January 11, 2012

White House Office of Science and Technology Policy
Comments on RFI: Public Access to Digital Data Resulting From Federally Funded Scientific Research Research.

Response from Arizona State University Libraries

There are four different areas in which the Federal government can make substantive contributions in the effort to ensure that publically financed research data are made widely available, and are preserved and curated for the long-term: technical (build robust, sustainable technical infrastructure); standards (establish national as well as discipline-specific standards for data and metadata); sustainability (determine what services should be supported as a basic part of a university research environment and what services need to be offered on a cost-recovery basis); and governance (ensure that solutions and initiatives undertaken by various agencies, universities and other stakeholder groups are coordinated and communicated widely, and that issues are prioritized and solved in ways that benefit the majority of stakeholders).

At present, most if not all research universities are struggling with similar issues: how to preserve research data and make it accessible for the long-term; how to support increasingly complex E-science research projects; and how to build and maintain a sustainable cyberinfrastructure in times of economic downturn and budget cuts. Universities have an important role to play in addressing these issues, as do the professional societies, the Federal government and private industry. While having grassroots solutions developed by several hundred universities has advantages, the concern is that the results will be diverse, incompatible solutions for the same problems. The Federal government can play a much needed role in coordinating the individual efforts of all stakeholders thus ensuring that the results are complementary, interoperable, and communicated broadly.

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?*

A necessary first step in preserving digital data, and making it accessible for the long-term, is to mandate that publically funded research data be deposited in open access or publicly accessible digital repositories, unless confidentiality or privacy concerns prevent this. Since Federally funded research data has been paid for by the public, it should be accessible to the public in as timely a manner as possible. Requests for short embargo periods (e.g., 6 months to 3 years) could be accommodated but recent experience with genomics data indicates that when research results are made public quickly, scientific advances proceed apace and new commercial applications and product development rapidly results (Williams 2010). The current norm, that individual researchers post their data on personal web sites, or respond to individual requests for

copies of their research data, is risky (from a preservation standpoint) and not sustainable over the long-term (what happens when the researcher retires?).

The second critical issue is the need for open access and not-for-profit digital repositories to be interoperable, perhaps linked by a Federal portal. Here, the important Federal role is to coordinate efforts across the various scientific disciplines to establish standard data and metadata formats, and mandate their use. At the very least, there should be a national clearinghouse for best practices for research data (data formats, metadata, data management using a lifecycle approach) from the various fields.

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?*

Balancing access and intellectual property rights in the digital age is complex. Texts and other digitally published works have the option to include license or access restrictions by utilizing a standardized infrastructure, such as that provided by Creative Commons. However, there is widespread concern that a Creative Commons model for research data is problematic, in part because facts cannot be copyright protected and in part because data – unlike a published text – is mutable (see, for example, de Cock Buning et al. 2009). A useful Federal role would be to coordinate existing efforts, like those of the Open Knowledge Foundation, to help establish best practices for public data access and licensing.

Intellectual property issues notwithstanding, a mandate that data be deposited in an appropriate digital repository is essential to ensure curation, preservation and to facilitate public access. Limited embargo periods, as mentioned in question #1, can be accommodated by digital repositories with the added benefit that the data are stored securely and preserved during the embargo period.

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?*

An existing Federal organizational structure is in place as a result of the disciplinary focused granting agencies (NSF directorates, NEH, NIH). Working with the professional and learned societies, this structure can be used to help ensure that disciplinary differences are recognized and accounted for. However, what is missing is a national coordinating body to facilitate communication among the various disciplines and representatives from the Federal funding agencies. Cross-disciplinary standards should still be a priority to help ensure that data are preserved and accessible for the long-term. These include minimum standards for descriptive and technical metadata, file formats, the use of disciplinary specific controlled vocabularies, and permanent digital identifiers (DOIs).

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?*

Ultimately, assessing the significance of research data sets is a task that should fall to the researcher, the repository, and be informed by standards and practices of the various professional societies. Federal agencies might usefully work with stakeholder groups to establish guidelines for evaluating research materials, and suggest best practices for archiving and retention schedules, similar to document retention guidelines for public documents and business records.

It is important to recognize that even after scientific data are no longer current, they retain important historical value. University institutional and digital repositories generally take the approach that all deposited materials should be preserved for the long-term. Like other cultural memory institutions (e.g., museums, libraries and archives) University digital repositories hold the scholarly and creative output from that institution and assume a stewardship role that lasts in perpetuity.

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

Since January 18, 2011, when NSF required data management plans for grant proposals in all directorates, followed by a similar announcement from the NEH Office of Digital Humanities (June, 2011), research universities throughout the country have expended considerable effort in providing support and best practice information to faculty and researchers. As noted in the second paragraph on page one, a diversity of results from individual efforts is valuable, but more valuable is the active coordination among institutions so that successes and lessons learned can be shared. The Federal government can play a much needed role in coordinating the individual efforts of all stakeholders thus ensuring that results from the various efforts are complementary and communicated broadly.

Data management planning, throughout the active research period of the grant-funded project and beyond, involves a partnership between data producers, repositories and Federal funding agencies. The "stick," i.e., that a data management plan is required by funders and evaluated as part of the proposal review process, is certainly important. However, digital repositories need to develop services and mechanisms that researchers consider valuable, to serve as a "carrot" that encourages researchers to describe their data with adequate metadata and deposit it in sustainable formats in publically accessible repositories. An example of a basic repository service would be to assign permanent, unique identifiers (DOIs) so that data sets can be cited unambiguously. Tracking citations, as well as data set views and downloads, are value added services because they provide useful measures of research impact, measures that can be very important in tenure and promotion decisions.

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

Digital repositories are expensive, with costs primarily falling into three categories: repository staff salaries, short-term file access and storage, and long-term data curation and preservation services (Goldstein and Ratliff 2010; Rumsey 2010). Universities can reasonably view repositories as a new and essential component of the research infrastructure, on a par with libraries, museum collections, and university archives. As such, basic funding should come, at least in part, from grant indirect costs. However, as with other research costs that exceed the basic campus infrastructure, some digital curation and preservation costs should be eligible for inclusion as direct costs, particularly for projects where large amounts of data are generated (e.g., terabytes or petabytes) and/or where the data are complex and difficult to archive. Currently, there is no clear understanding of how to establish the line between what can be managed as a part of the "basic" university research infrastructure, and what exceeds that threshold. We need a national referendum where stakeholders with experience in this area can discuss the costs of preservation and possible solutions for establishing sustainable repositories.

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?*

Adopting a research data lifecycle approach, such as outlined by the Digital Curation Centre in the U.K., is helpful. At a minimum, there are three points in the lifecycle at which compliance could be systematically evaluated: grant proposal review (data management plan), active grant research phase (data production and storage), and data publication at the end of the research project (deposit in a public access repository).

Grant reviewers hold researchers accountable for proposing a reasonable and feasible data management plan as a normal part of the review process. Agencies could provide guidelines to proposal reviewers enumerating the essential elements of a well-developed data management plan tailored to disciplinary best practices (e.g., NSF could develop such reviewer guidelines by directorate).
University sponsored program offices typically monitor researcher compliance with agency mandates during the active research phase, though this is typically limited to audits of financial expenditures. A useful Federal role might be to coordinate stakeholder meetings to develop systematic and automated workflows that could become part of the University's monitoring process. Minimally this might involve requiring project principal investigators to complete an annual or semi-annual compliance form, stating that they have implemented the steps they outlined in their data management plan.

Finally, if research data are deposited in public, open access repositories, agencies as well as university sponsored program offices could systematically verify that appropriate files had been deposited and were accessible. Linking repository based research data sets with publications, something the ecological sciences achieve using the Dryad repository, is one approach that has worked well (Beagrie et al. 2009). A necessary activity, however, is to provide opportunities for disciplinary stakeholders to discuss these compliance measures, thereby ensuring that sufficient attention is given to differences between scientific disciplines and different types of digital data.

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?*

One of the primary goals of the America COMPETES Reauthorization Act of 2010 (Gonzalez et al. 2010) is to improve education in science, technology, engineering, and mathematics (STEM) in support of innovative research in the physical sciences and engineering. If we mandate that publically funded research data be deposited in open access or publicly accessible digital repositories, we build a valuable reservoir of educational materials that are easily available to K-12 teachers as well as college and university instructors. The next step is to encourage the widespread use of these resources, by encouraging teachers and students to explore repositories and incorporate digital collections into the curriculum. Purdue University has successfully pioneered this approach using their hubZERO digital repository platform (Magana 2010).

In partnership with researchers, new repository services could then be developed, such as providing sample homework assignments, curriculum modules, and learning objects for a variety of grade levels. This would help satisfy the broader impacts expectation mandated in NSF funding proposals.

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?*

As mentioned in our answer to question #5, repositories can build services (such as assigning DOIs) so that data sets can be cited unambiguously. At present scholarly disciplines have well – developed procedures for citing published articles, books and other texts. Coordinating disciplinary referenda with the professional societies, universities, federal agencies and other stakeholders would allow for the development of data citation systems that are appropriate for the various scientific fields.

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.*

With the growth of disciplinary repositories, and data repositories linked to journal publications, the various scientific fields have begun an active engagement with topic of metadata and other standards. As with other digital data management efforts, however, these tend to be conducted in isolation, without as much national or international input as needed. Coordination by agencies such as the Library of Congress and the National Archives and Records Administration would be helpful in enhancing communication and reducing duplication of effort.

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?

During the 1990s the Federal Geographic Data Committee (FGDC) successfully established the Content Standard for Digital Geospatial Metadata, building on nearly two decades of prior work by various federal agencies involved in geographic information systems. Several factors led to the widespread adoption of this standard, including early and persistent engagement by a variety of federal and industry stakeholders, simple tools incorporated into GIS software by commercial vendors that allowed users to easily enter metadata, and additional software tools that updated metadata automatically in response to file changes. The FGDC standard is currently being revised to a North American Profile, allowing for important updates and changes to be made to keep pace with changes within both the GIS and technological realms.

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

The European Union has been quite successful in establishing standards among EU nations that result in applications that assemble disparate resources into a single web platform, easily accessible to users from around the world. For example, Europeana is a web portal that brings together digital cultural material from major European art museums, libraries, and other cultural heritage institutions. To leverage this experience, the U.S. should seek to establish working groups with European and Australian cyberinfrastructure partners. Increasingly, NSF has been coordinating grant programs with JISC (the Joint Information Systems Committee of the UK) and other others. Similarly, the recent National Endowment for the Humanities "Digging Into Data" challenge brought together a large group of international research funding agencies, representing Canada, the Netherlands, the United Kingdom, and the United States. Coordination at this level would seem to be helpful in fostering information exchange while still remaining sensitive to disciplinary differences and challenges.

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

As noted in our answer for question #7, a number of disciplinary repositories have successfully established links between repository data files (and other associated file types) and published journal articles, books and reports (e.g., Dryad). It is helpful if disciplinary repositories seek out partnerships with appropriate publishers and professional societies. There are mutual benefits from this kind of commercial / repository profit partnership. For example, tDAR (the Digital Archaeological Record), a not-for-profit repository for digital archaeological data, is able to link disparate information about an archaeological site, a research topic or a geographic area, by including metadata from commercial publishing firms with the metadata and documents in its repository (McManamon and Kintigh 2010). Publishers gain an inexpensive and easy way of advertizing their publications. Repositories gain additional digital resources that they can make available to users. The overall benefit is that available information is made more easily

discoverable, accessible, and usable. Users gain a "one-stop-shopping" experience that increases accessibility and expands the number of relevant search results for users.

These comments are submitted on behalf of Arizona State University Libraries by:

Mary Whelan
Geospatial Data Manager

Sherrie Schmidt
University Librarian

References Cited

Beagrie, Neil, Lorraine Eakin-Richards, and Todd Vision
2009. "Business Models and Cost Estimation: Dryad Repository Case Study." Society 1:1-6. Retrieved from http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beagrie-37.pdf.

de Cock Buning, Madeleine, Allard Ringnalda and Tina van der Linden
2009. The Legal Status of Raw Data: a Guide for Research Practice. Report of the Centre for Intellectual Property Law, the Netherlands. Retrieved from www.surf.nl/publicaties.

Goldstein, Serge J. and Mark Ratliff
2010. DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data. Report for the Office of Information Technology, Princeton University. Retrieved from: http://arks.princeton.edu/ark:/88435/dsp01w6634361k.

Gonzalez, Heather B., John F. Sargent Jr., and Patricia Moloney Figliola
2010. America COMPETES Reauthorization Act of 2010 (H.R. 5116) and the America COMPETES Act (P.L. 110-69): Selected Policy Issues. Congressional Research Service Report for Congress. Retrieved from: http://www.ift.org/public-policy-and-regulations/~/media/Public%20Policy/0728AmericaCompetesAct.pdf.

Magana, Alejandra J.
2010. "How Engineering Instructors Use NanoHUB Simulations as Learning Tools?" Retrieved from: http://nanohub.org/resources/8742 .

McManamon, Francis P. and Keith W. Kintigh
2010. "Digital Antiquity: Transforming Archaeological Data into Knowledge." SAA Archaeological Record 10(2):37-40.

Rumsey, Abby Smith (editor)
2010. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, Retrieved from: http://brtf.sdsc.edu/ .

Williams, Heidi L.
2010. Intellectual Property Rights and Innovation: Evidence from the Human Genome. National Bureau of Economic Research, Working Paper Series, Working Paper 16213. Retrieved from: http://t.co/9OYy9DjO.