Response to Request for Information: "Public Access to Digital Data Resulting from Federally Funded Research," November 2011 January 12, 2012

G. Sayeed Choudhury
sayeed@jhu.edu<mailto:sayeed@jhu.edu>
Johns Hopkins University Libraries
Baltimore, MD

Prudence S. Adler
prue@arl.org<mailto:prue@arl.org>
Association ofResearch Libraries
Washington, DC

Heather Joseph
heather@arl.org<mailto:heather@arl.org>
SPARC
Washington, DC

Summary
Thank you for the opportunity to comment on "Public Access to Digital Data Resulting from Federally Funded Research." These comments are submitted on behalf of the Johns Hopkins University Libraries, the Association of Research Libraries (ARL), and the Scholarly Publishing and Academic Resources Coalition (SPARC). The Johns Hopkins University Libraries have established a leadership role with digital data management through a long-term program of R&D, prototyping and implementation of data infrastructure. ARL is an Association of 126 research libraries in North America. These libraries directly serve 4.6 million students and faculty and spend $1.4 billion annually on acquiring information resources, of which 62% is invested in access to electronic resources. SPARC is an international alliance of academic and research libraries. Action by SPARC in collaboration with stakeholders – including authors, publishers, and libraries – builds on the unprecedented opportunities created by the networked digital environment to advance the conduct of scholarship.
Question 1) Whatspecific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientificresearch, to grow the U.S. economy and improve the productivity of the American scientific enterprise?
Comment 1) The most effective Federal policies in thisregard would mandate digital data deposit into publicly accessible repositories. In the absence of such policies, there are already cases of digital data which have been lost or remain inaccessible or accessible only with high barriers. While laudable efforts such as the NSF and NIH data management plans move the community in the direction of supportingU.S. economic growth and productivity, the reality is that many researcherscontinue to strictly interpret the requirement as sharing data based on specific requests or personal provisions. The Federal policy framework should move public access to digital data away from the current idiosyncratic environment to a systematic approach that lowers barriers to data access, discovery, sharing and re-use. Instead of relying upon individual investigators to interpret and support public access through a point to point network (e.g.,researcher provides digital data upon request), Federal policies should ensure that public access can occur through well managed, sustained, preservation archives that enable a legally and policy compliant peer to peer model for sharing. A useful metric for full-fledged public access to digital data is whether someone (or some machine) other than the original data producer can discover, access, interpret and use the digital data without contacting the original data producer. And such infrastructure becomes particularly important as science is increasingly interdisciplinary and global. Finally, fundamental to science is the ability to replicate. Researchers must be able to access data in order to

reproduce results and in the current economic climate, we will not see the same level of research funding so it is imperative that data be shared.

Question 2) Whatspecific 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?

Comment 2) Raw data are not subject to copyright. There are fundamental differences between peer-reviewed publications and digital data stemming from federally funded research in the context of copyright and intellectual property. The existing copyright framework and associated business models for peer-reviewed publications are not appropriate for digital data. Unlike publications, data are not processed or reviewed by publishers. Any potential assignment of rights should be applied only to derived datasets for which there is tangible intellectual or creative interpretation or processing. Even in such cases, copyright should be not assigned in an exclusive manner that would curtail or inhibit preservation, discovery and sharing of data. While the Creative Commons CC0 and CC-BY are legallydefensible licenses for publications, they would require augmentation to apply for all types of digital data. Their principles represent an appropriate foundation from which to build a license for digital data that acknowledges potential copyright issues while maximizing the prospects for both people and machines to build services that maximize utility.

If the US Government were to consider the limited use of embargoes, the one reasonable argument for embargoes relates to the unique effort exerted by the digital data producers or the original scientific team. It is true that the original digital data producers exert (often) unique effort that could be acknowledged in terms of a limited, fixed-term,exclusive access to this data. However, it is also important to note that such an arrangement should not confer copyright or preclude the deposit of the digital data into a certified digital data repository even during an embargo period, particularly to initiate archiving activities and (ultimately) subsequent sharing mechanisms.

Question 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?

Comment 3) There is ample evidence that different scientific disciplines present a variety of requirements for management of digital data. Fundamentally, there are two important considerations in this regard. First, there are still baseline conditions or requirements that apply to all data regardless of discipline, particularly as they relate to archiving and preservation. While there are notable exceptions, too many scientific disciplines have focused primarily on access or discovery rather than archiving or preservation. There are critical pieces of the digital data infrastructure that can support archiving, preservation or enhanced access (e.g., identifiers, fixity information) that should be applied to all scientific data. For example, the Public Library of Science assigns identifiers to figures within articles. By doing so, it becomes possible to discover, share and preserve data at a more granular level.

Second, while scientific communities are indeed the most qualified to decide regarding appropriate community practices and norms, the reality is that many scientists do not fully understand the implications oftheir preferences or choices or appreciate the choices available to them. In this context, social science and information science research has started to identify implicit and explicit issues that relate to differences and commonalities across scientific disciplinary data practices. As noted in the recent National Science Board report, “Digital Research Data Sharing and Management,” communities of practice should “take responsibility for determining its own standards and conventions for data stewardship and for coordination across the research enterprise.” It is equally important to ensure that when scientific communities identify and recommend data management practices, they are held accountable to the public access concept and focus on scientifically defensible criteria.

Question 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?

Comment 4) The reality is that different types of data and community-based requirements will introduce different, relative costs and benefits. However, it is most useful to consider requirements particularly as they relate to baseline services thatapply across all disciplines (e.g., archiving) and secondary services (e.g., specialized query capabilities). Agency policies might consider the relative emphasis between these two categories of services, especially as it relates to distribution of costs and benefits across the full array of stakeholders. For example, a federal agency might wish to fund research libraries to develop the capacity for digital data archives and look to new sustainable funding models for the long-term preservation and access to these digital resources. In this sense, an agency might provide seed funding to develop and establish the preservation infrastructure, provide ongoing funding to a scientific community to develop secondary services and explore new partnerships for long-term preservation and access.

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

Comment 5) There exists diversity in approaches for data management within various scientific communities, which is healthy for various reasons. In cases where communities have resources for data management, it is worthwhile to build upon existing infrastructure (e.g., Interuniversity Consortium for Political and Social Research for survey-based social science research). However, it is critical that even in these cases the community service provider demonstrates rather than assert capability. Far too often, terms such as archiving or preservation are being used loosely without associated evidence of meeting specific requirements. Cultural memory institutions such as archives, libraries and museums have an extensive track record with these functions and could serve the essential purpose of developing and/or implementing frameworks that thoroughly test and certify assertions. With a clearly articulated set of requirements, it will become possible to identify how various stakeholders can implement data management plans, noting that these roles will vary by discipline or community.

Since the National Science Foundation and the NationalEndowment for the Humanities announced their agencies' guidance on data management plans, a growing number of research libraries have collaborated with researchers and scientists on developing effective data management plans for proposal submission. It is certainly worth learning from and leveraging thelessons learned from these institutions and in particular, those institutions that directly handle data offer the richest experiences to consider. For example, Johns Hopkins University Sheridan Libraries has acted upon an agreement with the Astrophysical Research Consortium to archive and preserve the Sloan Digital Sky Survey (SDSS) datawhich are considered to be an exemplar collection within the scientific community. SDSS data, which comprise almost 140 terabytes, are also used extensively by citizen scientists who have even helped discover new astronomical objects. On an institutional level, the Johns Hopkins University Libraries have established a data management service that has helped over 35 research teams prepare data management plans for National Science Foundation proposals and committed to archive and preserve the data from these proposals. The Johns Hopkins University Libraries have already acquired the first datasets to be preserved and shared through this data management service.

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

Comment 6) Given the critically important role of digital data to the scientific enterprise, the most important step would be to acknowledge and communicate to federal grantees that the real costs of preserving and making digital data accessible are indeed legitimate and important costs of the overall research enterprise. Researchers do not generally object to including publication costs within their research proposals; it is important to assert that proper data management should be viewed in the same manner.

In addition, agencies could support funding of twenty-first century workforce development. There is some currently underway such as that provided by the Institute of Museum and Library Services but

additional support by other agencies is needed. This was acknowledged in the recent National Science Board report, Digital Research Data Sharing and Management.

Question 7) Whatapproaches 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?

Comment 7) One of the key points in this context is that it is easier to verify compliance through systematic approaches. It is easier to verify compliance of library-based or community-based data archives than to check thousands of individual researcher hard drives. Technical infrastructure components such as persistent identifiers and appropriate licenses represent critical mechanisms through which compliance and verification can be automated thereby reducing costs.

There are two milestone events that every researcher cares about deeply: proposal submission and publication submission. These two points of leverage represent the best instances to introduce or implement policies given the heightened attention of researchers. By embedding appropriate policy, license and infrastructure requirements or components into these workflows, the prospects for efficient compliance and verification are heightened considerably. Researchers will likely complain about the additional burden but as their institutions or their communities develop capacity to support and implement data management plans, those "burdens" can be shifted to entities that view such activity as part of their core mission (i.e., do not view them as burdens but rather core business). This was the case with the implementation of the National Institutes of Health Public Access Policy. Compliance is now considered routine and a key component of ensuring future grants will flow to researchers and the research institution.

Agencies could also provide guidelines to proposal reviewers highlighting the elements of a well-developed data management plan, noting that disciplinary or community practices may vary (e.g., National Science Foundation could develop such reviewer guidelines by directorate).

Question 8) Whatadditional 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?

Comment 8) Federal agencies could stimulate the development of public access digital data archives that support discovery and download and also support APIs that allow individuals and machines to develop new capabilities and services. In particular, this type of open system would facilitate new opportunities for all types and sizes of businesses including small businesses, perhaps somethinglike an app store for data. Thelicensing arrangements would be critical to ensure that one single entity or group does not secure an exclusive right to generate new business opportunities. By fostering a broader array of participation, federal agencies could help build upon citizen science efforts which, to date, have primarily engaged the public mainly through data gathering or data classification activities. Examples of such successful initiatives that offer rewards to individuals or teams working on projects include:

http://showoffyourapps.challenge.gov/

http://dev.mendeley.com/api-binary-battle

http://appsforscience.com/

Question 9) Whatmechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

Comment 9) While this topic remains an active area of research and consideration, one of the most important components is author and institutional identifiers (e.g., ORCID) that would support developing attribution and credit processes. Additionally, it seems unlikely that extending existing attribution and credit frameworks or mechanisms can be seamlessly or easily ported into thedata realm, particularly given the importance of machine-based access. Other metrics do exist such as those outlined at:

http://altmetrics.org

By providing all of these metrics through organizations such as ORCID, a greater level of attribution and the impact of the research can be measured.

Question 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.

Comment 10) There are many community-driven data standards for digital, scientific data, most of which deal with interoperability or sharing rather than archiving or preservation. One useful activity, perhaps through an inter-agency group or process, would be the development of an inventory of such standards identified or labeled by function. There are too many to list succinctly within this response. An example of a comprehensive list from the bioscience community is:
http://biosharing.org/standards
Additionally, the minimum metadata requirements for DataCite [1] require at least 5 standard descriptors but also feature an optional, additional 17 extra pieces of information that can be added at the discretion of the researcher.

Question 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?

Comment 11) There are examples within various domains such as FITS within astronomy and FGDC within the earth sciences. In each of these cases, there are undoubtedly several characteristics or reasons for the success (or alternately reasons why such efforts did not succeed in other cases). Social science or information science research offers the most promising means for rigorously studying such processes, especially toward generalizedlessons that may be applied across domains or toward policy development.

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

Comment 12) While there exist groups that work in this area (e.g. CODATA), it would be helpful for Federal agencies to support community-based efforts that connect nodes of data infrastructure development activities. For example, the European-based EUDAT project has already reached out to projects within the U.S. regarding a Data Access and Interoperability Task Force (DAITF) along the lines of the Internet Engineering Task Force. The National Institute of Standards and Technology could be helpful in this context especially toward an idea of a “data grid” that would operate in a similar manner to the power grid.

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

Comment 13) There is widespread consensus within the research community that it is essential to link publications and underlying or associated data. The peer-reviewed publication is viewed as the final “snapshot” of the research process and outcome. One of the most important considerations from a policy, practices and standards consideration is thatthere be a requirement to use persistent, unique identifiers for publications, data, authors, figures, etc. These identifiers not only bolster the linking of publications and data, but alsohelp foster the re-use and development of new services by people and machines. While there are multiple identifier schemes, at this point, perhaps the most important policy decision would be to require using persistent identifiers instead of relying upon existing mechanisms such as website URLs.

Thank you once again for this opportunity to respond to the Request for Information: Public Access to Digital Data Resulting from Federally Funded Research. Any inquiries related to this response should be addressed to G. Sayeed Choudhury (sayeed@jhu.edu).

_____

[1] http://datacite.org/schema/DataCite-MetadataKernel_v2.0.pdf