Thu 1/12/2012 11:50 PM
Response to RFI

Michael Carroll; mcarroll@wcl.american.edu

Professor of Law and Executive Director; Program on Information Justice & Intellectual Property, American University, Washington College of Law

Meredith Jacob; mjacob@wcl.american.edu

Assistant Director; Program on Information Justice & Intellectual Property, American University, Washington College of Law

We appreciate the opportunity to submit the following comments to the National Science and Technology Council's Interagency Working Group on Digital Data in connection with the Request for Information on Public Access to Digital Data Resulting from Federally Funded Scientific Research published in the Federal Register on November 4, 2011.

(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?

The question implies that changes to federal data sharing policies would be needed to improve access and reuse of data produced or collected from federally-supported scientific research. This implication is correct. As a general matter, existing federal data policy is uncoordinated, underspecified, and, frankly, incoherent. Notwithstanding the laudable goals articulated in the America Competes Act and in OMB Circular A-130, data produced or collected with federal support is subject to a range of possible rules regarding public access and terms of reuse.

We conducted a review of publicly accessible policies from agencies supporting scientific research to ascertain what, if any, data sharing requirements recipients of federal funds agreed to with respect to non-classified research.

1. What mandates are imposed by federal law through statute, regulation or policy on recipients of federal funds to provide public access to scientific data generated by federally funded research? (Both intramural research and grant or contract funded research will be examined.)

2. In the absence of a federal mandate for data sharing, what efforts do agencies take to promote or provide public access to data produced or collected through federally funded research?

3. What, if any, restrictions or requirements do agencies place on recipients of federal funds who make their research data public to use technological protections or contractual terms of use that limit reuse, reanalysis or redistribution of such data?

In sum, our results show that there are very few federal mandates requiring data-sharing, that agencies by and large have adopted an ad hoc approach to promoting data-sharing by federally-funded researchers, and that no policy that we could find recognized or addressed the common practice among federally-funded researchers to impose terms of use on data made public over the Internet without any federal input into the presence or substance of these terms of use.

We recommend that federal data sharing policy should be consistent across all agencies so that data availability is useful and predictable. This policy needs to be clearly set out, easily accessible online, and consistently enforced.

Federal funding for scientific research comes through a number of routes: through the direct employment of researchers; through grants to researchers at not-for-profit institutions; and through partnership and contract agreements with for-profit corporations. Whenever possible, the data that results from this federally funded research should be made available to the scientific community and to the public.

Ideally, data produced by employees, grantees, or contractors could be made available online to the public in a searchable, standardized form without any artificial barriers or limitations on reuse. Unfortunately, many Federal agencies only meet a minimum standard, complying with OMB Circular A-110 that requires that any data that were used for rulemaking be made available in response to a FOIA request. Additionally, A-110 gives the Federal Government right to "(1)obtain, reproduce, publish, or otherwise use the data first produced under an award; and (2) authorize others to receive, reproduce, publish, or otherwise use such data for Federal purposes."

Notice that this policy does not prevent the recipient of federal funds from imposing reuse limitations on other users via contract.

In practice, two agencies, the National Institutes of Health and the National Science Foundation impose the broadest requirements on grantee data-sharing. In other agencies, specific projects or institutes will impose data-sharing requirements on grantees, such as the Genomics:GTL project within the Department of Energy. Finally, many agencies, such as the Environmental Protection Agency and NASA have invested significant resources into making data publicly available, but neither has a policy of mandatory access to data generated by grantees. Data policy and location should be standardized across agencies so that researchers, policymakers, and lay people can local data and rely on its continued availability.

With rare exception, federal data policy defers to investigator preference as to whether data will be shared with the public. We recommend that this policy be revised because investigators may have a conflict of interest with the public interest and engage in competitive withholding for personal gain or may undervalue the reuse potential of research data by researchers in other disciplines.

While agencies may need to establish limited criteria to opt-out of data sharing, these criteria should be specific and should not simply rely on researcher choice. Federal policy on access to digital data should not use the NSF model that relies solely on the discretion of the researcher to determine whether data is made publicly accessible.

(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?**

Some care should be given to delineate the "intellectual property interests" referenced in the question. Copyright does not apply to factual data that are arranged in an unoriginal manner. Most data themselves are not patentable inventions. Data can be treated as a trade secret, but there is circularity in this determination. Information can only be a trade secret if it is not "readily ascertainable", and it is a matter of federal policy as to whether data should be made readily ascertainable. As a consequence, there is only an "intellectual property interest" to be protected if federal policy is that there should be such a private interest rather than a right of public access.

The issue that needs to be addressed is to what extent researchers may hobble or encumber access to data arising from federal funds through contractual restrictions or technological protection measures. Neither of these is an "intellectual property interest," but each can be an effective means for undermining the public's interest in data access and data sharing.

However, even when researchers use private databases, policies can protect public access to the results. In one example from the National Center for Environmental Economics:

> **(d) Data Plan (if applicable).** Provide a Data Plan (2 single spaced page limit) to make available to the public all data generated from observations, analyses, or model development (primary data) collected under an agreement awarded as a result of this RFP. The plan should describe how the applicant plans to make all data resulting from an agreement under this RFP available in a format and with documentation/metadata such that they may be used by others in the scientific community. This includes both primary and secondary or existing data, i.e., from observations, analyses, or model development collected or used under the agreement. *Applicants who plan to develop or enhance databases containing proprietary or restricted information must provide, within the two pages, a strategy to make the data widely available, while protecting privacy or property rights.* (emphasis added) National Center for Environmental Economics, Grant Solicitations *available at* http://yosemite.epa.gov/ee/epa/eed.nsf/pages/GrantSolicitations.html#bmk50

This example illustrates options that could be included in other guidelines to acknowledge the interaction between public and private databases, and the need to allow contribution to the private database while protecting public access.

(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 in digital data across disciplines by allowing data deposit in discipline-specific repositories while also requiring metadata necessary for indexing and search to be submitted and maintained in a central database that makes it possible to local all digital data resulting from federally funded data from a single central search.

Discipline-specific repositories can thrive with the support of a dedicated research community, but we should not lose the value in making that data also locatable and useable by the larger scientific community and the public.

(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?

The presumption should always be in favor of long-term stewardship and dissemination because the future utilization cannot be anticipated. Federal data access and preservation policies should protect the reuse of digital data for both researchers outside the original field and for future researchers even when the potential for reuse is not obvious within the field. Digital data is a resource which can be an input into a range of innovative activities, and it would be unwise to assume that we can predict the value of data as technological capacities evolve and as public access to research increases.

Discipline -specific repositories lack a central directory or access point for lay members of the public, or for researchers outside the field. Data.gov. or another central portal should provide a central search to locate data sets, even if they are deposited in discrete repositories. This would allow specialized repositories if necessary to adapt to the needs of a specific scientific community, while still insuring broad public access for novel or crosscutting research.

Currently, though large amounts of digital data are available online, there is no system for determining either data sharing policies, or data repository location, across agencies. Even within agencies, such as discussed below for the Department of Health and Human Services, data sets are scattered across agency websites, without a central index

If the public funds the cost of data collection and storage, then it is imperative that we have an efficient central index for locating these data sets across repositories so that they get the most possible use. Clear policy and a functional central search index helps the research community and the public get the highest possible value for the effort of data collection and preservation.

Finally, long term data preservation allows for the measurement of change over time, even when that was not the initial intent of the data collection. One example of this is Library of Congress-led project on the preservation of historical geospatial data that was initially intended for mapping and geologic studies, but can be used for other environmental, economic, and social research when preserved over a longer period.

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

Research communities can contribute to standards and best practices that allow collection, standardization, and deposit of high quality data.

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

The preservation, deposit, and hosting of digital data should be addressed throughout the research funding process. Funds should be allocated in proposals for data collection and management. Clear criteria should be given to reviewers for the evaluation of funding proposals. Finally, completion of data deposit with a public repository should be an enforceable requirement of Federal grants.

(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?

The burdens of compliance are actually reduced as the policy of requiring open access to digital data is standardized.

A key step to encourage compliance with Federal data stewardship and access policies would be to build in a focus on data access throughout the grant process. For the preservation and deposit of digital data to thrive, it must be seen as a core deliverable of a grant or contract. This focus on data preservation and data sharing should begin at the grant review phase, where clear guidelines should be given on how to evaluate data management plans. It should be followed by an approach, such as the one currently in place at National Institutes of Health (NIH), that views failure to implement the data management plan as grounds for enforcement, and as a barrier to future grants.

While agencies may need to established limited criteria to opt-out of data sharing, those criteria should be specific and should not simply rely on researcher choice, and the interests of researchers and the public are not completely aligned. Researcher may not be able to see the applicability of data sharing to fields other than their own and also may have self interest in delaying the publication by competitors in the field.

If this became a standard part of scientific research, systems could be developed to reduce inefficiency and automate the deposit of data in open repositories. The more data deposit is standardized and automatic, the lower the cost of enforcement 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**?

Data made available through data sharing should not contain any contractual preclusion on reuse. Repositories used for public access should not contain any terms or conditions that limit the free reuse of 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?

The Federal government should support initiatives to develop standards for data citation and data attribution.

Summary

The following principles should guide Federal data policy:

- Federal data sharing policy should be consistent across all executive branch agencies. Consistency across agencies is valuable for researchers and the public so that data availability is useful and predictable. Specific policies can be implemented at the agency or project if need be.
- Federal data policy and any agency or project level modifications should be clearly available online and specifically set out the location of data indices and repositories.
- A central index of all data or data repositories should be established, i.e. data.gov
- Federal data policy should require data sharing as the default for all federally funded research. While agencies can establish criteria to opt out of data sharing, these criteria should be publicly available as part of the data sharing policy. Researcher election alone should be insufficient criteria to opt out of data sharing.
- Data sharing guidelines should be built into the research grant process from proposal evaluation to completion of the grant. Data sharing should be seen as a enforceable requirement of the grant.
- Federal data sharing policy should recognize the value of research outside the original field, as well as the high potential for future, unanticipated use of data.

- Federal contracts should require the same data sharing policies as grants to non-profit institutions, unless they fall within criteria established by the contracting agency.
- Data made available through data sharing should not contain any contractual preclusion on reuse.

Best regards,

Michael W. Carroll
Professor of Law and Director,
Program on Information Justice and Intellectual Property
American University, Washington College of Law
4801 Massachusetts Ave., N.W.
Washington, D.C. 20016
vcard: http://www.wcl.american.edu/faculty/mcarroll/vcard.vcf

Research papers: http://works.bepress.com/michael_carroll/
http://ssrn.com/author=330326
blog: http://www.carrollogos.org/
See also www.creativecommons.org