This is a response to the Office of Science and Technology Policy's "Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research." Note that our responses cover questions 1-9. We do not respond to questions 10-13.

Request for Information:
http://www.federalregister.gov/articles/2011/11/04/2011-28621/request-for-information-public-access-to-digital-data-resulting-from-federally-funded-scientific

Responses written by:
Matthew Mayernik – mayernik@ucar.edu
Mary Marlino – marlino@ucar.edu
Karon Kelly – kkelly@ucar.edu

Affiliation:
NCAR Library
National Center for Atmospheric Research (NCAR) / University Corporation for Atmospheric Research (UCAR)
Boulder, CO

(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?

RESPONSE:

Federal policies can strongly influence how data resulting from federally funded scientific research are managed and preserved. Such policies should focus on creating institutional structures within and across disciplines such that researchers organize their research practices around data sharing and re-use.

Any new policies must recognize that providing access to and preserving digital data is a profoundly human process. Technologies facilitate the collection or creation of digital data, as well as the discovery, transmission, and preservation of data across space and time. But digital data can only be collected, accessed, and preserved through the purposive actions of individuals and organizations across the public and private sectors.

(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?

RESPONSE:

Intellectual property interests around digital data should focus on the creation of social norms within particular communities, not new legal protections. Facts, such as "the temperature in Boulder, CO, was 62 F on Jan. 10, 2012," are not copyrightable, thus most

forms of scientific data do not fall under copyright control. "Data organization," on the other hand, can be put under copyright licenses (Stodden, 2009). Free/Open copyright licenses, such as the GNU General Free Documentation License (GFDL) and the Creative Commons licenses, can thus be applied to data organization systems. Applying such licenses to data, however, will inevitably complicate data sharing and integration efforts of any large scale for a couple of reasons. First, there is an amorphous line between what can and cannot be copyrighted within databases. Within a database, what is an uncopyrightable fact and what is a copyrightable arrangement of facts? Second, different copyright licenses have different usage requirements. The implication of this is that querying across ten databases may return results with ten different usage licenses. The data user is then put in the difficult position of navigating complex legal regimes before bringing data together and releasing subsequent results.

Because of these difficulties in navigating intellectual property issues around data, the Science Commons project, an off-shoot of the Creative Commons organization, recommends that scientific data be assigned to the public domain rather than being placed under copyright of any form (Wilbanks, 2008). Their "Protocol for Implementing Open Access Data" (http://sciencecommons.org/projects/publishing/open-access-data-protocol) outlines how the public domain is the most appropriate way to enable the widest use of scientific data. Putting data in the public domain eliminates data use restrictions, it enables data integration in that data from disparate projects all have the same legal status, and it encourages non-legal means for resolving problems related to data use. In lieu of copyright-based methods of controlling data use, federal policies should promote norms within scientific communities as to how data should be made available, used, and attributed. Scientists should check with data centers for data use and attribution policies, and work with collaborators to ensure that the usage and attribution of others' data meets with community accepted practices. For example, the 2007-2008 International Polar Year (IPY) project required that data be "made available fully, freely, openly, and on the shortest feasible timescale…equitable, non-discriminatory access to all data preferably free of cost, but some reasonable cost-recovery is acceptable" (IPY, 2008, pg. 3). Similarly, the seismology community has developed a norm in which data are released to the broader community after a specified period of time. This norm is codified within the NSF Division of Earth Sciences policy: "For those programs in which selected principle investigators have initial periods of exclusive data use, data should be made openly available as soon as possible, but no later than two (2) years after the data were collected." (NSF Division of Earth Sciences, 2010, pg. 2).

Not all data can be assigned to the public domain. Data collected about individuals, medical data, classified data, and other sensitive data (such as the locations of endangered species), are, and should be, withheld from subsequent use unless measures have been taken to ensure their compliance with ethical and legal considerations, such as anonymization, declassification, or removing sensitive data by other means.

(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?

RESPONSE:

Disciplines do exhibit differences in their data management practices. For example, some disciplines have higher levels of adoption of data and metadata standards than others. However, recent studies have shown that the variation in data management practices is often as large within individual disciplines as it is between disciplines. This intra-disciplinary variation in data management practices can be seen in astronomy (Wynholds, et al, 2011), ecology (Mayernik, Batcheller, & Borgman, 2011), and the quantitative social sciences (Pienta, Alter, & Lyle, 2010), among others.

When looking at data management practices from an institutional perspective, however, it is possible to see that many important data management challenges span the academic disciplines. Many questions important to data management and preservation are discipline agnostic: What data management institutions exist (or do not exist) for particular disciplines? How well are they known by researchers within those disciplines? Do institutions exist that create and promote data format, transmission, and preservation standards? Do data centers/repositories/archives exist? Does the discipline have a tradition of working with trained data management experts within library and/or computing institutions? Is data management/sharing valued by the institutional structures that reward achievements within a discipline, such as graduate student advancement, and faculty tenure and promotion decisions?

(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?

Please see responses to question #3, #6, and #8.

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

RESPONSE:

Data management plans are the first step in the data management process. *Plans help collaborators orient each other to their data management needs and options, but only lead to effective data management down the road if data are actively managed as an ongoing process.*

The stakeholders listed in this question can engage with researchers from the beginning of the research planning process to promote and facilitate data management. Many university and research libraries are now offering data management planning services, in which they work with researchers to develop a data management plan for a research proposal. These planning services build relationships between researchers who create or collect data and the library and university institutions that have the expertise and (ideally) the capacity to ensure that data are made available and preserved over time.

As an enforcement mechanism, funding agencies and universities can reduce/withdraw funding from Principal Investigators if data management plans are not carried out. Similarly, universities and funders can deny funding future projects based on a Principal Investigator's insufficient data management actions in the past. An incentive-based approach to promoting data management would reward researchers for reusing and repurposing existing data collections, thereby increasing the demand for quality data collections.

Publishers can build relationships with data archives to build pipelines that enable researchers to deposit data with data archives as a part of the publication process. Publishers can also request that researchers provide citations to the data that were used to produce a publication.

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

RESPONSE:

Funding agencies should work with existing data archives to understand different cost and sustainability models for data management and preservation work. Numerous data archives exist in many disciplines. These organizations understand the costs involved in collecting data, organizing and providing access to data, maintaining data over time, and preserving digital resources. How much does it cost for them to do their work? Understanding current data archiving practices would greatly inform future funding for long-term preservation and access of data.

Another way to assess data management and preservation costs would be to fund select (but diverse) pilot projects where the economics of data preservation and accessibility are explicitly studied. For example, the NSF could explicitly study the costs of data management and preservation within the forthcoming National Ecological Observatory Network (NEON, http://www.neoninc.org/), or the Advanced Cooperative Arctic Data and Information Service (ACADIS, http://www.aoncadis.org/home.htm;jsessionid=B7A0C3ABCA80E00C83D4A316D76DE570), which is being managed by NCAR and the National Snow and Ice Data Center, both in Boulder, CO.

Critically, any attempt to assess data management, preservation, and long-term access must take a long-term view. The costs of data preservation and access cannot be quantified by looking at a two-to-three year window. The largest costs of data management and preservation are ultimately related to the long-term (and often open-ended) commitment required to ensure that data resources will continue to be available into the future.

(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?

RESPONSE:

Representatives from funding agencies should promote compliance by knowing the institutional landscape for data stewardship. What data centers are relevant for a particular project? Should a funded project be working with a particular data center, or using a particular data standard? Individual investigators may not have a wide enough view to know where their data might be submitted for long term preservation. Funding agencies can create relationships that may not exist yet between individual investigators and data centers by making introductions and providing financial support for researchers to prepare and submit their data to relevant data centers.

(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?

RESPONSE:

Within the research communities, agencies can develop funding programs that explicitly invite/require researchers to make use of existing data. That is, agencies can release calls for proposals wherein money is earmarked specifically for proposals that will leverage existing data. Currently, it is much easier for researchers in any discipline to receive funding for projects that produce new data. To stimulate innovate use of existing data, data reuse must be financially supported in the same way as original data production.

Second, agencies can develop and support education programs across the disciplinary spectrum that promote "data science" as a viable career path. This can include graduate and post-doctoral fellowships for data-related research and development in ecology, sociology, atmospheric science, library science, biology, etc, as well as educational initiatives that bring researchers from different disciplines together in order to foster collaboration and cross-discipline sharing of knowledge, technologies, and research opportunities. Funding for the development of educational programs that cross the information and scientific disciplines could serve to introduce data management techniques and practices into disciplinary curricula. Data management and curation workshops for undergraduate and graduate students might also bring more such activities within disciplines where those topics are not regularly addressed.

(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?

RESPONSE:

Within NCAR, we are promoting data citations as a way to assure that data producers are given appropriate attribution and credit for the use of their data for secondary and integrative purposes. Our data citation initiative is part of a broader movement in scientific and public policy circles. The interest in data citations is coming from many research stakeholders, including funders, policy makers, professional societies, research organizations,

and individual researchers (NAS, 2011; AGU, 2009; ESIP, 2011; Parsons, Duerr, & Minster, 2010), and is stimulated by the availability of new tools for identifying and linking to data in a web environment (Van de Sompel, et al., 2004; Bizer, 2009).

Data citations promote transparency in research by offering a direct pathway to the data so that the research can be validated or easily carried forward from a known starting point. They also raise the profile of data, that is, to make data as valued and rewarded in scientific settings as peer-reviewed publications. The benefits to scientific communities of data citations include: 1) formal citations give credit to scientists for their work in collecting and creating data, 2) formal citations will allow data center managers to track the use of data sets and gain the benefits of documenting their services and creating a foundation to design better services, and 3) formal citations will help accelerate scientific progress by tightly coupling scholary publications and data, so that two-way discovery and access are common.

In order for data citations to serve these desired roles, however, *there must balanced support for citation-linking technology, promotion of data citations within research settings, improved bibliometric measurements of data citations, and greater acceptance of data citations as an indicator of scientific impact within research organizations.*

References:

- American Geophysical Union (AGU). (2009). *AGU Position Statement: The Importance of Long-term Preservation and Accessibility of Geophysical Data.* http://www.agu.org/sci_pol/positions/geodata.shtml
- Bizer, C. (2009). The Emerging Web of Linked Data. *IEEE Intelligent Systems* 24(5): 87-92. http://dx.doi.org/10.1109/MIS.2009.102
- Federation of Earth Science Information Partners (ESIP). (2011). *Interagency Data Stewardship/Citations/provider guidelines.* http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines
- International Polar Year (IPY). 2008. *International Polar Year 2007-2008 Data Policy.* http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf
- Mayernik, M.S., A.L. Batcheller, & C.L. Borgman. (2011). How Institutional Factors Influence the Creation of Scientific Metadata. In *Proceedings of the 2011 iConference* (iConference '11). New York, NY: ACM (pg. 417-425). http://doi.acm.org/10.1145/1940761.1940818
- National Academy of Sciences (NAS). (2011). *Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop.* http://sites.nationalacademies.org/PGA/brdi/PGA_064019
- National Science Foundation (NSF). (2010). *Division of Earth Sciences Data Policy.* http://www.nsf.gov/geo/ear/2010EAR_data_policy_9_28_10.pdf
- Parsons, M.A., Duerr, R., & Minster, J.-B. (2010). Data Citation and Peer Review. *Eos Transactions, AGU,* 91(34). http://dx.doi.org/10.1029/2010EO340001
- Pienta, A.M., Alter, G., & Lyle, L. (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

- Stodden, V. (2009). The Legal Framework for Reproducible Scientific Research: Licensing and Copyright. *Computing in Science & Engineering* 11(1): 35-40. http://dx.doi.org/10.1109/MCSE.2009.19
- Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., & Warner, S. (2004). Rethinking scholarly communication: Building the system that scholars deserve. *D-Lib Magazine* 10(9). http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html
- Wilbanks, J. (2008). Public domain, copyright licenses and the freedom to integrate science. *Journal of Science Communication* 7(2). http://jcom.sissa.it/archive/07/02/Jcom0702%282008%29C01/Jcom0702%282008%29C04/Jcom0702%282008%29C04.pdf
- Wynholds, L., Fearon, D.S, Borgman, C.B., & Traweek, S. (2011). When use cases are not useful: data practices, astronomy, and digital libraries. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (JCDL '11). New York, NY: ACM (pg. 383-386). http://doi.acm.org/10.1145/1998076.1998146