TO: Office of Science and Technology Policy publicaccess@ostp.gov
FROM: Daniel Lee, Tucson, Arizona
RE: Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research
DATE: Wednesday, January 11, 2012

The following comments are in response to the request for information issued November 4, 2011, by the Office of Science and Technology Policy (OSTP) regarding recommendations on approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally funded scientific research. I would like to thank OSTP for the opportunity to respond and contribute to the conversation. My comments follow.

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?

- Federal policies that require making appropriate digital data resulting from federally funded scientific research openly available in repositories committed to preservation and access would be a significant stimulant to the American scientific enterprise and to the economy as a whole. (Classified data and data that include personally identifiable information would certainly be inappropriate for inclusion in such policies.)
- On the one hand, such policies would allow for verification of findings and analyses by outside researchers. Verification and reproducibility are the very heart of the scientific enterprise. By facilitating this activity funding agencies would be contributing to the credibility of the funded research and thus promoting further progress by researchers who build on these results.
- On the other hand, in many cases these same policies would expedite further research by saving subsequent researchers from recreating the data that was already collected and produced. There is clear potential for saving both funds and time, thus allowing research funding to go further and accomplish more.
- Open data policies would also have the added benefit of creating resources where businesses large and small would have information available to them to use as they see fit to create new products, services, and markets that can drive economic growth. Further, by providing broad, ongoing access to data, a wide range of research could be promoted including interdisciplinary projects apart from the expected uses intended by the initial 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?

- First, it is important to recognize that copyright rests with the author/creator unless and until s/he transfers that right to another party such as a publisher. A research funder could reasonably require authors and creators to adjust these rights in some way or other. Research funders that are the author's employer might indeed be the rights holder themselves and thus impose even greater control over how rights are managed. Given that scientist/authors are often more interested in spreading results (while getting credit) and having impact than in controlling rights, protecting the intellectual property of publishers doesn't seem like a helpful place to start.

- This isn't to say that publishers don't add value and that their investment in that value doesn't need protection of some sort. They do and it does. If publishers invest in hosting and providing access to the data that supports the papers they publish, that investment does deserve protection. One form this protection might take is treating the data as part of the publication that makes up the version of record and is cited as such. The formal recognition that comes from citation and the granting of authority that comes with it is indeed a form of protection that drives future business their way.

(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?

- One important way Federal agencies could account for disciplinary differences in regards to what data is usefully shared and how it should be presented is to build flexibility into an agency's policy that relies on the program directors and reviewers for each program area to largely define what data should be shared, in what time frame it should be shared, the mode it should be shared in, and where it should be shared within a broad mandate for making relevant data openly available.
- It is also important to recognize, though, that while allowing for differences in data types there also needs to be sufficient commonality to allow for and promote cross-disciplinary discovery and reuse. Much of the advantage of open data is creating the possibility of discovering data sets that were collected or created for one project in one field that serves useful for solving problems in a separate field unexpected and unintended by the original researcher.

(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?

- As in comment #3, the value inherent in differences in costs and benefits of long-term stewardship and dissemination will depend on differences of disciplinary needs. Agency policies should allow for those differences and allow for those in the research areas to help define those needs.

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

- Research communities can make a huge contribution to the implementation of data management plans by reaching consensus within each community on what makes up a high quality plan. Universities, research institutions, and their libraries can assist the communities in developing these best practices and share successful approaches in centralized web site linked to other research compliance support.
- Among the topics that would likely be included in support materials are guidelines for depositing the resulting files in the institutional repository where appropriate, promoting consideration of the issues around sharing data early on in project development to save time later having to re-work data into usable formats and forms at the end of the project, and advice on useful, practicable metadata templates and standards.

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

- Once funding agencies begin requiring preserving and making digital data accessible grant announcements should clearly indicate that the costs required to achieve these ends are expected to be included in the budgets as part of proposals and awards. Working with the constituencies, agencies can also promote efforts to create best practices in these areas that will help researchers understand and define the real costs. Among the issues to be addressed is the need to develop business models for one-time payment out of grants that account for the ongoing costs of continued, persistent access.

(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?

- This is a very important goal. Simplifying processes that funded scientists and their institutions go through to meet agency requirements for accountability would lower overhead costs and allow for more research to occur. Researchers want to do research and share the results, not satisfy bureaucracies.

- Simple procedures that fit into an existing workflow have the best chance of achieving desired ends with minimal additional burdens.

(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?

- Besides requiring data resulting for federally funded research, agencies could stimulate innovative use of research data by facilitating standardization of the infrastructure that supports data modeling and data management. Within the context of differences of data types, such standardization promotes findability and reuse, thus allowing entrepreneurs eased access to research findings to create new products, services and markets and to enhance existing ones.

(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?

- Cultures of citation and other forms of recognition develop and are enforced within disciplines. The greatest contribution agencies and others outside the specific fields could make would be to create and encourage standardized metadata schemas and templates that clearly indicate the responsible parties and that delineate the various roles in gathering and producing the data.

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

- I am confident other submitters will be more up to date and exhaustive than I can be on this issue.

(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 example of the Federal Geographic Data Committee (FGDC) is a model worth considering in other research areas. As a major factor in this research community, the FGDC was able to reach agreement on a standard that then was adopted by smaller venues in the field.

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

- Federal agencies could promote effective coordination on digital data standards with other nations and international communities by working closely with equivalent agencies and other scientific organizations as full participants in ISO (http://www.iso.org/iso/home.htm) standards development.

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

- This is a key issue to confront if we are to take full advantage of reusable data. Agencies could support linkage by requiring funded researchers to assign persistent unique identifiers to each data set resulting from the funded project and referring to the data through the unique identifier in all resulting publications. The California Digital Library has created one such tool call EZID (http://www.cdlib.org/uc3/ezid/). Such an identifier would link publications (ideally using Digital Object Identifiers) with uniquely identified authors using services such as ORCID (http://www.orcid.org/) with the data using EZID. Whether the data is archived by a publisher, a university, or a funding agencies wouldn't matter as long as the unique identifier stays with the associated file.