DUKE UNIVERSITY
LIBRARIES

January 10, 2012

White House Office of Science and Technology Policy
Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

RESPONSE from the Duke University Libraries.

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 regarding research data should start from the premise that more open is better, and then allow restrictions only as necessary to protect specific interests or provide specific incentives (and only if justified in the funding proposal). Open data can act a lever to maximize the investment made to create it, by allowing others to analyze it using methods other than those applied by the creator, and by allowing third parties (including entrepreneurial and commercial services) to combine it with data from other sources or layer innovate services on top. The classic example of this is data from the National Weather Service, which, because it is openly available, has provided the basis for an untold number of scientific and commercial projects, and created a whole new market for weather-related services that could not have been supported by the agency collecting the data on its own.

The best implementation approaches for such a policy would be those that take into account incentives that would encourage researchers and their institutions to preserve and share data, peer expectations being among the strongest of these for researchers. The data management plan requirement recently adopted by NSF is helpful in that it sets an expectation but doesn't require a specific implementation method, accounting for the variation in data types and practices across disciplines. Similar policies should be adopted by other federal funding agencies, encouraging broader access and preservation of research data to become an expectation in all disciplines.

However, setting policies will not be enough – it would be helpful for the federal government to stimulate the development of services that would make data sharing and

Box 90193 Duke University Durham, NC 27708-0193 Tel: 919-660-5800 Fax: 919-660-5923 library.duke.edu

preservation easier, and to foster standardization on a small set of generalized platforms and best practices to reduce the costs of managing research data.

(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?

Intellectual property issues around data are not well understood within the research community, and often barriers to access are put in place because of fear of potential misuse that could be allayed by the application of clear licenses that address specific concerns, like attribution. As part of their data policies, Federal agencies should recommend or require selecting from a specific set of data licenses that allow the openness that will promote innovation and scientific discovery while addressing legitimate concerns of the creators of the data and the agencies and home institutions that supported them.

Agencies should provide guidance to researchers at the proposal stage of a project on how to understand intellectual property issues related to their data, so that appropriate license selection and data management practices that take these into account can be implemented early on.

(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?

The policies and principles should be general, while the practices will need to be specific. Specific practices should emerge from specific communities, but providing common implementation platforms that can be the underpinning for variable practices would help.

Funding agencies should also be willing to provide funding to support data management expertise to be available locally at researchers' institutions (for example at their libraries) or through disciplinary repository services (such as NESCent Dryad) to assist researchers in applying data management approaches appropriate to their discipline. Examples of institutional services are the Distributed Data Curation Center at Purdue or the Scientific Data Consulting Group at the University of Virginia.

(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?

This is difficult to answer without some baseline data on what the relative costs and benefits are, and this baseline data will be difficult to come by unless comparable practices are put in place across different disciplines and outcomes are tracked over time. A good starting point might be to set a baseline allowable cost for data management and preservation (as a percentage of total project cost) for funding requests, and analyze after several rounds what approaches have been applied in different disciplines and how effective they are based on metrics like transaction costs for a third party to discover, retrieve, and make use of the data; verifiable integrity of the data at different year intervals, retrieval and use statistics, and so on.

If funding is provided to disciplinary repository services (such as Dryad, as mentioned above) they could be required to report on their methods and effectiveness, and comparing outcomes across different disciplinary repository services could be used as cost/benefit heuristics.

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

Researchers themselves should be able to keep their focus on what they know best – the subject and processes of their research. The organizations named in this question should provide support services (local and hands-on, where possible) to the researchers to facilitate best practices, find appropriate disciplinary standards and infrastructure, and implement approaches with a big-picture and long-term view in mind. These services should be made available at the beginning of projects (ideally, at the proposal stage) to facilitate best practices being put in place early and avoiding inefficiencies of retrofitting new practices mid way through or at the end of a project.

Currently, few organizations have the staff or infrastructure needed to provide this support. Federal agencies could provide funding and incentives to help build these support systems and encourage researchers to make use of them. The cost of developing these concentrated institutional support infrastructures will almost certainly be less than the distributed costs and inefficiencies of each researcher trying to figure out how to implement appropriate data management practices on their own, and likely doing it inconsistently or unsuccessfully.

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

It will be important to recognize that not all costs for data management, sharing, and preservation will be directly attributable to particular projects, and that as these expectations become more routine a larger proportion of the costs will need to be considered indirect costs. Data management and publishing and preservation services will become the equivalent of library stacks and services today, and will need to have a

basis for persistence beyond the life of any given project. While disciplines or projects with exceptional needs will be able to articulate clearly their specific data management needs and costs, the majority of research projects are not likely to be able to do so, and will need to rely on baseline services provided by their institutions or disciplinary organizations, with more general formulas for funding allocation.

(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?

As noted in the response to #4 and other questions above, develop some reporting metrics that can be used for early efforts to improve effectiveness of approaches to data stewardship, provide support for organizations to help researchers meet a baseline set of expectations, and only when these are in broadly in place and researchers have no excuse not to do the right thing, then become stricter regarding compliance and verification.

(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 encourage and support use of open data licenses and platforms that make it easier for researchers to share data in standard ways. If organized and described well and provided through documented APIs and open licenses, data may be combined and analyzed using new analytical tools, or used for purposes not envisioned by their original creator.

Agencies could support data hubs, providing a discovery and access service to data even if it is hosted in distributed disciplinary repositories. Such hubs could act not only as registries of available data and how to get it, but could also feature examples of innovative uses of the data, to stimulate others to envision similar or unique uses of available data.

(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?

Data citation standards (such as those being developed by the DataCite project) and researcher identifier standards (such as those being developed by ORCID) are important, to encourage consistency of practice and enable machine-actionable analysis. But widespread use of such standards will depend on community expectations. To encourage data citation norms to be adopted by disciplinary research communities, agencies should require disclosure of data sources (using common data citation and researcher identification standards) in grant proposals and reports, and should encourage authors to prominently display their data sources and data citations in their

publications. We need to reach the point where data citation has become an expectation similar to publication citation. Requiring particular data citation practices in places where requirements are possible will make the practices more visible and more likely to be adopted in places where they are not necessarily required.

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.

It's difficult to address individual standards because they will differ widely in different fields and require deep knowledge of practices in that field to be able to make reasonable recommendations. However, agencies should provide incentives and assistance to researchers to be aware of, choose, and use existing standards with broad community adoption rather than creating new ones. In proposals and reports, researchers could be required to justify what standard they have chosen, and agencies could encourage peer reviewers to look at these critically. Agencies should make sure to have experts who understand the commonly used standards for particular disciplines on review panels.

(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?

Some of the standards that form the basis of the Internet were created and are governed by collaborative processes through NGOs, with active participation from government agencies and the private sector. For example, groups like W3C, Apache, and Mozilla, have strong support from both public and private sector organizations and have developed open standards and systems that have formed the basis for the Internet economy. One of the key characteristics of their success is the commitment to openness, and that decisions are made based on consensus and ability to demonstrate technical merit and functional pragmatism, rather than the needs or business plans of any particular participant.

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

The key venue of decision making in research communities is not necessarily national boundaries but disciplinary communities. Agencies could make it possible for data experts (with a deep understanding of disciplinary needs) to attend disciplinary conferences, to seed and support standards discussions and possibly to staff standards development efforts. Agencies could look to how (and why) private sector companies

have supported development of open source software and open standards through groups like W3C, Apache, and Mozilla, and follow a similar model.

As they have with the Federal Agencies Digitization Guidelines Initiative http://www.digitizationguidelines.gov/, Federal agencies are well-placed to develop leadership roles in establishing best practices. While deeply engaged with disciplines, Federal agencies typically stand outside of them, giving the agencies the opportunity to provide a unique perspective.

Funding joint international demonstration or infrastructure development projects would also be helpful.

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

The answer to this is similar to the answer to #9 above – it's a combination of development and support for standards that meet this goal and encouraging broad adoption of the standard by embedding in in the publication, citation, and recognition norms of research communities. Ultimately, it will have to be something that will be easy for researchers to use - highly complex approaches may be technically superior but will likely not be used. The standard should be something researchers are already familiar with – the DataCite project is going with DOI, since these are already well understood in academia and actionable in many publication and discovery systems. Agencies could set expectations that authors should include with publications data citations and DOI links to supporting data (whether it is data they created or from another source) and could encourage publishers to make these links prominent in publishing systems and use the links in reporting and analysis.

A valuable analysis of many of this issue and many of the issues discussed above can be found in this blog: http://opencitations.wordpress.com/

These comments are submitted on behalf of the Duke University Libraries by:

Deborah Jakubs
Rita DiGiallonardo Holloway University Librarian & Vice Provost for Library Affairs

Kevin L. Smith
Director of Scholarly Communications

Paolo Mangiafico
Director of Digital Information Strategy