January 12, 2012

RFI: OSTP, Public Access to Digital Data Resulting From Federally Funded Scientific Research
Contact: Francis P. McManamon fpmcmanamon@asu.edu
Center for Digital Antiquity [Arizona State University]
Tempe, Arizona

Questions and Comments:

Preservation, Discoverability, and Access:

1) What specific Federal policies would encourage public access to and the preservation of valuable digital data resulting from federally funded scientific research, to grow the US economy and improve the productivity of the American scientific enterprise?

We suggest several specific Federal policies that would encourage preservation and access to data resulting from federally funded scientific research. These include:

(a) Require that data generated from federally funded research be archived in an appropriate trusted digital repository or archive dedicated to the preservation of and access to data and supporting documentation. Part of this requirement should include the creation of appropriate and sufficient metadata for discovery, so that data are not simply preserved, but also readily accessed and interpreted for future research. Regarding the kinds of digital repositories most appropriate for data archiving, we suggest that discipline-specific repositories provide a rich context of similar materials so that users, as they search, are provided with search results tailored to their expectations and needs. Metadata in disciplinary repositories contains phrases, keywords, and categories that match subject matter domains, making search results much more targeted to information resources that are especially useful. In addition, specialized digital repositories, as opposed to institutions or organizations with generalized missions, can increase efficiency and productivity of tasks related to digital archiving, and minimize the costs of data access and preservation to researchers and traditional archives maintained by libraries, museums, and universities.

(b) Encourage data repositories to include in their archive any documentation relevant to the original data set. For example, repositories should include reports related to the data. Repositories also should ensure appropriate linkage to metadata, which would ease searches among related data and make background research more efficient.

DIGITAL ANTIQUITY
School of Human Evolution and Social Change
an academic unit of the College of Liberal Arts and Sciences
Francis P. McManamon, Executive Director
PO Box 872402
Tempe, AZ 85287-2402
Direct: (480) 965-6510 Digital Antiquity: (480) 965-1369 Fax: (480) 965-7671
http://digitalantiquity.org http://tdar.org fpm@digitalantiquity.org

(c) Encourage the re-use of existing successful software tools across disciplines where available (to both decrease costs of implementing archival systems and to foster interdisciplinary and integrative research).

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?

Public access and use might be enhanced by expanding the "fair use" concept applied to sharing documents (even when copyrighted) and data, if the use is for non-commercial educational, scholarly, or scientific purposes.

In increasing public access, it is important to develop and employ administrative procedures, data organization, and/or publishing formats that control appropriate access to "confidential data," for example, very specific locations of archaeological sites that might be subject to looting if their locations were generally known. It also would be desirable to develop procedures to provide appropriate warnings concerning sensitive information that may be available in widely shared data, documents, or images.

In developing procedures for wider access to scientific data, agencies must balance the requirement of placement in digital archives with a short embargo within the archive to facilitate the completion of ongoing research while ensuring future public access via the archive. Creative Commons and other open access licensing formats have worked well for document and other text data, but are not completely appropriate for research data. There are several current efforts (e.g., the Open Data Commons Project) to develop appropriate licenses for research data. On balance, though, every effort should be made to encourage public and professional access to data through appropriate channels. Related to this wider access to data, policies and procedures are needed to ensure and encourage the appropriate citation, crediting and attribution for individual researchers and organizations that carry out research and produce the data being shared.

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?

Federal agencies should work with other agencies or organizations (like the National Science Foundation) that have well-established disciplinary directorates to develop policies relevant to their area of expertise. A number of disciplines are already addressing issues of data preservation and access specific to their research; it would be cost-effective and reduce duplication of effort to work with the professional societies in developing policies. For example, within the discipline of archaeology, the Center for Digital Antiquity (http://digitalantiquity.org), the Open Context repository (http://opencontext.org/), and the Archaeology Data Service in the UK have cooperated on various topics related to the digital archiving and providing for access and use of archaeological data.

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?

Considering the cost of lost data versus the costs of long term preservation, lost or inaccessible data essentially means that all Federal monies spent on a research project were expended for no result. In such instances, there is no potential for current or future benefit to the public or the American scientific enterprise. Alternately, preservation and access may require marginally more Federal funding to ensure that the research results are accessible and preserved, but those costs are amortized over a very long time period and will have a broad range of economic, scientific, and educational benefits.

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

Stakeholders are typically generalists (such as universities) or specialists in specific disciplines or topics other than digital archiving. Their efforts would be best steered toward implementation of policies that require submission of data to appropriate disciplinary digital archives. Stakeholders should participate in the development of guides to good practice for researchers contributing data, both broadly (libraries, universities) and in specific disciplinary research communities. Stakeholders should require authors and publishers to provide data related to their publications in ways that facilitate archiving and in standard file formats that are susceptible to archiving and ultimate migration as new formats develop. Researchers should be required to organize their data in ways that facilitate archiving and in standard digital formats that are amenable to archiving and conversion as new formats are developed. Likewise, authors should provide appropriate documentation for all archived materials to ensure that the information is useful to various stakeholders.

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

Computational science models predict that the cost of storage will continue to decrease over time. However, the cost for long-term archiving and curation are more complex because files have to be converted as software becomes obsolete and hardware and software advances are made. A "pay once, store forever" model (the so called Princeton model, developed by Goldstein and Ratliff, 2010 [http://arks.princeton.edu/ark:/88435/dsp01w6634361k]) is currently the most reasonable option. Grant funded projects should include an appropriate direct cost line item for long-term curation and preservation of digital research data.

Life-cycle cost analysis is a useful method of determining or estimating the real costs of preserving and making digital data accessible. Requiring funding mechanisms to incorporate ways to explicitly address the life-cycle costs associated with the maintenance of digital data will be an improvement over current approaches. Relying on discipline-specific and specialist repositories will both minimize costs associated with digital data maintenance (economies of scale) and those expert facilities will be well situated to realistically determine the costs associated specifically with the long-term preservation of and access to digital resources.

7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access polices for scientific research? How can the burden of compliance and verification be minimized?

Agencies can require that researchers submit as part of their invoicing for research grants or in their reports on grant expenditures, official documents from the digital archive in which their data and documentation has been placed. These documents need not be complicated, but should verify that the research data and documentation have been deposited in the digital repository along with appropriate metadata and that they are now accessible. The archive also should affirm that the deposit of the data and documentation means that these will be preserved for future use in the archive.

In the planning for a research project and as part of the data management plan submitted with the research proposal, the steps should be described to track digital data from its creation through placement in an appropriate archive.

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?

Agencies should provide incentives through grants specifically targeted for integrative research drawing on digitally archived data. These could have very high returns at small cost: new data would not need to be generated - but bigger-picture questions beyond any single project could be addressed.

Tools like the Sunlight labs have shown the value of data when available, by taking the step of requiring it to be captured digitally and available this would facilitate significant progress for entrepreneurship. Agencies also could offer grants to "rescue" important data in out-of-date software programs or media. This would be particularly important for lost or threatened data from past federally funded research. A systematic approach to digitally archiving legacy data would be an ambitious and extraordinarily important project for many disciplines.

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 a condition of grants, agencies could require thorough citation details as part of the metadata for archived resources. Appropriate attribution (whatever the source) is a fundamental part of the responsible conduct of scientific research and should be explicit in the training of students, scientists, and other professionals. Disclaimers and terms-of-use could be required by the digital repository laying out expectations of proper attribution if the data or supporting documentation is used in any fashion. Repositories also should be encouraged to offer services that provide, for example, a standard format for citing the research data set (similar to the formats used for citing published works). Included in this citation should be permanent identifiers that are associated with a particular data set (e.g., Digital Object Identifiers [DOIs]).

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-driving data standards effort.

The effort should start with better practices for the long term preservation of digital data. (an example would be the *Guides to Good Practice* (http://guides.archaeologydataservice.ac.uk/) developed for archaeological archiving and data files by the Archaeology Data Service in the UK and the Center for Digital Antiquity in the US.

Libraries and archives already have a number of standards to support basic interoperability, but discipline specific metadata requirements would be a significant first start.

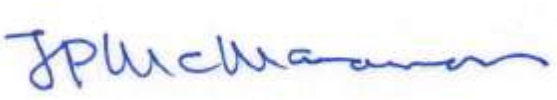
12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

Agencies could provide funds for cooperative activities between disciplinary archives in different countries, e.g., for archaeology, the Archaeology Data Service in the UK and the Center for Digital Archaeology in the US. Agencies also could fund archival projects that develop links among repositories, e.g., the TransAtlantic Gateway project (funded by the US National Endowment for the Humanities and the JISC in the UK).

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

Policies that encourage publishers, specifically those who publish top tier journals for specific disciplines, to work with established archives to link data with articles should be developed. This has worked well in the field of ecology, where the data archive, DRYAD, holds research data related to publications in various ecological journals. When an article is accepted for publications, the journal requires that the relevant data sets be deposited in DRYAD. Similar approaches should be developed and encouraged in other disciplines by agencies through funded initiatives.

Sincerely,



Francis P. McManamon, Ph.D., RPA
Executive Director and Research Professor