January 12, 2012

Brian Westra, Science Data Services Librarian, University of Oregon, bwestra@uoregon.edu
Eugene, OR

Comments in response to the Office of Science and Technology Policy Request For Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research
Federal Register Doc No. 2011-28621
submitted electronically to: digitaldata@ostp.gov

These comments include text and distill discussions and ideas from a number of data curation scientists and librarians at other institutions and the University of Oregon. The response also reflects the opinions of the compiler, Science Data Services Librarian at the University of Oregon, Brian Westra.

Preservation, Discoverability, and Access

(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 clear, government-wide policy supporting open data should be the foundation on which other policies are based. At a minimum, it should require that research data be made available via open access repositories and data centers. Allowing data to become a restricted access commodity held by commercial entities will only serve to enrich those commercial data providers, while impeding scientific research and economic development in all sectors. “Open data access with appropriate ethical restrictions can be viewed as a new core principle for developing a global data infrastructure” (Parsons, 2011). The requirement by the NSF to include data management plans in all proposals has been a good step in encouraging researchers to actively plan for and engage in practices that will enable sharing of research data. More specific guidance and requirements for sharing data will help. One example would be to directly tie data stewardship and open access to data, to future research funding.

Of course, data sharing requirements can only be met if there are sufficiently robust and diverse resources to accommodate the variety of data sets collected and generated by researchers across scientific domains. Funders should support the infrastructure for research data, from data centers and repositories for the data, to the metadata standards, and discovery and access tools necessary to find and use those data sets. Policies that minimize barriers to sharing should also attempt to address other issues, such as insufficient local and national infrastructure. For instance, see the NSB (2005 - http://www.nsf.gov/pubs/2005/nsb0540/) report on long-lived digital data: "Participants

agreed to a considerable extent on the main policy issues, even though there is one stark difference between NSF and many other agencies: the vast majority of long-lived data collections supported by the NSF are managed by external research organizations, while other agencies, such as the National Aeronautics and Space Administration (NASA) and the National Oceanographic and Atmospheric Administration (NOAA) focus more heavily on archiving and curating many such data collections themselves."

It may not be necessary for the NSF to hold the data sets themselves, but it is important that stable, long-term funding be made available to maintain data centers and repositories providing open access. Insufficient support for domain-based repositories forces researchers to adopt ad hoc approaches that are not optimal or efficient, and degrade over time. In addition, existing data centers, particularly those that receive federal funds, should be encouraged to accept and curate a broader range of data in their disciplines.

Funding in many cases has been relegated to the traditional term-limited model, which is not compatible with long-term preservation and access to research data. Data preservation and access efforts have been hampered by this instability, and data curated by term-funding in some cases have been put at risk when the repository funding is cut. Secure and stable funding of infrastructure and expertise not only facilitates access, but enables research and development of data tools and services which in themselves can lead to ground-breaking research and economic development. Open access provides new opportunities for commercial development with not only your own intellectual property but also that of others. It opens up opportunities for everyone.

Awareness of data preservation and access problems, benefits and strategies, remains limited amongst many communities. We recommend the addition of basic data management skills into the scientific curriculum in parallel with a campaign to promote a cultural shift in the sciences towards expanding the core principles of reproducibility, transparency of methods and evidence-based assertion.

One of the biggest obstacles to data reuse and preservation is the capture and standardization of metadata. Metadata is at the core of many data reuse issues, from discovery, to trust, and data quality. To realize the goals of growing the economy and improving productivity of the scientific enterprise, we must comprehensively approach metadata through not only technical systems, but also, critically, through social mechanisms.

Lastly, the data science field must continue to be developed, with the recognition that many of the questions in the RFI are the subject of ongoing research.

(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?

We feel it is important that intellectual property considerations be weighted to reflect the contribution of the researcher and the investment of public funds into the research endeavor. Data is fundamentally different from peer-reviewed publications in regards to copyright and intellectual issues. In some cases, embargoes on the public release of data may provide a sufficient accommodation to the needs of the original researchers. It is counter-productive to the research process to grant intellectual property rights over research data to publishers, since that approach discourages free access to the data.

Other steps that can be taken include the development and implementation of strong metadata standards, particularly as they relate to the provenance (history or chain of custody) of the data. Others have pointed out that the Creative Commons CC-BY license, with modification, represents a good foundation for a license for data that facilitates the development of services to maximize the utility of 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?

Policy and guidelines should be created with the full input of practitioners in the research field. In addition, data scientists/librarians and curators should also be an integral part of a collaborative effort to generate transparent and workable guidance that can be realistically implemented throughout the data life cycle. Supporting the work of data scientists is important, as they provide a translational layer that brings the data from an isolated study into the context of related data, ready for future work. The services offered by various data centers and communities, lessens the impact of the inherent differences between disciplines, and between the researcher and the data center or repository's curation practices.

(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?

Domain-based data centers with an ongoing mandate could help to ensure long-term access to data. Although the long-term value of a particular digital data set may not be known until well after its creation, some metrics, such as the cost of recreating a data set, might be considered as factors. Scientific research at this moment in many fields generates more data than can reasonably be expected to be preserved, and that pace is accelerating, so the research community will need to help establish processes to identify and "promote" data that are recognized as worthy of preservation. Preservation and archiving services and platforms are also evolving, so archivists, data librarians and curators should also be consulted in decisions about assigning costs and values to the facets of data stewardship. Data scientists/librarians and curators should be incorporated into the policy framework for data infrastructure. Policies should recognize the range of relative contributions of and

applications for data sets and support strategic delegation of resources by data scientists and curators.

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

In some instances, such as the ICPSR for social science research survey data, an infrastructure already exists. The roles of organizing, annotating and cataloging, preserving, and providing access to information are inherently part of the capacity of libraries. This means that many libraries may have the capability to provide consultation and support for the implementation of data management plans, in collaboration and coordination with other stakeholders. Depending on the resource, these services and skills can be applied to working with local, shared, consortial, or remote/external stakeholders in the process of supporting data stewardship throughout the data lifecycle.

For example, academic libraries and data scientists/librarians and archivists can connect researchers with campus services; collaborate with other stakeholders and the researchers on the development and operation of domain-based infrastructure; promote and support the use of existing infrastructure (assisting researchers in identifying appropriate infrastructure and data repositories, advising on best practices for preparing data and metadata for deposit, etc.); participate in the development of standards; and maintain a current awareness of institutional and funders' policies and assist researchers in meeting them.

For the successful implementation of data management plans, many stakeholders with a diversity of roles, responsibilities and expertise must come together under a common goal of ethical data sharing. Recognizing the need for a diversity of perspectives at the table – including those of the scientist and research team, data curator, technologists, metadata experts, digital archivist, data reuse support personnel, and others – is an initial step towards supporting DMP implementation. Roles and responsibilities must be clearly defined and delegated to the stakeholders with the appropriate expertise. All stakeholders should promote policy compliance along with community-based standards-making.

Specifically: Universities and research organizations - Clarify intellectual property statements regarding data and promote data publication and sharing as part of the tenure process considerations. Recognize and support centralized data management and infrastructure systems and services as valuable institutional facilities for researchers. Offer career paths for data managers and data scientists on campus. Include data training as part of the science curriculum. Research communities – Recognize researcher efforts to preserve and make data accessible and usable when considering rewards for professional achievement. Proactively contribute to standards-making discussions and the cultural shift towards data sharing and complete metadata capture. Libraries – Consider data as not only part of the scholarly communication cycle, but as a publication/resource (with equal

weight as traditional resources) to be curated. Apply extensive experience in cataloging and bringing together diverse, interdisciplinary resources to the data paradigm. Publishers – Enable and encourage ethical data sharing and attribution through citations within publications and research documentation.

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

In many ways the real costs of preserving and making data accessible are just beginning to emerge. The recognition that data preservation and stewardship are not solely technical problems is one step towards identifying and planning for the costs. The cost of data preservation for reuse includes not only the infrastructure and long-term system maintenance fees, but also the expenditure for capturing and structuring metadata, ongoing standardization work and user support. In cases of observation data from the 'small' sciences, the largest cost involved potentially comes when not only making data accessible over the long-term, but making data *useful* well into the future. The amount and quality of metadata required for reuse can potentially dwarf the metadata required for access and preservation. This is the topic of ongoing research.

A secondary issue that is currently not addressed in most research funding mechanisms is that data stewardship, particularly preservation and access (such as format migration) goes on well beyond the life of the grant. There is also a lack of agreement or established guidelines on how research funds can be allocated toward data curation services, since they involve more than direct hardware or technology costs. Clearer, more comprehensive guidelines and provisions in these areas would have a considerable impact on clarifying how the costs can and should be addressed, particularly at the academic institutional level.

(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?

Much as publication submission to an open access resource (i.e., PubMed Central) can be used as a metric for compliance, links to citable and discoverable data sets reported in a final grant report could be a verification mechanism for data sharing compliance. If a minimal set of metadata standards are established that can be harvested from the relevant data center(s) or repositories, that information could provide a very basic and low-cost verification measure. Although this does not directly address many facets of good data stewardship, it would provide at least a starting point. The adoption of data registration services and DOIs for open access datasets are a similar component that could be used to verify compliance, and provide persistent links from publications and reports.

As with other components of data stewardship, these services are likely to evolve, and should be seen as a starting point. Tieing future grant support to compliance, and educational provisions such as the Responsible Conduct of Research requirements are also

mechanisms that will improve compliance. As noted in other sections of these comments, federal support for the infrastructure for preservation and access to research data is necessary to reduce the barriers to compliance by motivated and interested researchers and other stakeholders. Ongoing research support in the areas of data curation will also be key to moving the requisite systems forward to match the volume and complexity of the data being produced.

(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?

Research awards can be directed toward innovative re-use of data, and toward proposals that support and promote standardization, normalization, and value-added services and tools. These kinds of approaches are key to leveraging the data deluge, and opening new "fourth paradigm" approaches to research. Collaborative proposal support between Federal agencies, such as the IMLS and NSF, could leverage the strengths of a broader community of partners to address these issues.

(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?

Several groups are already promulgating data citation standards (such as the Digital Curation Centre in the UK, http://www.dcc.ac.uk/), data set registration (DataCite, http://datacite.org), as well as author identifiers (ORCID,). Incorporating these into guidelines and as standard mechanisms and best practices will not guarantee appropriate attribution, but certainly will help. If federal agencies work with these stakeholders and others to adopt these practices and require them in reports and other documentation, it will lend credibility to these efforts.

Standards for Interoperability, Reuse and Repurposing

(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.

Because of the heterogeneous nature of research data across and even within research disiciplines it is unrealistic to expect to employ a single standard for all research data. Community-driven efforts are necessary and important within a domain context, while broader, more general standards are necessary to promote interoperability nationally and internationally. A combination of standards, refined for each data type, is one option for enabling reuse and repurposing. The development of these standards and involvement of the full range of stakeholders and expertise should be encouraged and supported. One

approach would be to spread support across ISO-level standards development while recognizing the role of community-driven standards creation. For top-down, high-level standards to effectively intersect with bottom-up, detail-oriented standards and best practices there will need to be a decentralized model of support and communication with a defined path towards formal integration.

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

There are a number of organizations and agencies in other countries that may be further along on the path to coordinated data policies. Within the United Kingdom, the Digital Curation Centre (http://www.dcc.ac.uk/) regularly provides and seeks out collaborations with U.S.-based organizations such as the Coalition for Networked Information (http://www.cni.org/). Some other likely partners are the Australian National Data Service (http://www.ands.org.au/), ISO, Dublin Core metadata iniative (http://dublincore.org/), and groups within the European Commission, for example.

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

Federal Agencies should support the work of the International DOI Foundation (IDF) and of DataCite and other organizations with the goal of making an unified international standard and support structure linking between publications and associated data. Access and use of this mechanism/standard should be easy for the data generator. There may be a role for a Federal Agency to act as a mediator or an issuer of DOIs.

References and Resources:

Parsons, Mark. (2011). Expert Report on Data Policy and Open Access. GRDI2020.