Response to RFI: "Public Access to Digital Data Resulting From Federally Funded Scientific Research" Office of Science and Technology Policy

From: Inter-university Consortium for Political and Social Research (ICPSR), Institute for Social Research, University of Michigan

Contact: George Alter, Director

January 11, 2012

About ICPSR

The Inter-university Consortium for Political and Social Research (ICPSR), a research center in the Institute for Social Research at the University of Michigan, is the world's largest archive of social science data. More than 100,000 users download data from ICPSR every year. Since our creation in 1962, we have expanded to provide quantitative data across all social science disciplines. The Consortium includes more than 700 universities and research organizations located around the world, and we disseminate data for a range of government agencies and other groups, including the Bureau of Justice Statistics, the National Institute on Aging, the Substance Abuse and Mental Health Services Administration, the Bill & Melinda Gates Foundation, and the National Collegiate Athletic Association. Our archive has more than 8000 research collections, some of which include hundreds of datasets. The American Educational Research Association (AERA) and ICPSR are currently working together to encourage broader use of NSF-funded data on education. AERA is offering small grants to young scholars for re-analyzing existing data, and data producers are being assisted in making their data publicly available through ICPSR. The highly regarded ICPSR Summer Program in Quantitative Methods offers more than fifty courses every summer, and almost 900 participants attended in 2011. ICPSR was also one of the founding members of the Data Documentation Initiative (DDI), which has become an international standard for metadata in the social sciences, and we provide the home office for the DDI Alliance.

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

ICPSR advocates Federal policies in these areas to improve the access and preservation of scientific data:

1. Require deposit of all scientific data resulting from funded scientific research in an appropriate repository
2. Long-term funding for specialized domain-specific repositories to distribute and preserve data
3. Consistent citation of data in scientific publications
4. Encouragement of standards for data and metadata
5. Including data re-use as a criterion in evaluating research designs

We explain these recommendations briefly here and include additional detail in our responses to subsequent questions.

1. Require deposit of all scientific data resulting from funded scientific research in an appropriate repository

   A general Federal mandate requiring grantees to archive scientific data for secondary analysis would promote re-use of scientific data, maximize the return on investments in data collection, and prevent the loss of thousands of potentially valuable datasets. We have surveyed NSF and NIH grantees in the social sciences to learn what happened to data created on their projects. One quarter of these grantees reported that the data are now lost, and only 14% archived their data at an established repository (Pienta, Gutmann, and Lyle 2009). Our research also shows that sharing data increases scientific productivity: twice as many scientific publications resulted when data were shared (Pienta, Alter, and Lyle 2010; see also Piwowar 2011; Piwowar et al. 2007).

   In our experience, broader access to scientific data is found in research communities that have developed a culture of data sharing. This occurs when leading scientists share their own data, funding agencies commission datasets for general use, and younger scholars can establish their careers analyzing data produced by others. In contrast, domains that condone secrecy create a culture in which researchers seek a competitive advantage by hoarding data and resist scrutiny of their work. Although researchers in these fields sometimes say that they fear being "scooped" with their own data, we consider such concerns unfounded. In our study of NIH and NSF grants, researchers who shared their data had more publications of their own than those who did not share (Pienta, Alter, and Lyle 2010).

   Precedents for a data archiving requirement are available in both the U.S. and abroad. The National Institute of Justice requires archiving of all data resulting from their funding (see http://www.nij.gov/funding/data-resources-program/applying/data-archiving-strategies.htm). In the United Kingdom, grantees

of the Economic and Social Research Council must offer any data resulting from an award to the UK Data Archive (http://www.esrc.ac.uk/funding-and-guidance/guidance/grant-holders/open-access.aspx). The UK Data Archive operates a self-archiving system providing all ESRC award-holders with a way to meet the data archiving requirement. Deposits in this system are reviewed, and high value datasets may receive additional processing to improve accessibility to the research community.

A data archiving requirement does not imply that all data must be preserved in perpetuity. Repositories can offer a limited preservation commitment, perhaps five to ten years depending upon the scientific domain. During this time, datasets can be selected for long-term preservation based upon use by other researchers and judgments of experts in the field.

2. Long-term funding for specialized domain-specific repositories to distribute and preserve scientific data

We advocate the creation of long-lived, sustainable institutions for archiving, preserving, and disseminating data in each specialized scientific domain. Domain specific repositories are needed to solve both technical challenges related to data preservation and re-use and to champion data sharing within their disciplines. For fifty years, ICPSR has been performing these functions for the social science community, and the relatively high level of sharing and re-use of data in our research community would not be possible without the decades of leadership by ICPSR and our peer institutions in the U.S. and abroad. In addition to ICPSR and our peer institutions in the social sciences (see http://www.data-pass.org/), domain repositories exist in a few other domains (e.g., the Protein Data Bank, Dryad), but wide areas of science lack basic long-term infrastructure. New digital repositories need not be free-standing organizations, like ICPSR. They can also be formed within the framework of existing repositories that have strong, long-term institutional commitments. It is essential, however, for these institutions to have governance structures that make them responsible to the communities that they serve.

Domain repositories are needed to mediate between the specific needs of scientific disciplines and the rapidly developing world of digital preservation. The distribution and preservation of digital assets is a complex and rapidly developing area, and each type of scientific data presents its own problems. The requirements for social science data are very different from those for large-scale experiments in the physical sciences. The development of the Data Documentation Initiative (DDI), an XML standard for social science metadata in wide use around the world, is an example of a domain-specific initiative that was promoted primarily by a coalition of domain

repositories. Specialized repositories can monitor and focus attention on issues relevant to their communities.

Our focus on domain repositories is not meant to exclude other institutions from playing a role in distributing and preserving scientific data. We believe that libraries, archives, and other memory institutions have an important part to play, and a number of universities have created institutional repositories for digital objects. These institutions are in a position to provide general services (such as expertise in data management) and personalized assistance to researchers. The main weakness in the institutional repository model is their lack of experience with data. Most institutional repositories developed out of libraries, and their core competence is in the management of digitized text. We believe a partnership between institutional repositories and domain repositories is needed (Green and Gutmann 2007). For this reason, ICPSR has been actively developing ways to work with institutional repositories under a grant from the Institute for Museum and Library Services (see http://www.icpsr.umich.edu/icpsrweb/IR/).

We are very concerned, however, about the role that scientific journals are playing in distributing data within some disciplines. Some journals have a longstanding practice of accepting data as a supplement to published articles. We see several problems with this model. Journal publishers have neither expertise nor financial incentives to redistribute scientific data in forms that will be most useful to the research community. Data are sometimes published in a very limited format like pdf, which is not intended for extraction of numeric data. Publishers also have no obligation to preserve data to provide long-term access for future researchers. Preservation requires accurate and complete documentation and attention to formats, which become obsolete and inoperable. The enormous volume of data being generated in some fields also raises questions about how long publishers will be willing to pay rapidly rising storage costs to make data available.

3. Consistent citation of data in scientific publications

Scientists who create and share data have a right to expect credit for their efforts. Today, merit for academic advancement is measured by citation counts and "impact factors," and the contributions of scientists who create important datasets should be counted. We strongly believe that datasets should be cited in scholarly publications in the same way that other scholarly products are cited. Unfortunately, citation of data in most scientific publications has been incomplete, inconsistent, and unreliable. With our partners in the Data Preservation Alliance for the Social Sciences (Data-PASS), ICPSR has been urging professional associations to adopt and enforce standards for citing data in their journals. The response of these

associations has been positive, and we note that the American Sociological Review revised their guidelines to authors to require citing data in the reference list of every article. Their new guidelines also require a persistent digital identifier, such as a digital object identifier (DOI), which is an important step in facilitating the capture of these citations by indexing services.

4. Encouragement of standards for data and metadata

   Data access is meaningless without documentation (metadata) describing the contents, context, and origin of each digital object. Standards for data and metadata allow developers to create tools for discovery, access, and analysis of shared digital resources. Standards are especially important for long term digital preservation to assure that data will be accessible and comprehensible ten, twenty, or fifty years from now. As mentioned above, ICPSR has been an active participant in the Data Documentation Initiative Alliance, and we are now beginning to realize the benefits of DDI for facilitating data discovery, providing more detailed documentation, and the standardization of access and analytical tools.

5. Including data reuse as a criterion in evaluating research design.

   Federal agencies that support the collection of scientific data can increase access and availability of data for re-analysis and re-purposing by including re-use as a criterion in evaluating research designs in grant and contract proposals. Scientific review panels should be encouraged to consider whether design features (such as the sample size, representativeness, compatibility with earlier studies for meta-analysis) will affect access to data for secondary analysis. For example, samples drawn from one or two locations are much more difficult to share than national samples, because it is much easier to re-identify subjects when the location is known. It is clearly less expensive to collect data in only one location, but the evaluation of a research proposal should consider potential for future analysis of the data. Public-use datasets are much more likely to be re-analyzed than data only available under a data-use agreement. Consequently, the benefit to cost ratio (e.g., publications per dollar invested) may be much higher for a national sample than for a sample based in a single location. In evaluating the overall scientific value of a proposed project, scientific review committees should consider potential for secondary analysis by future researchers.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

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

Researchers often desire time to complete their own publications before releasing data to others, and a short delay in the public release of data is consistent with an open data policy. ICPSR sometimes defers the release of data for a limited time (usually 6 to 12 months).

Embedding scientific data in publications is not necessary to make data available to other researchers. Datasets in online repositories are assigned unique persistent digital identifiers, which can be cited in publications. As we argued above, repositories are in a much better position to assure access and preservation of data than publishers.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

We agree that scientific data are becoming more diverse. Important differences include:

- Size. Storage requirements have become problems in some disciplines with massive instrument arrays, video, transactional data.
- Confidentiality. Protecting the privacy of subjects is an essential consideration in biomedical, behavioral, and social research. (See National Research Council 2003 and 2005.)

- Obsolescence. Many types of data remain valuable to researchers for a long time, but in some disciplines improvements in instrumentation make data obsolete in a few years.

We believe that a network of domain-specific repositories would be valuable in creating policies to serve the needs of different disciplines. Repositories in constant contact with their communities are in the best position to understand the unique requirements of their disciplines.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

Scientists are not trained in data management, and they often think about the process too narrowly. Few scientists understand the difference between "backup" and "digital preservation," which have very different meanings in the community responsible for digital libraries and repositories. Research communities can benefit greatly from the expertise of librarians and information scientists, and we have noticed the rapid expansion of positions in "data curation" and "data stewardship" in university libraries and research centers. As noted above, we believe that partnerships between organizations with domain-specific and institution-specific mandates are the best way to provide services to diverse and dispersed scientists.

There is a broad need for training in data science to educate stakeholders about the importance of sound data management across the data life cycle and emerging best practices. ICPSR is developing a course on this topic for inclusion in its 2012 Summer Program in Quantitative Methods of Social Research.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

The central problem in funding of digital repositories is that preservation requires a long-term commitment and most Federal funding agencies provide only short-term funding. It is not possible for a repository to assure long-term preservation if funding is provided only in the form of short-term grants.

ICPSR, which is celebrating its 50th anniversary, has developed a sustainable business model based on two sources of funding. First, we have a base of 700 member

institutions that pay for access to data. Second, we distribute data under grants and contracts for twenty different Federal and private funding agencies. Data archived with member dues is only available at member institutions, but access to data supported by external sources is usually open. This model works for ICPSR, because we have a large collection of data that is only available to member institutions, and because we have a diversified portfolio of other funding sources. However, ICPSR cannot provide unlimited open distribution to non-members. If a data distribution agreement ends, the long term preservation of that data is supported by the ICPSR membership, and data access is limited to members.

Two changes in Federal funding models would help to sustain access and preservation of digital data. First, in addition to grants and contracts for data distribution, Federal agencies should be able to pay data archives and institutional repositories for long-term preservation. This could involve a single payment for the estimated present value of future distribution and preservation, which repositories could annuitize in some way.

Second, Federal agencies should make commitments to long-term funding of necessary digital repositories. A number of other countries consider data archiving an essential aspect of their research infrastructure and have made long-term commitments to digital repositories for scientific data. A Federal program to establish and support long-lived institutions is needed to create repositories capable of providing preservation.

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

Federal agencies can assure that data from funded research are accessible and preserved by requiring grantees to report a persistent digital identifier pointing to the data in an established digital repository. Most repositories already assign persistent digital identifiers to objects, and these identifiers can be included in citations. Compliance will be easy and inexpensive to verify, because a persistent digital identifier works like a URL pointing directly to a digital object.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

In our opinion, the most important barrier to broader use of data is lack of standardization in data formats and metadata (documentation). By reducing development costs and broadening the range of compatible data sources, standardization will stimulate innovation.

It is particularly important to develop robust, machine-actionable standards for metadata. We are very concerned that inadequate documentation will result in misinterpretation of important policy-relevant data. Modern surveys involve complex "skip patterns" so that respondents only answer relevant questions. For example, married subjects answer different questions than unmarried people. It is very easy to reach incorrect conclusions if the "universe" of each question is not available. Standards for metadata (such as DDI and SDMX) provide ways for data producers to specify background information that is critical to data users.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

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

Federal agencies should be assigning citations and persistent identifiers to the data they distribute across the federal statistical system. This would ensure proper attribution and credit for data producers and would also help agencies track data reuse to better understand the impact of their funding decisions and data programs. Appropriate attribution language that can be easily inserted into manuscripts should be included with all documentation. This language will make giving appropriate credit easier.

Publication authors should acknowledge original data producers by including citations to the data in the references section of their papers. Treating data citations as first-class references provides attribution and recognition of the importance of data as an intellectual product. Journals and other publishers should require data citation and persistent identifiers as part of their submission criteria.

Standards for Interoperability, Re-Use and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

In the social sciences, many data producers and data archives have converged on the Data Documentation Initiative (DDI) metadata standard – see www.ddialliance.org. Currently expressed in XML, the DDI specification provides a mechanism to document data in a structured, machine-actionable way. This structure enables metadata-driven survey design and processes along the entire life cycle of research data generation.

In the DDI model, metadata needs to be entered only once and then can be referenced and reused later, resulting in greater efficiency. Metadata creation should ideally begin at the conceptualization stage, when survey questions are being designed. Moving this step "upstream" in the data production process leads to greater cost savings for data producers as metadata can be reused.

DDI is being taken up in many countries (see map at http://www.ddialliance.org/community) and by many projects (see a sampling of projects at http://www.ddialliance.org/ddi-at-work/projects), including the National Children's Study and other large-scale efforts.

A Federal commitment to DDI and emerging standards for other types of data would go a long way toward lowering the costs of data management by promoting convergence on these standards and encouraging the development of tools.

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The Statistical Data and Metadata Exchange (SDMX) standard for aggregate time series is another example of a community-driven standard. Eurostat, the European Central Bank, and other partners have developed the standard, also expressed in XML, to share and exchange data.

Making such standards efforts effective and successful requires a defined community of practice whose members are engaged and invested in the outcome. Seed funding can be very important to these efforts. In the case of DDI, the National Science Foundation provided initial funding that supported meetings of the DDI committee developing the specification and beta-testing. This was key to developing momentum. In 2003 the DDI committee reorganized itself as a self-sustaining membership organization to provide modest ongoing funding for standards development.

Federal agencies might consider investing in standards development with the goal of interoperability.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Initiatives like DDI and SDMX are already international in nature. Federal agencies generating data in related disciplines should become members of these efforts in order to have a say in shaping the standards and in coordinating their development internationally. (The Bureau of Labor Statistics is already an associate member of the DDI Alliance.)

The UNECE High-Level Group for Strategic Developments in Business Architecture in Statistics (HLG-BAS) -- http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+Strategic+Developments+in+Business+Architecture+in+Statistics+%28HLG-BAS%29 -- oversees various groups that help to coordinate the development of interoperable metadata across agencies and countries. Federal agencies should encourage and participate in these efforts and the Generic Statistical Business Process Model (GSBPM).

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

Federal agencies should craft policies requiring that their data have citations and persistent digital identifiers and that publications based on them use these citations properly. Disciplines tend to develop their own standards for the elements that belong in a data citation, but in general this is a small set of items. Organizations like ICPSR and DataCite can consult in this area.

Once persistent digital identifiers are part of citations and are integrated into the scholarly publication process, it becomes much easier to automate the harvesting of citations for online indexes and to understand the links between data and publications.

ICPSR has a Bibliography of Data-Related Literature that contains over 60,000 citations to publications based on data in the ICPSR data holdings. This permits two-way linking from the publication to the data and from the data to the publications. Most of the work in associating data and publications for the Bibliography has been manual in nature, but greater use of data citations and unique persistent identifiers should make automated harvesting of this information easier.

Agencies could consider providing such linkages for the data they fund. They currently require acknowledgment through grant numbers in publications. Using data citations could become another such requirement. It would also be welcomed if large publication databases like PubMed would integrate links to the underlying data in their systems.

References

Ann G. Green, Myron P. Gutmann. 2007. "Building partnerships among social science researchers, institution-based repositories and domain specific data archives", *OCLC Systems & Services*, Vol. 23 Iss: 1, pp.35 – 53.

National Research Council. 2003. *Protecting participants and facilitating social and behavioral sciences research*. Washington, D.C.: National Academies Press.

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

Pienta, Amy M., George Alter, and Jared Lyle. 2010. "The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data." Presented at the BRICK, DIME, STRIKE Workshop, The Organisation, Economics, and Policy of Scientific Research, Turin, Italy, April 23-24, 2010 (http://hdl.handle.net/2027.42/78307)

Pienta, Amy, Myron Gutmann, & Jared Lyle. 2009. "Research Data in The Social Sciences: How Much is Being Shared?" Research Conference on Research Integrity, Niagara Falls, NY.

H. A. Piwowar. 2011. "Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data." PLoS ONE 6: e18657.

H. A. Piwowar, R. S. Day and D. B. Fridsma. 2007. "Sharing Detailed Research Data Is Associated with Increased Citation Rate." PLoS ONE 2: e308.