Association of College & Research Libraries
50 E. Huron St. Chicago, IL 60611
800-545-2433, ext. 2523
acrl@ala.org, http://www.acrl.org

TO: Office of Science and Technology Policy digitaldata@ostp.gov
FROM: Association of College and Research Libraries
RE: Recommendations on Public Access to Digital Data Resulting from Federally Funded Scientific Research
DATE: Thursday, January 12, 2012

The Association of College and Research Libraries (ACRL) writes in response to the request for information issued November 10, 2011, by the Office of Science and Technology Policy (OSTP) regarding recommendations on approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally funded scientific research.

ACRL, a division of the American Library Association, is a nonprofit professional organization representing more than 12,000 academic and research librarians and interested individuals. ACRL is the only individual membership organization in North America that develops programs, products and services to meet the unique needs of academic and research librarians. Our initiatives enhance the ability of academic library and information professionals to serve the information needs of the higher education community and to improving learning, teaching and research. ACRL publishes scholarly, peer-reviewed journals in the field of library and information science.

ACRL appreciates the opportunity to comment on increasing public access to digital data that result from federally funded scientific research. Many of our individual members and their libraries will also submit detailed comments to OSTP. ACRL has long believed that ensuring public access to the fruits of federally funded research is a logical, feasible and widely beneficial goal. We have endorsed "The Federal Research Public Access Act of 2009" (S. 1373) noting, "It reflects ALA policy regarding access to federal government information by providing for the long-term preservation of, and no-fee public access to, government-sponsored, taxpayer funded, published research findings."

ACRL offers comments on the first nine questions posed in the request for information, as we would most like to express our position regarding policy for preservation, discoverability and access. We are refraining from addressing the last four questions on standards for interoperability, reuse and repurposing, which seek specific examples of good data management. Our comments to the specific questions follow:

*(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 most effective federal policy to encourage public access to, and preservation of, research data would be to require that data generated in the course of federally-funded research be deposited into publically accessible repositories. The National Science Foundation (NSF) 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.

- The NSF approach, which has been to set an expectation but not to require a specific method of implementation, is the right one. 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?*

- It is important to recognize that raw data is not subject to copyright under U.S. law, 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 manage locally (if at all) and idiosyncratically, should be changed to openness and standardization.
- Theykey 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.
- The federal government could assist in making data preservation and sharing as seamless as possible by working with publishers and other stakeholders to reduce the burden of handling “supplemental materials” and make it easier to integrate data into their publishing platforms and access systems.
- 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 begin.

*(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?*

- Since there is ample evidence that different scientific disciplines present a variety of requirements for the management of data, the policies and principles that underlie a data sharing mandate should be relatively general, while the practices within each discipline will need to be specific. 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. 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.

*(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 particularized and ad hoc nature of so many approaches to research data until now makes it difficult to assess relative costs and benefits for different disciplines. 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?*

- 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.
- 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.

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

- Perhaps the most important step would be to acknowledge and communicate 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?*

- 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.
- 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.
- Agencies should develop guidelines for those who review grant proposals 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?*

- 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.
- 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.

*(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?*

- Support should be given to ongoing efforts to develop data citation standards (such as the DataCite project) and author and institutional identifiers (such as those being developed by ORCID).
- Agencies should require disclosure of 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.