UNIVERSITY OF MICHIGAN
Library

Paul N. Courant
University Librarian and Dean of Libraries
Harold T. Shapiro Collegiate Professor of Public Policy
Arthur F. Thurnau Professor
Professor of Economics and of Information

818 Harlan Hatcher Graduate Library South
Ann Arbor, Michigan 48109-1205
734 764-9356 pnc@umich.edu

11 January 2012

Office of Science and Technology Policy on behalf of
National Science and Technology Council
Attention: Ted Wackler, Deputy Chief of Staff
digitaldata@ostp.gov

Re: Response to Notice for Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research (FR Doc. 2011-28621)

Dear Mr. Wackler:

Since 1838 the University of Michigan Library has been serving the research needs of students, faculty and the public. Over its many years of operation the library system has acquired an enormous wealth of diverse resources and continues to be a springboard for research, invention, and learning. Today, the lifeblood of much research and inquiry is data. Scholars, inventors, economists, medical researchers, social scientists, astronomers – all disciplines - look to data to find patterns, make predictions, identify stories of the past and present that may help us make a better future.

American taxpayers invest a tremendous amount in research, reflecting our national commitment to education and fundamental research. The resulting data are paid for by taxpayers and should be made available for further inquiry (along with publications produced as the result of that research). The success of the NIH mandate and PubMed Central as a free, publicly accessible, reliable source for NIH-funded research provides an important practical and philosophical model for making data produced by taxpayer-funded research broadly available.

My response to the questions raised in the Invitation to Comment follows. Thank you for the 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?

A policy that ensures the ability of scientists, federal agencies, and the public to make use of data produced by federally funded research will foster economic growth and, by enabling other scientists to re-use that data, will improve our scientific productivity and increase innovation.

Data sharing reduces redundancy and eliminates wasteful uses of federal funding, since lack of access to scientific data makes it impossible for scientists and funders to answer important questions without needless duplication of research. It will help to ensure the scientific integrity of federally funded research by enabling the verification of published results and by exposing errors when they occur. It will improve sponsors' ability to allocate funding efficiently by providing better metrics for the measurement of scientific influence and progress.

Finally, funders will gain an enhanced ability to demonstrate the positive impact and value of work they support, helping to ensure the continued availability of federal funding for scientific research.

From the perspective of scientists and their institutions, sharing of scientific data has the potential to promote the development of broad-based metrics for measuring scientific influence. This will improve the current credentialing process, which relies almost exclusively on formal publication and its attendant metrics (such as a journal's "impact factor"), freeing institutions to promote broader means of disseminating the knowledge they create.

A key requirement of a public access policy that functions as a driver of the broader economy is to make sure data are openly licensed in a way that permits widespread reuse, including commercial as well as non-commercial uses. These open licenses could address possible integrity questions for compilations or other bodies of information. That said, data itself are in the public domain and not subject to copyright at all under US law. It is critically important that data be available in a manner sufficiently unencumbered to allow for innovative uses, reuses, and recombinations permitting new insights. This way, the public can benefit from the research that it funds. To be used, data also need to be discoverable, whether by a person or a machine, without the need to contact the original data producer.

Because scientists need strong incentives to share, the policy should encourage the use of appropriate licenses and discoverability-supporting standards through detailed guidance provided to them. Awards must be contingent on verifying that data from past projects are available and discoverable. Finally, and to the greatest extent possible, the policy should be implemented in a consistent way across all funding agencies.

(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?

Data licenses must be crafted to protect the rights of scientists who create the work and the rights and interests of the agencies and taxpayers who fund that work. This must be the first priority of any policy.

The policy should assure that patent rights, the right to enter into publishing contracts, and all other intellectual property rights are retained by researchers. It should also assure that the chosen licensing scheme does not limit unnecessarily the ability of other parties to make use of the data.

(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?

Current data sharing practices are diverse, making a single, mandated approach to data management difficult. Funders should encourage scientists in each discipline to develop their community's standards and norms. They should also set benchmarks to ensure the preservation and availability of data and to guide scientists toward consensus in disciplines where it has not yet been achieved.

One of the biggest obstacles scientists currently face when contemplating how to share their data is a lack of specific funder guidance. To enable scientists to plan adequately for data preservation and sharing there must be clear guidance provided on certain fundamental aspects of data management. Federal agency guidelines might include:

- Definition of research data
- Data sharing and access policies (including preferred timelines for sharing relative to the time of publication and the conclusion of the award)

- Minimum data retention periods
- Preferred disciplinary repositories
- Preferred file formats for specific types of data
- Preferred metadata standards
- Preferred access mechanisms (modes of data delivery) and licensing schemes
- Admissible exceptions to data sharing requirements (for e.g. privacy or security reasons)

As described under question 4 below, the creation of domain-specific data repositories where none currently exist would go a long way toward ensuring the long-term availability of valuable data, and would even out some of the differences in data sharing and archiving support for various disciplines.

(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?

While it would be ideal if we could consistently and reliably identify data for which the benefits of preservation exceed the costs, in reality it is difficult to know in advance what the value of a scientific dataset might be several years from now. And although we can make assumptions about the long-term costs of data preservation, we cannot know for sure what they will be. So both sides of the cost-benefit equation are unknown.

However, at present the greatest barriers to data preservation and sharing have little to do with our ability to make this determination, so creating the right incentives and support structure to facilitate the sharing of research data must be our first priority.

Facilitating the creation of disciplinary data repositories in areas where they are needed is one way to help achieve this. Another is to make sure that scientists are given enough guidance to be able to identify the communities for which their own data are potentially relevant and to budget realistically for data preservation and sharing. Identifying at-risk datasets seems particularly urgent in light of the difficulties often faced even by major research collaborations in assuring continued access to experimental data that are unique and non-repeatable (see "Data Preservation at LEP" for one interesting case study). Clear guidance from funders will help to ensure that scientists identify communities of interest and budget appropriately for long-term preservation and access.

"Data Preservation at LEP." André G. Holzner, Ryszard Gokieli, Peter Igo-Kemenes, Marcello Maggi, Luca Malgeri, Salvatore Mele, Luc Pape, David Plane, Matthias Schröder, Ulrich Schwickerath, Roberto Tenchini, Jan Timmermans. arXiv:0912.1803v1 [hep-ex].

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

A successful approach to data management requires cooperation among many stakeholders, but here we will limit ourselves to discussing the role of libraries and scientific publishers. Libraries can help fill gaps in the data preservation infrastructure by preparing and archiving data that doesn't have a home elsewhere. We can connect scientists with existing services and in some cases we can manage data locally. We can serve an advisory role, recommending best practices when it comes to basic digital preservation strategies, but we will probably not have the domain knowledge needed to curate every dataset.

Scientific publishers have made significant contributions to the cause of scientific data sharing by requiring authors to make supporting data available, by crafting advice on data archiving options, and by making efforts to link datasets with the published literature. These efforts should be encouraged, but with the understanding that scientists cannot rely on them exclusively for data hosting and dissemination via supplemental materials published in traditional peer-reviewed journals. This would already be problematic simply by virtue of the fact

that most journals are only available to subscribers, rendering these datasets off-limits to most taxpayers. But through the materials made publicly available by the joint NISO/NFAIS working group on journal supplemental materials, it has become apparent that many scientific publishers see data as a liability, not an asset. The group is concerned first and foremost with limiting the scope of what can be considered a supplemental material, a measure aimed at reigning in costs associated with data hosting and review of submitted materials. The publishers in this group have stated explicitly that raw datasets fall outside of their purview, and so cannot be relied upon to curate original data even in the short term. They have also stated that they intend to manage supplemental materials under the same rights regime they commonly apply to published articles, i.e. exclusive copyright is to be signed over to the publisher.

As a result, while we should encourage and support all scientific publishers who are willing to open their archives to the general public to do so, such archives cannot and should not be the only archives or long-term stewards of research data. Those whose mission is to act on behalf of the public interest, be they Federal agencies or universities and colleges acting on their behalf, must have direct involvement in every aspect of assuring public access to data. Publishers currently have few incentives for making data broadly accessible, so creating those incentives and ensuring shared and co-equal ownership, storage, and responsibility for access is essential in any public-private partnership.

NISO/NFAIS Supplemental Journal Article Materials Project. http://www.niso.org/workrooms/supplemental

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

The single most important step funders can take to improve scientists' ability to budget for data preservation and access is to assure that agencies set minimum data retention periods (per above, these may differ by discipline, but should be made as consistent as possible across all agencies). Doing so will allow scientists to plan for how long they will need to store their data after the conclusion of the award, and thus estimate the costs of doing so. Funders should also encourage the inclusion in data management plans of an explicit data sharing timetable describing when data will be prepared, deposited, and their availability verified before the end of the award, as well as the use of publicly accessible data repositories whose hosting costs are known in advance.

One way to address the considerable costs in time and effort of preparing data for archiving is to compensate scientists for this labor by providing better incentives for data sharing. In parallel, we must lower the cost of doing this work, a goal that can be achieved in a number of ways. Since paid data curators take much of the burden of preparation off scientists' hands, facilitating the existence of well-funded disciplinary data repositories for most disciplines will lower the costs of data preparation and hosting. Further, encouraging scientists to make good data management practices an integral part of their research process will reduce the burden of data preparation and description at the conclusion of the project. This can be done through specific guidance and through more systematic approaches to verifying compliance at key points during the course of the project.

(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?

Implicit in any discussion of a policy or mandate is the need to provide incentives for compliance. Currently, some agencies and directorates have made an explicit commitment to monitor compliance with data sharing policies through annual and final reports, while others have left it an open question as to how or whether compliance will be monitored or acted upon. As with the creation of data management guidelines, mechanisms for verifying compliance don't need to be uniform across all disciplines but they do need to be implemented consistently for all scientists.

However, the best way to ensure that data are shared is to give scientists positive incentives to share. Scientists value data sharing (Tenopir et al., 2011), yet many do not share data because doing so requires significant work and offers little reward. To make data sharing worth their while they must be able to demonstrate the value of their data to funders and to their fellow scientists. Although technologies and standards to support data attribution exist, they are not widely implemented and don't yet carry the weight that publication does. Thus, mandates alone will not create the incentives needed to foster a culture of data sharing; it must be valued sufficiently by the funding agencies and the scientific community at large to justify the effort required. Increased citation to publications as a result of data sharing has been demonstrated, but by itself it's not enough to incentivize scientists; they must receive credit directly. One key to achieving this goal is the development and use of new scientific performance metrics ("altmetrics") that take dataset citation and usage into account.

In many science disciplines the traditionally inseparable communication and credentialing functions of scholarly journals have long since become decoupled, with the communication function shifting to more timely venues such as the preprint archive (Gentil-Beccot et al., 2009). But the continued reliance of the academic credentialing process on formal publication and its attendant metrics has left scientists with little incentive to share data and other research products. A more inclusive and accurate way of measuring scientific progress would benefit all stakeholders, bringing the credentialing function of publication back in line with the communication function. Scientists should be rewarded for doing more and better science, which means sharing data as well as publishing papers. Not only will this benefit science as a whole (through better verification of results and a reduction of duplicated research), it will also improve sponsors' ability to allocate funding efficiently. By the same token, funders will gain an enhanced ability to demonstrate the positive impact and value of the work they fund, helping to ensure the future availability of public funding for research.

To lower the burden of verification and compliance a systematic approach to data management is key. Rather than making data available by request via a wide array of idiosyncratic means, researchers should be encouraged to make use of standard technical components including managed repositories, persistent identifiers, and standard licenses attached to datasets. Metadata could include a reference to a grant number or other means of identifying the sponsor, and could be made available through a machine-readable interface. Such mechanisms would facilitate the automation of compliance and verification tasks, thus reducing costs to sponsors and researchers. Making metadata available and machine-readable will also facilitate the growth of altmetrics and new business models centered around them.

"Sharing Data: Practices, Barriers, and Incentives." Carol Tenopir, Carole L. Palmer, Priyanki Sinha, Jeffrey van der Hoeven, and Jim Malone. *ASIST 2011*, October 9-13, 2011, New Orleans, LA, USA. http://www.asis.org/asist2011/proceedings/submissions/26_FINAL_SUBMISSION.doc

Altmetrics. http://altmetrics.org/

"Citing and Reading Behaviours in High-Energy Physics. How a Community Stopped Worrying about Journals and Learned to Love Repositories." Anne Gentil-Beccot, Salvatore Mele, Travis Brooks. arXiv:0906.5418v2 [cs.DL]

Thank you for your consideration.

Sincerely,

Paul N. Courant