Stephanie Wright
swright@uw.edu
Data Services Coordinator / Atmospheric Sciences Librarian
University of Washington Libraries
Seattle, Washington

January 12, 2012

Office of Science and Technology Policy
The White House

RE: Comments in response to Office of Science and Technology Policy Request for Information: Public Access to Digital Data Resulting From Federally Funded Research
Federal Register Doc No 2011-28621
http://www.gpo.gov/fdsys/pkg/FR-2011-11-04/html/2011-28621.htm

Thank you for the opportunity to comment on this issue of such great importance to the future of scientific research in this country. I have provided my responses to those questions below to which I felt I had the most relevant expertise.

(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?

At the most basic level there needs to be a federal policy that mandates federally funded research data is deposited in a manner that makes the data openly accessible and fully reusable, with exceptions to this rule being applicable only in instances where access to the data poses privacy or security concerns. Exceptions should be based on ethical, not proprietary, considerations. "An open data regime not only maximizes the benefit of the data, it also simplifies most of the other issues around effective research data stewardship and infrastructure development."[1] There is already research that shows open access to research data increases citations and reuse of data.[2]

While immediate access by the public to this data would be ideal, an embargo period would not be out of the question to allow the original researcher/s to capitalize on the publication opportunities. That being said the allowable embargo period should not be so long as to unnecessarily restrict access for an extended period of time and slow down advancements that can be made through the reuse of that data. This will maximize the return on federal investment in the original research.

The federal policy should also include a clearly defined provision for free and open re-use of the data either by mandating that the data be in the public domain or at most maintaining that

[1] Parsons, Mark. (2011). Expert Report on Data Policy and Open Access. GRDI2020. http://www.grdi2020.eu/Repository/FileScaricati/e31a1aab-b01e-4e7e-9b10-0fd93d4b710f.pdf Accessed 12 January 2012.
[2] Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

subsequent users of the data must provide attribution along the lines of the requirements for the Creative Commons CC:BY license.[3]

Free and open access to research data provides new opportunities for commercial development of not only one's own intellectual property but also that of others. It opens up opportunities for everyone and accelerates scientific and commercial innovation.

(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?

Permitting free and open use of research data through a license similar to the CC:BY license, will allow credit to go to the researchers who invest significant effort into the research and collection of the data while also allowing for any subsequent researchers to be clear on the rights surrounding its reuse. This will also minimize the reluctance to reuse data by alleviating fears of lawsuits as current copyright law is poorly understood by most researchers. Even the current copyright law stating that facts are not copyrightable leads to misunderstanding surrounding its availability for reuse. Clear licensing of the data in this manner will minimize the complexities in instances where data is being used by researchers in international collaborations from different countries with significantly different or conflicting copyright law.

There would need to be development on a standard for data citation, preferably one that allows for tracking of citations between publications and their related datasets. This standard for citation would also need to take into account reusability, merging of datasets and versioning so credit can be given and shared appropriately.

(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?

The National Digital Information Infrastructure and Preservation Program (NDIIPP) 2010 Report proposed that legislation be enacted that would permit the Library of Congress "to designate specially qualified institutions as agents for mandatory deposit for specific types of content."[4] By taking this further and making the qualified repositories domain specific, it can simplify identification of the appropriate repository for a particular dataset and improve the findability of relevant datasets, much as subject classification does for print materials. It would also minimize the number of different types of data any one repository would need to accommodate and maintain and can build on existing infrastructure. This could lead to enhanced collaboration between stakeholder communities and increased likelihood of standard creation relevant to the domain.

---

[3] http://creativecommons.org/licenses/by/3.0/

[4] National Digital Information Infrastructure and Preservation Program. (2010). Preserving Our Digital Heritage: The National Digital Information Infrastructure and Preservation Program 2010 Report. Retrieved from website: http://www.digitalpreservation.gov/multimedia/documents/NDIIPP2010Report_Post.pdf. Accessed 12 January 2012

When it comes to creation of a federal policy, allow funding agencies to develop relevant guidelines for researchers in different domains such as NSF's data management plan guidelines which vary by directorate.

(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?

Grant funding to researchers and their institutions will need to be increased to help allay the costs of maintaining the qualified repositories mentioned in Question 3 with different amounts of funding based on the data type requirements of those domain specific repositories. This will help maintain the sustainability of these repositories by sharing the burden between multiple institutions, the organizations that support the domain as well as the federal government, all of whom would benefit from access to the data in the repository.

For domains where there are not already established repositories, the funding agencies can provide startup money to libraries and their institutions to develop a relevant archive with the stakeholders in that domain (research institutions, publishers, other domain-related organizations, related industries) developing a plan to cover long-term costs of the repository.

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

Shared responsibility among all the stakeholders is the most effective way to achieve the goal of maximizing the investment into research and its outputs. All the stakeholders can work together to develop, promote and support a domain-based infrastructure. They can collaborate on developing standards and tools for metadata capture and shifting the current data-silo culture to a one of data sharing.

Publishers and research organizations can enable and encourage ethical data sharing by providing recognition to researchers through data citation and cross linking. Universities can increase the incentive to share and reuse data by tying both of those to the tenure and promotion process. They can also make sure that data management is applicable earlier in the career of the researcher by making data management plans required for dissertations and theses with related research data. Libraries and universities can educate researchers on the importance and value of proper data management. The government can encourage this by providing funding for creation of data management curriculum.

(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?

Along with those suggestions provided in Question 2, there could be development of author and institutional identifiers as is being attempted through the ORCID (Open Researcher & Contributor ID) organization.[5] Buy developing a method of citation that incorporates data

[5] http://orcid.org/

creator identifiers, recognition could more easily be given to those who not only did the original work but subsequent researchers who add value to the original data as well.

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

In conjunction with suggestions made in Question 2 and Question 9, it would be in the best interest of the researchers, their institutions and publishers to work together to create a standard method of tracking and linking citations of datasets with the related works published in their publications. They could work with an existing organization such as DataCite to extend the functionality of their EZID permanent identifier service to meet that need.[6] By allowing cross linking of datasets and their related publications, not only is the dataset given increased visibility but so is the publication derived from the data.

[6] http://datacite.org/cdl_launch_ezid