UNIVERSITY OF
NORTH TEXAS
DISCOVER THE POWER OF IDEAS

UNIVERSITY LIBRARIES

Dr. Martin Halbert
UNT Libraries
1155 Union Circle #305190
Denton, TX 76203-5017

January 12, 2012

To the Office of Science and Technology Policy:

Thank you for the opportunity to offer comments in response to your RFI, Public Access to Digital Data Resulting from Federally Funded Scientific Research. This issue is of major long term significance to the long-term success of scientific research. The University of North Texas is very concerned with the survivability of scientific research results, and has undertaken a research project aimed at understanding and identifying solutions for the complex problems that you have marked. This research, still in early stages, is funded by the Institutes of Museum and Library Services (IMLS). In this brief response, we would like to a) share the early results of this research project, and b) offer our own recommendations to the questions articulated in your RFI. Our recommendations are informed by both our research effort and our perspectives as members of the National Digital Stewardship Alliance (NDSA) and the Scholarly Publishing and Academic Resources Coalition (SPARC). We participate in the organizational discussions concerning this issue within NDSA and SPARC, and endorse the comments which they have separately provided in response to your RFI.

UNT Research on Data Management

Our research project is entitled "DataRes: Research on Emerging Research Data Management Needs" and will be documented at URL http://research.library.unt.edu/datares/. The central questions of our research include: How are universities actually responding in terms of policy to data management requirements from funding agencies? What are the practical needs of researchers to meet the demands of those requirements? And, finally, how can university libraries, and the library and information sciences field at large, address the needs of researchers for data management, retention, and sharing in terms of services and support, training, and infrastructure?

In our initial findings, we have learned that of the top 200 NSF and top 200 NIH awardee schools (approximately 220 institutions), only 22% have published policies supporting the retention and sharing of research data; of the top 50 NSF awardee schools, only 50% have published such policies. In a recent focus group with NSF Program Officers, they clearly articulated the importance of university-level support in terms of policy and infrastructure to facilitate the recent NSF requirement for data management plans in funding applications.

In general, we have found that university libraries have proactively stepped forward to meet the needs of researchers in preparing and implementing data management plans, sometimes in the absence of top-down institutional support. In the coming weeks, we will conduct a wide ranging survey of stakeholders in research data management – researchers, librarians, office of research administrators, provosts, and others – to further determine the needs and perceptions of this diverse community, as well as conducting further stakeholder focus groups at professional meetings and conferences. We will also conduct targeted surveys of administrators at those institutions without published data retention and sharing policies to better understand the decisions behind the absence of such policies.

Based on our preliminary findings, we believe that data management planning requirements on the part of federal funding agencies have the potential to stimulate significant changes in the way research communities think about the retention and sharing of their data. It is, however, critical, that funding agencies recognize that the preservation of research data does not come without costs in terms of staffing, infrastructure, and ongoing maintenance and repository services; as such, some guidance for incorporating these costs into research proposals may be in order. Likewise, a broader statement of support for open access to research data and published outputs from above the agency level would be extremely valuable to help facilitate the socialization of these policies in the broad research community, and in related support industries, including libraries and publishing.

Responses to RFI Questions on 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?

Federal agencies should require and enforce stronger mandates concerning the long-term stewardship of research data. A very simple step would be to a) require all federally funded research projects which produce research data to report to the granting agency a permanent URL representing a location in which the data produced in the research project may be found, and b) publish these URLs in public registries or existing web locations which list the grant awards made by the agency. This requirement could be easily audited on an ongoing basis by automated URL checking software to ensure that data continues to be publicly available.

(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?

Only digital data which is funded by public funds should be required to be maintained as publicly available. Copyright and other appropriate guarantees of intellectual property are distinct from this class of content and should be addressed separately.

(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 form and standards associated with data from different fields differ, and should be considered on a case by case basis. The recommendation in (1) above only makes the case that a permanent access point for research data should be provided, not the form the data should take.

(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?

Disciplinary differences are real, and (as above) must be considered on a case by case basis. Establishing a fundamental requirement for long term access to research results remains the first priority. Strategies and mechanisms for ensuring the long term survivability of research data is a matter for institutional innovations and is a worthy subject for grant-funded competitive experimentation. Federal agencies should reserve some funding for precisely this kind of long term sustainability research in order to increase the likelihood of successful innovations.

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

Best practices are still emerging, and practical innovations in this area of expertise should be explicitly fostered and catalyzed through federal dollars. The National Digital Infrastructure and Information Preservation Program (NDIIPP) was a fruitful initial program that involved NSF research efforts. Further federally-funded research in this emerging area is needed to make progress in long term sustainability of digital data.

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

One of the priorities for research needed in sustaining research data concerns setting reasonable cost standards for such expenditures. Without such guidelines and concomitant best practices, there will likely be unhelpfully confused variation in requests and responses.

(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?

See the response to (1) above. Automated compliance verification is a manageable burden under this strategy.

(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?

Standardized access registries as described above would greatly improve access to research data by industry for new and existing markets. Further, innovations in access mechanisms for searching and culling this registry data would be a new growth industry in its own right.

(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?

Provenance of research data for attribution and scientific credit purposes will be improved by publicly verifiable registries as described here. Minimal requirements for attribution could be required as part of the URL strategy described here by linking the permanent data accessibility URLs to registry entries.

Responses to RFI Questions on 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.

This is again a disciplinary-specific topic. Interoperability standards arise naturally for specific fields as researchers seek to share data. It is an area that could be catalyzed for some fields by making competitive grant awards available for applicants that adopt or offer to collaboratively establish emerging standards.

(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?

Obviously there are many examples to cite, here is one: The ISO WARC standard format for web archives emerged from discussions in the web archiving community on simple strategies for making the results of web crawls more manageable. The content of web crawls is stored in standard formats for long-term access and preservation. This community standard is publicly documented, see the URL http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717 for full information.

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

A first step is to convene international meetings to explore this topic. A good example of such an exploratory meeting was the Aligning National Approaches to Digital Preservation (ANADP) conference recently held in Europe (see http://www.educopia.org/events/ANADP). Such meetings should have associated specific outcomes set forth in the initial planning.

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

Many protocols are in development for this purpose. A promising candidate is the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) specification, documented at URL http://www.openarchives.org/ore/.

Thank you again for the opportunity to respond to your RFI. This issue is of tremendous importance to American institutions of higher education, from administration, researchers, students and libraries. If you have any questions concerning any of these responses, do not hesitate to contact us.

Sincerely yours,

Martin Halbert, PhD, MLIS

Dean of Libraries and Associate Professor