The Alexandria Archive Institute
Comments on the Request for Information on Public Access to Digital Data
Office of Science and Technology Policy
(FR Doc No: 2011-28621)

January 12, 2012

Overview

The Office of Science and Technology Policy (OSTP) recently issued a Request for Information welcoming comments and recommendations for ensuring long-term stewardship of, and broad public access to, digital data resulting from federally funded research. The Alexandria Archive Institute (AAI) commends the OSTP for further exploring this topic.

The AAI (http://alexandriaarchive.org) is a non-profit organization that works to promote the dissemination and curation of digital scholarly resources. To this end, we developed Open Context (http://opencontext.org), a free, open access system for the publication of research content. Open Context demonstrates readily achievable ways to cultivate a distributed foundation for digital scholarship. Its methods for data portability enable researchers to work across silos and use a host of visualization, search and analysis tools. By leveraging archival and identity services offered by the University of California's California Digital Library (CDL), Open Context gains a strong institutional foundation for permanent citation and archiving.

We are delighted to have the opportunity to weigh in on the topic of "public access to digital data." Our responses to the questions posed in the RFI are based on ten years of exploration of issues around open access to digital data in the scholarly community. Below, we list our primary recommendations for encouraging public access to and preservation of digital data resulting from federally-funded research. Our responses to each of the RFI's questions follow the recommendations and provide more details to support each recommendation.

The OSTP request is the most recent development in broad moves to foster improved access, transparency, and stewardship of scientific data. The National Science Foundation (NSF) and private foundations have invested in developing technologies, standards, and datasets to support research. While we applaud recent developments promoting scientific data integrity and accessibility, policy provisions need to be strengthened. Data sharing remains at the margins of professional practice (*Nature* Editors 2009). The scientific community needs to put greater emphasis on data access and reuse to promote more robust, analytically rigorous, and more replicable scientific inquiry. To do so, the OSTP should

adopt a number of policies to clarify key requirements for maximizing the value of scientific data.

Summary Recommendations

Our recommendations are as follows:

- Cultivate a distributed information ecosystem: Integration, synthesis, analysis, and visualization of scientific data can foster tremendous opportunities across the commercial, not-for-profit and academic sectors. Agencies should foster an "open playing field" encouraging innovation in scientific data management and fresh ideas to advance new workflows, organizational forms, and technologies. To cultivate an open playing field, agencies need to promote the free flow of scientific data across multiple platforms and applications employing widely-used open and non-proprietary standards and formats.

- Cultivate a robust preservation infrastructure: Qualified digital libraries and digital archives are needed to maintain the integrity and longevity of scientific data. But not every participant in science data sharing needs to be a repository. To encourage innovation and experimentation, "sustainability" should not be required of every dissemination, visualization, analysis or aggregation platform. Rather, sustainability efforts should focus on digital libraries and archives. Since our understanding of how to best preserve digital data continually evolves, policymakers need to encourage innovation and collaboration across a broad spectrum of public interest organizations, particularly libraries and museums dedicated to playing stewardship roles. Multiple models, approaches, and organizations should play a role in scientific data stewardship to encourage continual learning and innovation in data longevity practices.

- Encourage data professionalism: Federally-funded research both creates and reuses data. Scientific integrity requires proper publication (including documentation) of data, and proper attribution and sourcing of reused, reanalyzed datasets. Data publication (including various models of peer-review and disciplinary archiving) and citation practices need to be mandated for federally funded research.

- Require non-proprietary data: The purpose of public support of science is to expand human understanding, not to subsidize particular commercial publishing models. In general, primary scientific data should be as free as possible from intellectual property and proprietary encumbrances. Such encumbrances create legal risk and complexities

that inhibit innovation around scientific data. Datasets should be in the public domain or under an open copyright license (such as the Creative Commons Attribution License) to widely encourage innovative approaches to data preservation and reuse.

- Data ethics: At the same time, the general need for minimized legal encumbrances should be balanced with data privacy and sensitivity issues. Privacy, research ethics, environmental and public health security concerns, and cultural property and indigenous rights needs, all require consideration. Defining ethical practices for data preservation, dissemination, and reuse will require broad-based, multi-stakeholder negotiations for different types of data in different scientific domains.

Responses to Questions

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?

Recent moves by NSF and NEH requiring Data Management Plans of all grant-seekers demonstrate how data sharing is becoming an expected outcome of the research process. These new requirements have the potential for improving transparency in research. Shared data also opens the door to new research programs that bring together results from multiple projects.

The downside to these new requirements is that grant-seekers may lack expertise and technical support in making data accessible. Thus, the new data management requirements will initially represent something of a burden, and many grant seekers may be confused about how to proceed. However, we expect the benefits of greater data accessibility, quality, and longevity will greatly outweigh any costs as expertise, support services (including the "Data Management Plan" tool offered by the California Digital Library and partners), and infrastructure mature.

(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 imperative that private intellectual property interests do not come at the expense of the public interest in promoting open and reusable scientific data. The public interest is best

served if scientific data can freely flow within a rich and competitive ecosystem that includes commercial publishers, aggregators and others, including nonprofits and academic institutions that contribute valuable services. Commercial, institutional, or other private interests should not have exclusive rights over scientific data created through public financing, since exclusive rights would only impede scientific inquiry and public transparency by imposing higher costs and legal risks. Public support for science should not serve as a subsidy for commercial (or not-for-profit) publishing interests.

The information ecosystem needed for scientific data management should make provenance and attribution of datasets easy to establish. This will promote citation and credit, both of original data creators and down-stream agents (commercial or non-commercial) that further enhance value. Citation is absolutely integral to scholarly practice. It enables and expresses collaborative knowledge production across space and time, and it is the foundation on which evidence and arguments are identified, assembled, reused, and critiqued. A major element on which careers are made and judged, citation is a key aspect of bringing digital communications into professional reward and incentive systems. But citation should not require complicated licensing or other encumbrances that would only make data expensive and legally risky to reuse.

Reliable and robust citation systems are key requirements for publishing data (Altman and King 2007). The DataCite project is establishing dataset citation standards and systems. DataCite promotes simple and readily adopted metadata requirements. To ensure persistence in citation, DataCite also promotes institutionally backed persistent identifiers, such as DOIs (Document Object Identifiers) and ARKs (Archival Resource Keys). The DataCite consortium can help establish the policy and technical requirements needed for efficient processes related to data citation, provenance, and credit.

(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?

The field sciences (archeology, field biology, and other areas of the environmental sciences) can be highly heterogeneous in terms of institutional context, methodology, and disciplinary perspectives. Because of this heterogeneity, it is often hard to establish pre-planned or top-down "cyberinfrastructure" that can meet researcher needs. To address this gap, policy efforts should encourage more bottom-up "Web-style" approaches as appropriate. Effective data management needs wide and open participation among diverse domain researchers to develop standards, data dissemination systems and work practices, as well as funding streams to support research that reuses existing datasets.

Data sharing advocates agree that data sharing requires more than "dumps" of raw and undocumented data on the Web. To be useful and used, data need adequate documentation to facilitate discovery and intelligibility. Data must also be disseminated through trusted

channels with clear versioning, quality control, and preservation mechanisms. Offering data with sufficient quality and levels of documentation requires expertise and effort. Doing so implies greater professionalism than encompassed by the term "sharing."

The term "publication" often better captures the effort, thought, and professionalism needed to make data dissemination meaningful (Kansa 2010). Publication models can align professional and career interests with public and scientific interests (see Costello 2009; Griffiths 2009; Piwowar et al. 2007). Policy efforts should promote innovative forms of professionally edited and reviewed data publication. Different data publication outlets can play an important role in shaping and communicating standards and expectations of data quality. The effort of cleaning and documenting data can be spread across multiple dissemination venues, each with editorial processes or other quality control mechanisms to improve quality, documentation, and alignment to expected standards.

In terms of technology, systems for reusing and analyzing shared data should be open for a variety of approaches. Experimentation can explore different mixes of "Linked Data" (Semantic Web) and more widely used "Plain Web" (Wilde 2008) approaches (favored by many commercial Web and mash-up developers) that may be appropriate in different circumstances. But all of these experiments must be supported by a distributed preservation infrastructure that safeguards data for the long term (Kansa 2011).

(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?

It is difficult to anticipate the impact of datasets over time. We need to experiment and develop a much richer body of experience to better understand which datasets to prioritize for dissemination and archiving. Data dissemination practices should experiment with ways of tracking the impact and use of different types of data, both with citation impact measures and alternative (Web, social media) impact metrics. Thus, different research communities will learn from experience about which types of data to prioritize for dissemination and archiving.

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

All aspects of the data lifecycle must be understood and supported. For example, AAI is now working on a data publication workflow tools and practices in order to help shepherd datasets from the hands of the author, through a data clean-up, editorial, and documentation process, to a university-backed digital repository where the datasets receive permanent identifiers and can be discovered and used in different ways. In addition to our

approach, many other models of scientific data dissemination should be explored, since optimal approaches will be highly context dependent and will no doubt evolve as methods, expectations, and technologies change.

Demonstrated scholarly outcomes will help make a compelling case for sharing datasets. Graduate students, undergraduates, high-school students, and informal-education learners need training opportunities that offer Web data skills so that the next generation of researchers can best make use of emerging data resources.

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

Funding agencies should require a line item in proposed project budgets that commits a certain proportion of the project budget to publication (ideally in open access venues), as well as data dissemination and preservation. Dissemination and data preservation services and costs should be outlined in a project's data management plan and justified in budget justifications. Reviewers should be asked to carefully consider whether a project's budget seems appropriate to the proposed data management plan. Reviewers must understand key issues in data management so that they can better evaluate data management plans in proposals. Finally, funded projects should be required by the funding agencies to provide details in their interim and final reports about how project data have been disseminated and archived according to the data management plan.

In general, scientific knowledge and its underlying foundation of data can be expensive to produce. It requires expertise and often great effort. But to make data integration and reuse feasible, data need to flow freely and interchangeably. Unless there are strong overriding ethical or security concerns (chiefly privacy), access and use should be free-of-charge and free of legal encumbrances (especially proprietary IP interests). Therefore data dissemination and archiving services should focus cost recovery on accession fees (budgeted in grant proposals), not on fees for later access or use.

The key point is that publicly funded research should create public information goods. The outputs of publicly funded science should be available in a robust, expanding, and open public commons. Commercial interests can and emphatically should build upon this commons of public information goods. But, particular commercial or even nonprofit interests should not be able to monopolize the public commons or exclude others from freely drawing upon the fruits of publicly financed research.

(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?

Participation in suitable, disciplinary dissemination outlets (data publication venues) and digital repositories should be easy to verify. In order to serve any scientific purpose, these systems will provide datasets with enough metadata documentation to make it easy for officials to verify compliance.

(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?

Granting agencies should develop funding streams specifically targeted toward encouraging graduate student use of publicly accessible data. These funding streams will help cultivate the needed Web data and analytic skills required to use public datasets effectively. They will also help change scientific cultures in ways that encourage greater openness, transparency, and participation in scientific data sharing systems.

In addition, to cultivate commercial innovation in this space, agencies can develop SBIR-type granting programs that fund innovative commercial projects that draw upon and add value to publicly accessible research data without monopolizing or excluding others from those 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 best protection from "data theft" is clear dissemination in widely accessible, easily searched, professionally recognized venues. Professional social norms of citation need to be promoted rather than encumbering and legally complex forms of licensing. Clear citation practices, supported by appropriate technical infrastructure, will promote proper attribution and professional rewards. Agencies can also encourage participation in open-access "data publication venues" with review and other editorial processes recognized by the researcher community. Publication of data, in recognized forums, where citation impacts can be tracked (like article impact metrics) can provide appropriate professional recognition for data creators.

Standards for Interoperability, Re-Use and Re-Purposing

(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.

Standards continually evolve and will vary by research community. Rather than dictating specific standards, agencies should set policies that promote adoption of "good practices" while recognizing that such practices continually evolve.

A key requirement for data dissemination should be data portability. Data should not be trapped in a given system or repository. Rather, data need to freely flow into different applications and systems. This requires that data have open licensing or lie in the public domain. Also, various machine-readable representations of data need to be available, though the specific formats and standards may vary across and between different research communities. In some cases, more elaborate standards may be required. In other cases, overly complex standards requirements may inhibit adoption. Simple and lightweight technical and semantic standards that yield immediate and tangible benefits may be most suitable for scientific domains with little funding or technical support.

The World Wide Web is an obvious choice for dissemination. As much as possible, data dissemination systems should adopt best practices in Web architecture. Systems need to adopt non-proprietary open standards and offer data in multiple machine and human readable representations. In many cases, Web-based dissemination should also emphasize and support "Linked Open Data" methods.

In our own efforts with Open Context, we have emphasized low-barrier to entry approaches to sharing machine-readable data. We offer data in widely used and recognized formats, including JSON, the Atom Syndication Format and other XML vocabularies. We are also incrementally adopting more Linked Data standards and services for interacting with RDF data. Our experience shows the need to continually adapt to expose data in new formats and services as expectations and needs evolve over time.

(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?

Some of the best examples of standards development come from the public Web. The Atom Syndication Format is a major example, but other successes include GeoRSS and GeoJSON. Development of these standards was largely "bottom up" where software developers (the stakeholders that actually implement the standard) play a key role in shaping standards development. Keeping complexity and barriers to entry at a minimum is vital in any standards building effort.

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

Federal agencies can offer funding support for efforts that help build ties between international collaborators. The "Digging into Data" challenge organized by NSF, IMLS, and NEH represents a good, though under-funded model (grants awards were quite small relative to the costs and complexity of multinational collaborations).

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

The Web is the most successful distributed computing platform ever developed, and principles of Web architecture need to be used when linking between scientific datasets and other scholarly publications. Unfortunately, many scientific publications are merely electronic analogs of "paper," typically with little linking to other resources published on the Web. We need to see far more innovation in scientific publication processes, but this innovation is hamstrung by professional reward and evaluation structures and by the fact that scientific publishing is dominated by a few monopolistic commercial publishing houses.

In addition, the divide between a scientific "dataset" and "publications" is not always hard and fast. Many important scientific inferences and analyses could be conducted through large-scale text analyses and mining of scientific literature. However, access to this literature is highly restricted through very expensive and tightly guarded commercial channels. These restrictions make it difficult for the scientific community to explore ways to better use and understand vast bodies of scientific literature.

The NIH requires public access to peer-reviewed outcomes of NIH funded research (usually delayed by one year, so commercial publishers temporarily enjoy exclusive dissemination rights). Similar requirements should be made by other federal agencies. In this way, publications can be used to generate additional datasets, and datasets can be fully understood and contextualized by accessible publications.

References

Altman, M., and King, G. 2007. A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine, 13(3/4). Retrieved from http://dx.doi.org/10.1045/march2007-altman

Kansa, E.C. 2011. New Directions for the Digital Past. In *Archaeology 2.0: New Tools for Communication and Collaboration*, edited by E.C. Kansa, S.W. Kansa and E. Watrall. Los Angeles: Cotsen Institute of Archaeology Press, pp. 1-25. Retrieved from http://escholarship.org/uc/item/1r6137tb#page-17.

Kansa, E.C. 2010. Open Context in Context: Cyberinfrastructure and Distributed Approaches to Publish and Preserve Archaeological Data. *The SAA Archaeological Record* 10(5), 12-16.

Costello, M. J. 2009. Motivating online publication of data. *BioScience* 59, 418–427.

Griffiths, A. 2009. The Publication of Research Data: Researcher Attitudes and Behaviour. International Journal of Digital Curation 4. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/101.

Nature Editors. 2009. "Data's shameful neglect." *Nature* 461, 145 (10 September 2009).

Piwowar H.A., Day R.S., Fridsma D.B. 2007. *Sharing Detailed Research Data is Associated with Increased Citation Rate*. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308 Retrieved from http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0000308

Wilde, E. 2008. *The Plain Web.* pp. 79-83. In: Proceedings of the First International Workshop on Understanding Web Evolution (WebEvolve2008), 22 Apr 2008, Beijing, China. ISBN 978 085432885 7.