they are subjected to statistical analysis which generally involves comparing the mean value of the dependent variables across different conditions to identify changes that are correlated with changes in the variables of interest. Prior to analysis, data are often aggregated across trials, or across participants, depending upon the type of statistical analysis attempted.

Within the research community, there is little consensus as to which of these multiple types of data should be shared. While some researchers advocate sharing only data that have already been processed on the grounds that without such processing the raw data are uninterpretable [2], other researchers argue that for a database to be scientifically beneficial it must contain the primary (raw) data, that is, all data necessary to interpret, analyze and replicate a study [6].

### Are raw source data useable or interpretable? How can data provenance and workflow be characterized adequately?

There are few technological resources to support the broad sharing of the complex datasets that arise from behavioral studies [5, 4]. For raw data to be interpretable, one must be able to associate the experimental data with the information regarding the experimental conditions, the tasks participants were asked to perform, and the stimuli to which they were exposed during data collection, i.e. with the *metadata* [6]. For metadata, the lack of standardization in data storage formats, specification of experimental designs, and stimulus definitions is a particularly difficult problem to overcome [4].

It may be necessary to create new database formats and analytical tools that will allow data repositories to archive large complex datasets that incorporate both data and metadata, and that will allow users of the database to easily access its contents. We also need to establish standards for the organization of data and metadata, and conventions for file naming, so as to allow researchers to easily submit their data to these repositories.

### What is the overall goal of data sharing from the perspective of your community?

Decisions about what kinds of data should be shared depend upon the goals of sharing the data and the uses to which the data are to be put. One goal of sharing data is to enable meta-analyses, in which the aggregation of data from multiple studies leads to greater statistical power than is possible with any individual study. Another is to make it possible for other researchers to fully explore the data, often in novel ways that were not originally envisaged by the researchers who collected the data. Data mining of this kind may result not only in new findings, but would also allow researchers to investigate the robustness of published conclusions to different analytical methods and statistical significance thresholds [5]. Data sharing has the potential to reduce research misconduct (which is thankfully already quite low) as researchers become aware that reanalysis of raw data may reveal hitherto undetectable anomalies [1]. On a more positive note, data sharing may lead to the development of a database that is more comprehensive than any single laboratory could develop, and allow for the testing of more sophisticated theoretical models [1].

### What are the incentives and disincentives for data sharing from the point of view of individual researchers?

Unfortunately despite the many advantages to be gained from data sharing, it has not always been easy to get researchers to agree to it. The experience of the creators of the fMRIDC is instructive. The fMRIDC aimed to establish a publicly accessible repository of peer-reviewed fMRI studies that contained all data necessary to interpret, analyze, and replicate the published findings. The *Journal of Cognitive Neuroscience* required all authors to deposit their data in fMRIDC and encouraged other journals to adopt the same policy [7].

However the fMRIDC creators found themselves surprised by the negative and hostile response to their efforts [7]. Researchers objected to the perceived loss of competitive advantage, voiced concerns about how to recognize the value of data sharing in terms of promotion and tenure requirements, raised the possibility that the field would become mired in disputes regarding specific analyses in published papers, resented the extra effort or money necessary to convert the data set into the required format, and secretly feared that with dissemination of their raw data, their published analyses may be found to contain errors or fail to replicate when subjected to different analytical methods and statistical significance thresholds [1].

These types of concerns need to be addressed before data sharing will be accepted by the research community. Institutions should provide incentives for data sharing and make the necessary financial support and technical expertise available to researchers. Incentives should include mechanisms via which the data generator receives credit through authorship in publications by data users [3].

## How can long-term access be financed, especially for data that are expensive to maintain or manage? Who should pay for data curation and storage?

Databases are fiscally problematic undertakings for funding agencies. Typical research projects last for a few years and then end, freeing up funds for new projects. In contrast, successful data repositories require ongoing and often continuously increasing funding. Thus, it is not feasible to expect funding agencies to provide long-term support for data repositories. On the other hand, scientists who contribute to databases expect some assurance that their contributions to these repositories and their efforts to convert their data into the appropriate organizational structures and file formats (efforts which can be quite extensive) will not be rendered worthless in a few years' time [2]. It is not immediately clear how scientific databases can achieve financial self-sufficiency. One possibility is to explore subscription-based funding analogous to those that have long been adopted by literature-indexing services and scientific journals [2].

## Relatedly, how can file formats be kept up to date long-term?

In individual labs, data are represented in an enormous range of different file formats, from raw data files in proprietary formats (such as those generated by the popular stimulus presentation program E-prime$^{TM}$) to spreadsheets or word processing documents again in proprietary formats (such as the .xls and doc files generated by Microsoft software), to raw text files (such as those with .txt and .cvs extensions). File formats can become obsolete if file formats are upgraded, if software that supports the format is withdrawn from the market, or if the format falls into disuse or becomes incompatible with current software. In these cases, it may no longer be possible to access the file, read the file or reuse the data. Given these dangers, it would be wise to consider what types of file format will be best for long-term preservation and use formats that are published, open, not protected by patents, and royalty-free.

## References

[1] Beth A. Fischer and Michael J. Zigmond. The Essential Nature of Sharing in Science. *Science and Engineering Ethics*, 16(4):783–799, November 2010.

[2]Peter T. Fox and Jack L. Lancaster. Mapping context and content: the BrainMap model. *Nature Reviews Neuroscience*, 3(4):319–321, April 2002.

[3] Krzysztof J. Gorgolewski, Daniel S. Margulies, and Michael P. Milham. Making Data Sharing Count: A Publication-Based Solution. *Frontiers in Neuroscience*, 7, February 2013.

[4] Maarten Mennes, Bharat B. Biswal, F. Xavier Castellanos, and Michael P. Milham. Making data sharing work: The FCP/INDI experience. *NeuroImage*, 82:683–691, November 2013.

[5] Nature Neuroscience. A debate over fMRI data sharing. *Nature Neuroscience*, 3:845–846, 2000.

[6] John D. Van Horn, Jeffrey S. Grethe, Peter Kostelec, Jeffrey B. Woodward, Javed A. Aslam, Daniela Rus, Daniel Rockmore, and Michael S. Gazzaniga. The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large–scale databasing of neuroimaging studies. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 356(1412):1323–1339, August 2001.

[7] John Darrell Van Horn and Michael S. Gazzaniga. Why share data? Lessons learned from the fMRIDC. *NeuroImage*, 82:677–682, November 2013.

## NSF Data Access/Management Workshop

**Matthew Woollard**
**UK Data Archive, University of Essex (United Kingdom)**

7 January 2016

### Preliminary (personal) note

I'm currently Director of the UK Data Service, which is a distributed data service infrastructure funded by the Economic and Social Research Council. This service has integrated policies for Data Collection and Data Access (and Data Preservation) that all dovetail with the ESRC's own Research Data Policy. I was involved in the construction and revision of the ESRC's Research Data Policies (2009 and 2014) and worked with the ESRC on a unified series of principles and guidance for data policy across all UK Research Councils (2015). I'm a co-author of: *Managing and Sharing Research Data. A Guide to Good Practice* (Sage, 2014), and I'm currently serving on a panel developing a national Concordat on Open Research Data.

### Broad themes/Questions

*What is data?*

> *Research data are defined for the purpose of this document as information relevant to, or of interest to researchers, either as inputs into or outputs from research. They are research materials resulting from primary data collection or generation, or derived from existing sources intended to be analysed in the course of a research project.* (ESRC Research Data Policy)

The various social science research communities, ranging from anthropologists to experimental economists, will consider research data in different ways. Increasingly the data service infrastructures deprecate the term "raw data", because of its difficulty of definition. Primary data may possibly be better terminology.

*Why share data?*

(NB. The term 'share' may not be understood in quite the same way across all stakeholders.)

The OECD principles (p.10), give a non-exhaustive list of reasons; evidence for some of these statements can be quite anecdotal. For example, does data sharing really lead to "The creation of strong value chains of innovation"? However, theoretically these are valid principles.

Corti et al (pp.10-11) list some of the well-known disincentives for data sharing including:

- Loss of first use / Loss of IP
- No resources to prepare to share
- Subject confidentiality

Note these and other reasons are often anecdotal and seldom supported by evidence in practice, though the *Sowing the Seed* report (referenced below) covers a lot of this ground systematically.

The ESRC Research Data Policy adds: “ESRC endorses the RCUK position on the exploitation of research results and positively encourages the exploitation of the results of research it supports, as a contribution to enhance the quality of life, sustainability and competitiveness of the UK.”

Positive as well as negative incentives are required to promote a change in culture and practice and these must be clearly signposted to researchers. Negative incentives, like losing part of one’s grant, or becoming ineligible for future funding *may* have more obvious initial impact than positive incentives, but creators of quality data who happily share also need positive incentives. (But on the other hand negative incentives may lead to resentment!)

Increasing access to data is not quite the same as making the data more widely used, but it is a necessary precursor to it. One method of increasing access to data is to identify and remove irrelevant restrictions to its reuse. Thus, for publicly funded data, restrictions on use should be guided primarily by the content of the data (legal, ethical (and commercial)).

*Scope of data management policy: What kinds of data does it apply to?*

There is no generic answer to this. Different communities will have different answers. Why not start from the principle that all relevant inputs and outputs of research are in scope, unless they are not…

The current version of the Open Research Data Concordat (not public yet) reads:

> *Research Data can be defined as evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical forms). These might be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, interview or other methods, or information derived from existing evidence. Data may be raw or primary (e.g. direct from measurement or collection) or derived from primary data for subsequent analysis or interpretation (e.g. cleaned up or as an extract from a larger data set), or derived from existing sources where the copyright may be externally held. Data may be defined as ‘relational’ or ‘functional’ components of research, thus signalling that their identification and value lies in whether and how researchers use them as evidence for claims.*
>
> *The purpose of open research data is to provide the information necessary to support or validate a research project's observations, findings or outputs.  Data may include, for example, statistics, collections of digital images, sound recordings, transcripts of interviews, survey data and fieldwork observations with appropriate annotations, an interpretation, an artwork, archives, found objects, published texts or a manuscript.*

Thus for most of the questions under this heading there are no “one size fits all” policy statements to be made. If (obviously) depletable data are part of a project then its maintenance and access conditions will need to be formalised. (cf. http://www.ukbiobank.ac.uk/principles-of-access/).

However, access can be allowed/increased while subjects are protected appropriately. The combination of

- good research design (i.e., taking potential sharing into account);
- good archiving procedures (i.e., ensuring anonymisation is robust) and

- selective data access mechanisms (i.e., including a range of access methods which are fit for the content of the data – open on the web to highly controlled within a RDC.)

applied *in harmony* would seem to provide the maximum level of protection at the lowest total cost.

*Are there new or emerging sources of data or methods in your field that are or likely to be constrained by existing data policies, norms, or practices?*

Yes. New and novel (aka Big Data) are likely to require a slightly different approach. Within the social science communities, these tend (but not always) to exist without any informed consent for redistribution/sharing. This can be enlarged on in discussion. Similarly, in the age of big date there will be a change in the emphasis of the traditional relationship between the data producer/data creator and the data repository. Again this can be enlarged on in discussion, but this could have profound effects on the manner in which access to data which is (or is perceived as) sensitive..

*Who should be allowed to access data? Is it global? Do you need ethics training? Does it depend on the nature of consent given by participants? What kind of institution is the applicant in?*

Again, there is no single answer. However, it would seem sensible to start from a principle like: "Everything is available to all" and work out exceptions rather than work the other way around. So, some publicly-funded data should be globally available, some should not. Advanced ethics training may be required in some specific cases, though training in general ethical review standards and legal obligations may be more generally pertinent.

The type of institution a researcher is employed in doesn't necessarily give a clear indication of the type of research which is being carried out. Commercial organisations carry out research in the public good; universities may operate commercial enterprises. Whether the research is in the "public good" may be a more pertinent question to answer. This point could be enlarged in discussion, but opinions of what "public good" is are varied, and establishing "public good" in a transparent and verifiable manner could also be difficult.

Participants' rights must (legally) be protected as well as (ethically) be respected. Some data may be so sensitive that it is impossible to share widely. However, proper planning can almost always obviate the no sharing barrier. So, in a project where personal data is required for analysis (and thus validation) the relevant consent form could be altered to allow for this, e.g., "Information that can identify you individually will not be released to anyone outside the project except for the purposes of independent validation of research, and then only under highly controlled conditions." (I made this up on the spot, so it's probably not watertight, but this is an example.)

*Logistics: if you have a plan, who implements it, pays for it, enforces it, etc.?*

All tricky questions, none of which can be answered categorically. The model that I am most familiar with is a centrally-funded repository for collecting, ingesting, archiving, curating and providing access to data. There are economies of scale in such a model, which allow for standards relating to trustworthiness (ISO 16363, Data Seal of Approval, etc.) to be applied. Any such repository has to have a clear mandate, a business model which provides for sustainability across time and a collections development policy (to define what is in and out of scope). The conundrum of data archiving (as with much archiving) is that demand for reuse over time is difficult to predict [more research needed!], and undertaking a cost-benefit analysis on the indefinite curation of a data collection has a good chance of being wrong. All stakeholders (funders, data creators, data users, etc.) need to be aware of the differential costs of 'short-term' and 'permanent' curation solutions, and make judgements with this knowledge. If the *primary* driver for maintaining access to data is for research integrity reasons (verification, or replication in certain communities), then more expensive long-term curation solutions may not be sensible, etc., etc.

Embargos. Again, there's probably no one size fits all policy. However, it would seem wise not to include the publication of relevant metadata within an embargo policy. Principle 5 of the ESRC's policy states: "This period of privileged use shall not preclude the publication of metadata at the earliest opportunity."

The implementation guidelines for the ESRC Research Data Policy Principle 5 says: "Where a delay in dissemination of deposited data is needed to allow grant holders to publish their research findings, an embargo period can be applied to the data. This embargo period is generally no longer than 12 months from the end of the grant, but may be longer depending on circumstances. The ESRC's data service providers will publish guidelines to ensure transparency."

*Who pays?*

The million dollar question. Theoretically, there are three (or four) stakeholders who can pay.

- Research funders --- they mandate data sharing, so they should carry the cost.
- Researchers (and their institutions) --- they are a primary beneficiary of data sharing, so they should carry the cost.
- Data "re-users" --- they are getting access to data which reduces the costs of their research/development activities, so they should carry the cost.

- Publishers of research --- since they make a 'profit' from research they should carry the costs.

The rationales presented above are highly simplified and represent extreme views which need to be unpacked carefully. In reality, a combination of all three/four stakeholders need to make some form of contribution, but since (and this is even more of simplification) almost all research funding (within the UK, at least) comes, in one form or another, from central government, then the most efficient method of paying would be centrally. In practice however, research funding is much more complex than this, and any model needs to take into account all of those who are benefiting from data sharing, while ensuring that the curation activities are properly resourced (both in terms of level (amount) and predictability (frequency of grant)).

**References:**

L. Corti et al, *Managing and Sharing Research Data. A Guide to Good Practice* (London: Sage, 2014).

ESRC Research Data Policy (2014)
http://www.esrc.ac.uk/files/about-us/policies-and-standards/esrc-research-data-policy/

OECD (2007): OECD Principles and Guidelines for Access to Research Data from Public Funding.
http://www.oecd.org/sti/sci-tech/38500813.pdf

RCUK Common Principles on Data Policy (2015)
http://www.rcuk.ac.uk/research/datapolicy/

RCUK Guidance on best practice in the management of research data (2015):
http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/RCUKCommonPrinciplesonDataPolicy.pdf

RCUK: Draft Concordat on Open Research Data (2015)
http://www.rcuk.ac.uk/research/opendata/

Sowing the seed: Incentives and motivations for sharing research data
http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf

UK Data Service Data Access Policy (2015)
https://www.ukdataservice.ac.uk/media/455247/dataaccesspolicypublic_2_00.pdf