Response to Request for Information: "Public Access to Digital Data Resulting from Federally Funded Research," November 2011
January 12, 2012

Wendy Pradt Lougee
University Librarian
McKnight Presidential Professor
University of Minnesota Libraries

Thank you for the opportunity to comment on “Public Access to Digital Data Resulting from Federally Funded Scientific Research.” These comments are submitted on behalf of the University of Minnesota Libraries. The University of Minnesota is one of the leading public research institutions in the United States, and a key contributor to the entrepreneurial economy of the state of Minnesota, as well as to scholarship both nationally and internationally. We strongly advocate for a policy that ensures public access and long-term preservation (“stewardship”) to digital data resulting from federally funded scientific research (“data”). We believe that such a policy would provide immeasurable public benefits far outweighing any costs or burdens such a policy might impose.

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?

Comment 1:
We recommend a Federal policy that would mandate data deposit into publicly accessible repositories as quickly after publication or shortly after the grant-funding period as possible, recognizing limitations that may be imposed due to confidentiality or other legally protected data. This policy step goes beyond the data sharing requirements outlined in the OMB Circular A-110[1], which many researchers interpret as data sharing only by request, would prevent the loss of potentially valuable data, and remove barriers of access due to variability in data managed in local or individual environments.

Data stewardship policies that encourage public access can also position data for re-use through curation and management techniques that ensure long-term access. Characteristics of a sound Federal data deposit policy might include:

- a requirement for a data management plan with all funding proposals that describes how the data will be deposited for public access within appropriate federal, disciplinary or institutional data repositories.

[1] The Office of Management and Budget (OMB) Circular A-110 was revised in 1999 to provide public access under some circumstances to research data through the Freedom of Information Act (FOIA). http://www.whitehouse.gov/omb/circulars_a110

- a post-award review process and merit considerations for future funding based on a successful history of data deposit. The NSF's Computer & Information Sciences and Engineering (CISE) directorate[2] provides a good example.
- recognition that not all data can be open due to security or confidentiality interests, and that some categories of data will need to be appropriately prepared before release or qualify for an exemption. When open access to data must be delayed, deposit with access restrictions could still be required in order to ensure preservation and long-term access.

(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?

Comment 2:
Intellectual property issues associated with data differ significantly from those associated with publications. Established federal law does not recognize an ownership interest in raw data, reflecting an understanding that data may also be most productive and fruitful throughout the economy when access and use are available to many parties. Intellectual property rights may exist in certain types of data, or compilations of data, but these are well addressed by current law. Federal policies should:

- value and reward data stewardship for re-use -- e.g., by incorporating data stewardship as a consideration at reviews and in applications for future grant awards.
- prohibit constraints on data due to overly tight coupling with related publications, or by publisher-imposed limitations on data access. Requiring open licensing for the data that is covered by intellectual property laws will foster productive reuse while allowing credit to be given to those who did the work.

(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?

Comment 3:
Recognition of the inherent differences between disciplines is critical, particularly with respect to decisions on archiving, shareability, and discovery. The NSF DMP requirement offers a good model by setting a minimum policy, and yet allowing each directorate to build from the base. The Federal agencies could take this model one step further by outlining a set of data stewardship principles and requiring that each agency provide base-line principles for data management, preservation, and sharing of data in their respective disciplines, such as establishing a disciplinary metadata standard, a shared data repository, appropriate maximum embargo periods, etc. Researchers could then model their data stewardship on these standard requirements.

[2] The NSF CISE Directorate Guidance for CISE Proposals and Awards requires that Data sharing progress and outcomes must be reported in award annual reports, the final report and subsequent proposals by the PI and Co-PIs. Last updated September 15, 2011 at http://www.nsf.gov/cise/cise_dmp.jsp.

Since disciplinary repositories are developed to take into account the unique structures and characteristics of the target data, Federal agencies should support protocols to enable cross-discipline and cross-repository discovery. The University of Minnesota’s NSF-funded TerraPop project, a joint undertaking of the Minnesota Population Center and the Institute for the Environment, represents a good example of integration of cross-disciplinary interests (in this case demographic and land-use data over time).

(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?

Comment 4:
In addition to differences between disciplines and associated data, there are differences associated with the data themselves that relate to cost-benefits of management and archiving – e.g., can the data be reproduced, do the data serve a canonical reference value? Consequently, minimum policies about data management, sharing, and preservation may not adequately ensure an overall cost-effective and sustainable data environment. Federal agencies and associated policies for federally funded research could:

- set data retention guidelines based on replicability, importance and potential use
- support establishment of discipline-based repositories with an ongoing deposit mandate to ensure long-term access to data at a scalable cost. See the NSB report[3] on long-lived digital data for examples.
- address the fact that existing disciplinary repositories do not always accept the full range of data generated by researchers in their field. For example, CUAHSI's Hydrologic Information System[4] only accepts geo-referenced data, yet researchers may do lab-based hydrologic research and produce relevant data. Policies should encourage existing data repositories that receive federal funds to accept and curate a broader range of data in their disciplines, and agencies should fund development of new repositories that bridge disciplinary gaps.

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

Comment 5:
Data Management Plans (DMPs) have had a positive impact on highlighting the issues of stewardship of digital data, however the benefits and strategies involved still remain limited within many research communities. Requiring a DMP for funding applications is a good start, however more specific guidance (see comment 3) and stronger incentives to openly share (see comment 1) would contribute to the implementation. All stakeholders should promote policy compliance along with community-based standards making, specifically:

- Research communities can

[3] NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century (2005) http://www.nsf.gov/pubs/2005/nsb0540/

[4] Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) http://his.cuahsi.org/

    - collaborate on the development and operation of domain-based infrastructure for data stewardship.
    - recognize researcher efforts to preserve and make data accessible and usable when considering rewards for professional achievement.
    - create tools for sharing and preserving data and new methods, standards, and tools for metadata capture.
- Universities and research organizations can
    - clarify organizational/institutional intellectual property policies regarding data.
    - promote data publication and sharing as part of the tenure process considerations.
    - recognize and support data management systems as contributing to enterprise-level research infrastructure.
    - require DMPs and data sharing implementation for all doctoral dissertations.
    - offer career development paths for data managers and data scientists on campus.
    - include data training as part of the science curriculum. “Open data access with appropriate ethical restrictions can be viewed as a new core principle for developing a global data infrastructure” (Parsons, 2011)[5]
- Libraries and librarians can
    - curate data as part of the scholarly communication cycle
    - collaborate in development of campus infrastructure for discovery, management, and distribution of data. Provide educational and consulting services about data plans, data management, and access to data repositories.
- Publishers can
    - enable and encourage data attribution through citations within publications and research documentation.
    - eliminate barriers to data access through pay-wall or copyright restrictions to “data supplements” in research articles.

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

Comment 6:
In many ways the real costs of preserving and making data accessible are just beginning to emerge, with growing recognition that human capital and expertise are often as essential to maximizing the value of data as technical infrastructure. Several areas for agency action include:

- funding research and education on the costs and benefits of data stewardship.
- supporting development of community standards and cost-effective infrastructure such as tools for deposit, management and discovery.
- encouraging greater specificity in data management plans to address sustainability.

[5] Parsons, Mark. (2011). Expert Report on Data Policy and Open Access. http://www.grdi2020.eu/Pages/SelectedDocument.aspx?id_documento=e31a1aab-b01e-4e7e-9b10-0fd93d4b710f

(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?

Comment 7:
Standardizing data stewardship policies across granting agencies and providing models of best practices would simplify and streamline both compliance and verification. Agencies have leverage at the time of submission and completion of grants, and mechanisms to track compliance could occur at both points. Strategies might include:

- tying new grant awards to prior data stewardship (for example, the NSF Computer & Information Science and Engineering Directorate Guidance for Proposals and Awards requires that data sharing progress and outcomes must be reported in award annual reports, the final report and subsequent proposals by the PI and Co-PIs).[6]
- enabling better tracking by implementing a data ID registry that is linked to stewardship best practices. For example DataCite is issuing persistent data identifiers for data, however, this is only one component of a verification system. Issuing a persistent and exclusive data ID, that is only given to data in repositories that meet minimum standards for openness and preservation, would provide clear evidence of data stewardship. Models exist, such as TRAC, for trustworthy repository standards that could be adopted.[7]

(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?

Comment 8:
Federal agencies can stimulate use of data by supporting the development of public access data archives to enable discovery and download. Public APIs could allow individuals who may not have specialized data analysis software to access and make use of the data, and could automate collection and aggregation for new capabilities and services. Specific agency actions might include:

- providing funds to discipline repositories and data producers to create additional metadata, visualization tools, and augmented holding records to allow users in unrelated fields and without the necessary software or expertise better access and to encourage reuse. For example, the Minnesota Geological Survey proposes to augment their public data collection including surficial geology data that is of interest to citizens and scientists alike who do not have access to the specialized and expensive GIS software to read the professional data. The aforementioned TerraPop project will similarly make data and general/specialized tools available to diverse communities.

[6] CISE Proposal and Award Guidance. Last updated September 15, 2011 at http://www.nsf.gov/cise/cise_dmp.jsp.

[7] Trustworthy repositories audit & certification (TRAC) criteria and checklist. http://catalog.crl.edu/record=b2212602~S1

- establishing awards for web-based visualization tools of existing data –e.g., the successful *Digging into Data Challenge*[8] that NSF has partially sponsored over the past three years. Also, Data.gov[9] has transformed into a “cloud-based Open Data platform for citizens, developers and government agencies” through web-based apps and clear re-use policies.
- promoting standardization of metadata using ontologies or RDF to incorporate data into new domains or use for non-traditional research (i.e. visual arts).
- provide awards for innovative re-use of research data. Alternatively, provide funding to repositories to host a data challenge with monetary incentives,
- stimulating new research in semantic technology, the underlying technology that supports linked open data.
- enabling citizen science efforts which have engaged the public in data gathering or data classification activities –e.g., Galaxy Zoo[10], is a Citizen Science Alliance project that has engaged over 250,000 members of the public in cloud-based scientific discovery.

(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?

Comment 9:
Strong attribution norms already exist within the academic and professional communities with regard to research publication, but are not as well-developed with respect to data. Policies embracing certain forms of open licensing (such as Creative Commons licenses) might strengthen attribution practices. Federal policies could contribute to developing strong data attribution practices by:

- encouraging development of data citation standards and practices. This would improve capabilities for tracking re-use and also provide impact measure of individual researchers’ contributions.
- requiring data citation following the above standards in publications resulting from federally funded research (as opposed to simply mentioning data sources in the bodies of articles.)

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.

Comment 10:
One of the biggest obstacles to data discovery, access and reuse is the lack of standardization of descriptive metadata adoption and practice within a specific scientific domain. This constrains the ability to search effectively across federated data repositories, as well as to ensure search effectiveness within a central index to data assets. Community-driven data standards efforts are essential within a domain context, while more general standards are necessary to promote interoperability nationally and internationally, and at more general levels of discovery (i.e.,

[8] The Digging into Data Challenge is an international collaboration administered by the Office of Digital Humanities at the National Endowment for the Humanities. http://www.diggingintodata.org/

[9] Data.gov - Open Federal Data. http://explore.data.gov

[10] Galaxy Zoo Project. http://www.galaxyzoo.org/

cross-domain searching). The catalog of community-driven domain-based data standards are too numerous to list out here, but prominent examples include standards produced by the Federal Geographic Data Committee (FGDC) for digital geospatial metadata, and Darwin Core, which functions as an extension of Dublin Core for biodiversity informatics applications.

Community-based expertise should be used to develop standards and conventions for data structure and metadata management specific to a discipline's research output. However, certain areas of metadata are of particular cross-disciplinary interest. Measures and representations of time and space are important cross-indexing aspects of many datasets. Researchers and disciplines should be encouraged to ensure inclusion and interoperability on these measures through use of standard models for recording these data. Repositories should also be encouraged to explore further the use of semantic web technologies (RDF and URL-identified entity and relationship vocabularies) and linked data to leverage discovery.

Recognition of a scientific data standards registry, engagement of national and international standards organizations and inter-agency participation in the creation and endorsement of standards are ways to encourage requisite coordination of standards.

In addition to the attention on standardized descriptive metadata, *technical metadata* is of increasing importance and utility. Technical metadata can provide information concerning the instrumentation and methodology used to create the data, and may be critical to ensure validity, reproducibility, and re-use by users not involved in generating the original data set. This should be linked to the data as part of stewardship. Currently some disciplines separate this out and describe data creation methods in a research article, but this information must be linked or kept together in an open data environment for the data to be usable and trustworthy.

(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?

Comment 11:
There are examples within various domains that have been successful in producing data standards for interoperability and re-use –e.g., the astronomy community’s development of the National Virtual Observatory (NVO) and the associated International alliance. The data standard such as FITS allows for data from different instruments to embed metadata elements that can be read by open source software. The success of this relies on ease (if not automation) of metadata creation, as the instrument or software embed metadata, such as celestial coordinates, into the image. This is similar to the GIS communities’ standard of FGDC, but is also software dependent.

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

Comment 12:
Federal agencies should support community-based efforts that promote and coordinate standards across international communities -- e.g., aligning policies with the processes supported by such bodies as the Committee on Data for Science and Technology (CODATA). Further, support for repository “nodes” of data-sharing infrastructure, such as the NSF DataNet

project DataOne, will bring together disciplinary communities that might not have the capacity to or budgetary incentives to combine efforts.

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

Comment 13:
Linking publications with the underlying or associated data provides significant benefit for research. Dryad is a good example of a discipline repository linking the underlying data supporting journal publications in the biosciences, with plans[11] to integrate data deposit for society and for-profit journals in the field within this publicly accessible data archive. Requirements for unique and persistent data identifiers will aid in this linkage and enable tracking of re-use. Federal agencies should support the work of the International DOI Foundation (IDF), DataCite, and other organizations with the goal of making a unified international standard and support structure linking between publications and associated data.

[11] Dryad Wiki (update January 4, 2012) Submission Integration. http://wiki.datadryad.org/Submission_Integration