12-Jan-2012

TO: Science and Technology Policy Office

FR: Edward Van Gemert, Interim Director, University of Wisconsin-Madison Libraries
and Bruce Maas, CIO and Vice Provost for Information Technology, University of
Wisconsin-Madison

Response: Office of Science and Technology Policy Request for Information: Public
Access to Digital Data Resulting from Federally Funded Scientific Research
RFI Document Citation: 76 FR 68517
RFI Document Number: 2011-28621

The University of Wisconsin at Madison's General Library System and the Office of the
CIO, in consultation with its Research Data Services staff, strongly endorse OSTP's
interest in preserving and providing public access to digital data from federally-funded
research. Responding to OSTP's specific questions:

(1) Blanket preservation policies should apply to digital data arising from Federal
grants. These policies should define as clearly as possible, taking disciplinary
differences and research workflows into account, which data are of sufficient
importance, quality, and reusability to warrant the cost of preservation. Policies
should authorize either specific preservation retention schedules or periodic
reassessment of preserved datasets so that obsolete ones may be discarded. Access
policies, which must be considered separately from preservation policies, should
require public access except in clear cases of human-subjects confidentiality, national
security, or similar. Institutional Review Boards may be suitable arbiters of human-
subjects questions surrounding access to data, but clear Federal guidance will help
them considerably.

Granting agencies requiring data management plans should strive for consistency in
terms of the data plan requirements, with each plan addressing data preservation,
data security, and access. To the extent possible, such consistency will encourage
easier compliance resulting in improve access to a greater amount of material over
time. These requirements should be integrated in the grant submission guidelines,
clearly outlining the purpose and elements of the data plan. At the time of the
award, grant recipients should have a documented and clear understanding of their
responsibilities with respect to data retention including retention schedules, which
data are to be retained (e.g. raw data, summaries, etc.), access rules, and so forth. An
additional suggestion pertaining to IRB policies should be considered: as part of IRB
policies, study participant consent forms should provide information indicating that
certain data they provide could be used in other contexts.

(2) It is vital to remember, and for Federal policies to state clearly, that many datasets do not meet the originality standard for copyright. For such data as do have copyright or patent encumbrances, however, and to accommodate most disciplinary cultures, Federal policies should allow delayed (but not indefinitely-delayed) public access to data. Deposition into suitable data archives should be as immediate as possible, as this best protects dataset viability, but Federal policy should permit embargoed access until after publication, after patent application, and/or for a discrete length of time after grant end. Federal agencies should insist that data be licensed for reuse, commercial and non-, via licenses such as the Open Data Commons Public Domain Dedication and License (opendatacommons.org/licenses/pddl).

(3) The National Science Foundation's implementation of its data-management-plan policy is an excellent example: the NSF's broad policy guidance has been interpreted and expanded upon by each directorate in disciplinarily-appropriate fashion. Federal standards agencies may also wish to endorse suitable data and metadata standards that arise from research and library communities and informatics initiatives.

(4) This question is extraordinarily complex and difficult, and of course discipline-dependent. One relatively simple answer would be to track dataset reuse, and publications based on given datasets, plotting these data against cost data to decide about continued preservation. We also hope that federal agencies will continue to play an active role in funding research pertaining to long-term sustainable data standards and formats given their potential to reduce the costs of storage, facilitate discovery, and improve upon the interoperability of research data sets from heterogeneous sources.

(5) We believe that roles and responsibilities around data preservation and access are very much in flux, and that this very uncertainty is contributing to valuable research and innovation in both the public and private sectors. We therefore suggest that Federal policy mandate *ends, not precise means to those ends*, whenever possible. Here at the University of Wisconsin-Madison, researchers, librarians, the School of Library and Information Studies, and IT professionals are working together to raise consciousness of data-management issues and provide expert consultation and training in responsible data stewardship. We believe that our year-old Research Data Services (researchdata.wisc.edu/), while not a comprehensive solution to the broad panoply of data-management challenges, is a promising example for other stakeholders.

(6) The research and library communities frequently lament that research grants are of finite duration, while preservation responsibilities last indefinitely. Moreover, some researchers perceive preservation costs as subtracting from the pool of research funds available, and may oppose data preservation policies on that basis. Federal policy should therefore consider strategies for ensuring that preservation is

considered during the earliest stages of grant development. We encourage Federal agencies to:

- regularly and consistently fund national-level disciplinary data centers, existing and new;

- provide portable funding sources to "endow" preservation of and access to specific datasets;

- provide separate budget lines and guidelines to fund preservation and access, rather than lumping them in with overhead costs; and,

- authorize or mandate the engagement of data management professionals as part of the grant submission process. Clarify what is meant by "incremental" costs for data management and specific types of costs agencies are willing to fund (e.g., costs for storage, backup, consultations, metadata development, etc.). It is presumed that funding applied to data management services would enable institutions to grow their cyberinfrastructure and expertise, which in turn, would enhance a given institution's ability to assist PIs in their efforts to be responsible stewards of data generated in federally funded research.

(7) Federal data-management policies should insist that persistent, Web-compatible identifiers (such as DOIs, ARKs, PURLs, and handles) be provided to grant agencies for applicable datasets, much as the NIH Public Access Policy now insists upon PMCIDs/NIHMSIDs in grant reports and subsequent grant applications. Data archives should provide identifiers for embargoed datasets, and be willing to certify to Federal agencies that the dataset is indeed present in the archive. Grant agencies should develop policies that clearly articulate preferred repositories which will aid said agencies with respect to auditing and other compliance issues.

(8) Data registries help connect data creators with data users. Quite a few state and local governments have successfully stimulated dataset-based innovation by holding developer contests, as well.

(9) Dataset and author identifier-assignment and citation standards are under construction, notably the ORCID (orcid.org) and DataCite (datacite.org) efforts. Funding these standards, and insisting they be employed in communication with Federal agencies around grants, will help assure appropriate attribution and credit.

(10)   Almost any digital-data standards will be helpful! Presently, many disciplines utterly lack such standards; others have developed them, but not managed to implement them discipline-wide owing, in part, to lack of incentive or funding for researchers to use them. Federal attention to developing, promulgating, and insisting upon use of standards should be a clear priority! Discipline-specific standards developed in cooperation with (or by) researchers in those disciplines are more likely to gain adherents, thus building momentum around a given standard's adoption from funding agencies, publishers, and professional societies. In turn, widespread adoption of standards will clearly enhance our collective ability to

provide for the preservation, discovery, and reuse of research data within and across disciplines.

(11)   Standards sometimes arise from a widely-acknowledged need to share data, as happened with the International Virtual Observatory Alliance (ivoa.net); they also spring naturally from the establishment of discipline-dominant data repositories such as the Interuniversity Consortium for Political and Social Research (icpsr.umich.edu) and Long Term Ecological Research Networks (http://www.lternet.edu/). Should Federal policy jumpstart broader data sharing as well as more disciplinary-data repositories, standards development is likely to follow naturally. That said, Federal policy can help by providing funds for standards development, and one or more registries of relevant standards.

(12)   International standards coordination is the natural role of Federal standards bodies such as NISO and ANSI, as well as the Library of Congress.

(13)   As mentioned in our response to question 7, persistent dataset identifiers are a necessary prerequisite to citation. We do not believe Federal policy need endorse one identifier scheme over another; a list of acceptable identifier types will do. Citation of datasets from published papers is a somewhat harder problem, governed as it is by style guides firmly mired in the 20th century. We suggest instead that Federal policy require a set phrase with a list of dataset identifiers for papers published from Federally-funded research and datasets, much as is often done now for acknowledgement of Federal grants in published papers.