UNIVERSITY LIBRARIES

January 12, 2012

Office of Science and Technology Policy
The White House

Re: Request for Information: Public Access to Peer-Reviewed Scholarly Publications Resulting from Federally Funded Research

The University of Maryland Libraries write in response to the Request for Information, published in the *Federal Register* on November 4, 2011, by the Office of Scientific and Technology Policy regarding public access to peer-reviewed scholarly publications resulting from federally funded research. We appreciate this opportunity to comment.

(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?

UMD Answer: NIH Public Access policy (NOT-OD-05-022) and the NSF Data Management Plan requirement.

(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?

UMD Answer: A variety of steps can be taken to protect the intellectual property interests of data creators. Some examples might include allowing for the ability to "embargo" works, much as we do with theses and dissertations, for a set amount of time. We can design preservation architecture that allow for permissions management. In addition, researchers should have some leeway in how they package the data that they wish to have publicly disseminated. For example, they may choose to make a subset of the data fully available, but withhold other elements of the data that while not crucial to the final results, could require future users to come up with their own methods of display and analysis. In some cases, however, having a deadline by which data must be made publicly available can actually serve as an incentive to researchers to publish their most critical findings first and in a more timely manner.

(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?

UMD Answer: Federal agencies need to be flexible and design preservation systems that can manage diverse types of data. They also need to be highly aware of the costs and size of the data. Some scientific research can generate hundreds of terabytes of data in a short amount of time and on a regular basis, while others, only a few gigabytes over long periods of time. First, researchers should not necessarily feel that they must restrict their research or computations based solely on size of data, but second, accounting for larger datasets is something that should be taken into account when asking for funding.

Since presentation of and access to data from various disciplines vary widely and since customizing access for various purposes requires expertise about the data depending on the discipline from which use cases develop, it might be best for Federal agencies to limit their services to presentation and access of whole datasets in their raw form. Instead, Federal agencies can enable, through crowdsourcing tools and other means, the user communities to develop special and customized presentation and access services.

(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?

UMD Answer: It is tricky to talk about the costs of long-term stewardship and dissemination without somehow placing restrictive quotas or penalizing those disciplines that naturally create large datasets of digital data. It is most important that agencies first recognize that there are differences in costs of different types of data, and then second perhaps conduct some focus groups and studies about the usefulness of different types of data to future researchers. Much of this discussion has to come from within the disciplines and with the data creators themselves. The big question is an appraisal one – what of the data needs to be retained? Of 100 terabytes of astronomical data, perhaps only 10% is useful or worth retaining. But without the other 90%, it may take a future researcher an incredible amount of time to use the data or reconstruct an experiment. Packaging data for reuse and preservation may be an extremely time-intensive process and something that should be accounted for in the costs when assigning grants, as well as when writing data management plans.

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

UMD Answer: Stakeholders can best contribute to implementation of data management plans by working together to provide guidelines, templates and services to better prepare data for long-term preservation. By working with the data creators from the beginning of their research, and by collecting relevant information to enable future access and use of the data, libraries, research institutions, etc. can ensure that they have the correct tools in place to fulfill the data management plans. In addition, stakeholders can provide staffing, in the form of data scientists, archivists, etc. who can assist in the process of preparing the data for final deposit/dissemination.

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

UMD Answer: Allow or require that the data management plans included in grants contain a line-item in the budget to account for start-up and maintenance costs of digital data.

(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?

UMD Answer: Registries would be one way to measure and verify compliance with Federal data stewardship and access policies for scientific research. The caveat is that registries are often dependent on self-selection and user submission. Adding a level of prestige or legitimacy to inclusion in a registry would encourage more active participation. Peer review is another mechanism to improve compliance by having panels of peers review and monitor, by audits and tests, compliance with policies. In addition, more direct involvement in the compliance process from agencies (e.g. National Archives and Records Administration, Library of Congress) for whom data stewardship is an important part of their mission and existing skill set, would help greatly in legitimizing the verification process. In general, use of public data over time by peers will serve as a natural audit mechanism, as well. Data that is important to a group of researchers will be used heavily and with scrutiny. Any failure in preserving the data will be flagged by users. It is important that Federal and other stakeholders monitor and act on the flags, to help recover lost data and in cases where data is not recoverable, learn from lessons over time and reduce risk of data loss.

(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?

UMD Answer: In order to stimulate innovative use of publicly accessible research data in new and existing markets, agencies could devote more time and funding to communicating availability of the data and provide, as part of the communication, an effective description of the available data. Stimulus grants, such as the National Science Foundation's Digging into Data Program, would help to jump start the process.

(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?

UMD Answer: There are currently no standard and reliable mechanisms to ensure this happens with non-digital data. The best mechanism is to instruct researchers on how to properly cite and document their work. Use technology (like watermarks for images) to ensure that data can always be traced to its source. Most technologies do not lend themselves well to the entire array of scientific data. The various disciplines likely have some way of ensuring this, and the most useful thing would be for the Federal policies to document all of this, so that it becomes very clear how secondary use of data is cited.

(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.

UMD Answer: This is an area where librarians can step in and help facilitate a process. Issues to address would be determining common metadata, and minimum metadata required. In cases where each discipline has solutions for these problems, librarians can develop processes and mechanisms to ensure that access and preservation systems can interpret various types of data standards, rather than requiring all to fit within a narrow scheme.

(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?

UMD Answer: The library and archives professions have created a host of long-lasting and effective standards over the years: MARC, Dublin Core, Resource Description and Access (RDA), OAI-PMH, Encoded Archival Description (EAD), to name a few. What makes these standards successful is widespread adoption by their communities, extensive documentation on usage, their ability to set "minimums" to enable compliance at various levels, their interoperability, their appropriateness to the material that they are describing, and metadata creation tools that allow for consistency and uniformity.

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

UMD Answer: Some existing programs that can help with this are the National Digital Information Infrastructure Program (NDIIP), and international bodies such as Digital Preservation Europe, who are already working with international communities within each discipline. Another example is the program to design a Microbial Research Commons as

described in a recent BRDI (Board on Research Data and Information) and BLS (Board on Life Sciences) symposium.

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

UMD Answer: Embrace and encourage open access. Standards for minimal set(s) of metadata for the datasets that have clear semantics and are normalized, at least within a discipline, bibliographic and keyword or subject heading standards for the publications that are compatible with the dataset metadata standards.