J. Willard Marriott Library
THE UNIVERSITY OF UTAH
295 South 1500 East Salt Lake City, Utah 84112-0860 (801) 581-8558

(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?

Experience so far indicates that to be effective, federal policies must mandate the data created in the course of federally funded research be deposited into publically accessible repositories. The National Science Foundation requirement of a data management plan is a laudable step toward awareness of the need to manage data, but a mandate will be required to create the critical mass of available data that will support rapid scientific innovation and encourage the commercial reuse of data that can underlie economic growth.

One aspect of the success of the NSF approach has been to set an expectation but not to require a specific method of implementation. Because of the variety of approaches and types of data across different disciplines, flexibility in compliance is called for, even within the context of a mandate. This flexibility can help the scientific community come to view data preservation and sharing as an issue of principle, necessary for good research and scientific accountability, rather than as merely a burdensome compliance issue.

A policy mandating data deposit will need to be accompanied by the development of standards and services that make data sharing economically feasible and data reuse as accessible as possible. Incentives are as important as requirements if the goal is to make usable data available for scientific verification and commercial reuse. By creating systems that are as simple, standardized and open to reuse as possible, the maximum potential of economic growth will be achieved. A useful metric for public access to data is whether someone, or some computer, can discover, access, interpret and use the data without having to contact the original data producer; such access is both economically beneficial and less burdensome to the data producers.

(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?

An important aspect regarding data is that under U.S. law, raw data is not subject to copyright and that rights even to protectable collections of data usually remain with the data producer, rather than being subject to transfer to publishers. So the rights issues are not as complex for data as they may be for publications.

The basic premise of federal policy should be that more openness is better, and restrictions should be applied only when genuinely necessary, for example, when the data makes it possible to identify a particular person involved in the research study. Basically the default, which is currently that data is managed locally (if at all) and idiosyncratically, should be changed to openness and standardization.

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

The key to convincing data producers, who are also the holders of whatever IP rights exist, to participate is to provide easy roads to compliance and incentives, usually in the form of norms and expectations within their disciplinary communities, to comply. Other incentives involving credit for their work, easy citation methods, and ability to demonstrate the impact of their work by using metrics such as the number of times their data set has been cited are important incentives and need to be built into any plan for compliance.

The federal government could assist in making data preservation and sharing as seamless as possible by supporting the creation of successful data management systems that resemble currently successful programs such as those at the National Library of Medicine's National Center for Biotechnology Information. There are four different methods for submitting data into GenBank, along with a number of other tools to make this process easy. Working with publishers in this process to deposit and store data may initially seem the obvious path, but publishers are in a business to make money. Publishers do not have the commitment to preservation or open access that other stakeholders, such as libraries, have already demonstrated in their work with books, journals, video, audio, maps, microfiche and other rare materials.

There is a reasonable argument for embargoes, in some cases, based on the unique effort exerted by the data producers or original scientific research team. Although effort alone cannot justify copyright protection (based on the Supreme Court's 1991 decision in *Feist Publications v. Rural Telephone Service Co.*), the need to protect data for some short period of time while the team or lab completes its own analysis could be respected by allowing a fixed-term period of exclusive access. Such an arrangement, however, does not preclude the deposit of the data into a certified repository even during that embargo period, particularly so that archiving activities can be begun.

(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 can take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data by adopting a relatively general mandate for data sharing while requiring more specificity for the practices within each discipline. This is because there is ample evidence that different scientific disciplines present a variety of requirements for the management of data. Data sharing policies should be viewed as flexible requirements that remain open to modification as problems arise or best practices emerge from within specific communities of scientific practice.

Some baseline conditions or requirements, especially related to archiving and preservation, can be applied across the board. This is a vital place to begin, since many scientific disciplines have focused on access or discovery rather than preservation, yet the latter is key to fostering efficiency and innovative reuse.

In some disciplines, a funder requirement will serve as a first step toward creating awareness of the fundamental need for data management, preservation, and access. We have seen this take place among working scientists as awareness of the NSF data management plan requirement has spread, and further mandates will facilitate this awareness.

Funding agencies should be willing to provide funding for data management expertise that is available locally at researchers' institutions, and/or through disciplinary repository services (such as the DRYAD repository at the National Evolutionary Synthesis Center). Such support will assist researchers in applying data management approaches that are appropriate to their specific disciplines.

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

(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?

In general it would be difficult for agency policies to consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research because of the particularized and ad hoc nature of so many approaches to research data up until now. It may be most useful to think in terms of baseline services that should be supported across all disciplines (i.e. archiving) and more particularized secondary services (such as specialized query capabilities). Agencies might consider the relative emphasis that is appropriate in the area of research that that agency funds, and what areas are appropriate for local institutions to assume responsibility for. Thus an agency might provide seed funding to institutions for preservation, but recognize the need for ongoing funding to a scientific community to develop secondary services.

A potential technique to establish a baseline cost would be to set an allowable cost for data management for funding requests, then analyze, after several rounds, what approaches have been applied and how effective they have been based on metrics such as use statistics, the verifiable integrity of the data over time, and third-party costs to discover, retrieve and use the data. If funding is provided to disciplinary repositories, reports based on these metrics should be required.

The benefits of shared data will also be difficult to measure, but they are nonetheless real. Accountability and the ability to verify scientific results are vital, but hard to quantify. Other benefits, such as the support provided for reuse by different teams of researchers or by commercial enterprises, will be easier to track. The opportunities for innovation and commercial exploitation of shared data will be evidenced by increased growth within a sector of the economy.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?
Stakeholders can best contribute to the implementation of data management plans by having access to or creating easy and efficient methods for themselves and others in their field for the storage, description, and sharing of the data they collect.

It is important to keep researchers focused on research; this is vital if a data sharing requirement is going to support innovation and growth and not hinder it. The stakeholders named thus have the important role of providing the services, standards, best practices and infrastructure that make data sharing simple and efficient. Insofar as agencies can provide funding and other incentives to support those functions, they will contribute to the implementation of data management plans.

The best approach is to build on existing infrastructures and practices, learning from what works well while being sensitive to disciplinary differences and the evolution of scientific disciplines over time.

While successful practices should be the model for policy implementation, it is important that success be demonstrated and not merely asserted. Each agency, as part of its data sharing mandate, should identify metrics that are important within that field by which the success of a plan or services can be measured. Those metrics will evolve over time, but with a clearly articulated set of requirements it will be possible to identify how various stakeholders can contribute to the successful implementation of data management plans.

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

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

Funding mechanisms can be improved to better address the real costs of preserving and making digital data accessible by acknowledging and communicating that the real costs of preserving and sharing digital data are indeed legitimate and important costs of the overall research enterprise.

It is important to recognize that not all costs associated with good data management will be directly attributable to specific projects. As data management expectations become more widespread and routine, an increasing proportion of the costs will need to be considered indirect costs. While some disciplines or projects may present exceptional needs, many other research projects will likely rely on baseline services provided by institutions or disciplinary groups that need more general formulas for funding.

(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?

There are several approaches agencies can take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research. For example, reporting metrics should be developed and applied to early efforts at improving data stewardship, and the results shared broadly. As best practices emerge and community norms support good data management, researchers will have an incentive to preserve and share their data.

In addition, compliance should be verified through systematic approaches, which can be much easier and efficient for the agency and less punitive for researchers. Most researchers pay special attention to two milestone events in the research process – the grant proposal and publication. Policies and metrics that are embedded at these points will get the attention of researchers and make compliance more likely.

Finally, agencies should develop guidelines for those who review both grant proposals and final reports to the funding agency that highlight what to look for in a well-developed data management plan within the specific discipline.

(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?

There are several additional steps agencies can take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy. First, the use of open data licenses and platforms that facilitate sharing in standardized ways will make it easier for other researchers and industries to reuse data and increase the return on investment for funded research projects.

Second, support for well-documented APIs that allow individuals and machines to develop new capabilities and services is key to fostering innovation. One of the benefits of the broadest possible access and opportunity for reuse is that federal agencies could help build on "citizen science" efforts, which have up until now largely focused on data gathering and classification. Open licensing and usable APIs will ensure that the maximum number of creative imaginations are looking for innovative ways to use research data.

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

(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?

There are several mechanisms that can be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported. One such mechanism is to support ongoing efforts to develop data citation standards (such as the DataCite project) and author and institutional identifiers (such as those being developed by ORCID).

Another mechanism would be to require agencies to disclose data sources using common data citation and researcher identification standards in order to build community norms that reward good attribution practice, as is the case for research articles.

Nevertheless, it should be recognized that existing attribution standards for published articles will not translate seamlessly into the world of research data, especially given the importance of machine-based access and reuse. As in so many other areas, this is a case where standards will have to develop as reuse and innovation grows, and agency mandates should remain flexible while publicizing and encouraging best practices.

(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.

There are an astounding number of data standards available depending on the data in question and the subject area. The MIAME example is typical of the development of data standards among researchers in a particular area. A quick search in Google Scholar displays several other articles on a similar theme:
Taylor, et al., 2007, The minimum information about a proteomics experiment (MIAPE). Nature Biotechnology 25, 887.
Bustin, et al., 2009, The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clinical Chemistry. 55(4) 611.
Novere, et al., 2005, Minimum information requested in the annotation of biochemical models (MIRIAM). Nature Biotechnology. 23(12) 1509.
Field, et al., 2008, The minimum information about a genome sequence (MIGS) specification, Nature Biotechnology 26, 541

(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?

Another example involves the development of Common Data Elements, which are jointly created by users of the caBIG database, hosted by the National Cancer Institute (NCI). There are 20 different Common Data Elements in use and NCI is accepting others for review:
https://cabig.nci.nih.gov/workspaces/VCDE/Data_Standards/

One possible element that has lead to the success of many of these efforts to create data standards is the creation of the standards by those working in the field working with the data and attempting to share or use data created by others.

Some of the best examples of data management, preservation, access, and use can be seen at the National Library of Medicine's (NLM) National Center for Biotechnology Information (NCBI).

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.

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

Most research projects are completed in teams across institutions and a portion of those teams are international. These teams face the need to create standards for sharing data at earlier points in the process and are often poised to make the recommendations for data standards in their field of study. The fact that the articles listed in Question 10 are authored by international teams is a good indication of a process that works well. Even GenBank from NCBI is an international effort of data sharing, preservation, and storage.

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

The National Library of Medicine provides an excellent model of linking between publications in PubMed and data located in a whole host of databases and then back to the publications with single click capabilities.

With gratitude and recognition to Kevin L. Smith for many of his ideas expressed in this document.