**Responses to the 11/4/11 OSTP Data RFI**
**from the Science and Technology Office**
**by David B. Lowe, Preservation Librarian**
**<david.lowe@uconn.edu>**
**University of Connecticut Libraries, Storrs, CT**
**1/12/2012**

## 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 policies and programs that sponsor or otherwise facilitate the **creation and maintenance of discipline-specific repositories for research data** could encourage public access as well as preservation. Progress on this effort would need to start with a proper comprehensive inventory of such repositories, perhaps followed by a certification or at least vetting and recommendation process on behalf of federal funders, ideally involving professional associations and societies. Where the inventory reveals gaps in the spectrum that existing repositories cover, it would then be proper and desirable for federal funding to attempt to foster initiatives to cover these lacunae.<br><br>The reason for this need is related to a lesson that libraries and archives have learned over the millennia of gathering and organizing information objects of cultural significance: collections contain context. Context is crucial in identifying knowledge entities relative to one another, establishing hierarchies, and assigning priorities that are prerequisites for progress with scientific methods in particular, not to mention with any intellectual endeavor in general. Research assumes such structures as a basis upon which to progress and build further, relying on the credibility and veracity of past work, and our new digital environment should not be an exception.<br><br>In addition to the context intrinsic to a collection, its niche market tie to its clientele is also critical, in that—out of all the knowledge resources in the world—it can be positioned closer to those who are most familiar with it. This will be especially |

| | |
|---|---|
| | important in the future as we confront migration issues. The best strategies and solutions will involve knowledgeable users who are closest to the content and who can help ensure its viability into the future. |

| | |
|---|---|
| *(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?* | Access controls that allow in vetted researchers during an embargo period if they agree to respect intellectual property constraints via attribution and citation could be a solution for stakeholders' concerns about their data in a repository as mentioned in the response to question #1. From talking with researchers as part of formulating our institution's response to the NSF Data Management Plan requirement and also as part of an eScience Institute sponsored by the Association of Research Libraries (ARL), I understand that the most common model that has developed in the academic community over the years has been one of willingness to share data when asked. As we transition to better infrastructure for preserving and also sharing via open access, the part of that established model that could be lost would be the fact that the researcher and the requester are aware of each other, at least at some minimal level of identity, in a relationship of professional trust. Establishing a certain reasonable embargo period, matched to community norms and during which this controlled access could take place, would restore some of that identity clearance, which could take place on a case-by-case basis for individuals, or could be open to established researcher groups, which in turn might have their own certification processes that their communities can trust. No one questions that federally funded research should be made public eventually, except in a relatively few cases of privacy or security, so the problems to solve revolve around the timeframe of active projects and the 3-5 year window that follows. As I write this, the news features stories about the suppression of some of the specifics of federally funded avian flu research, which is just one case related to national security. The point here is that with a robust system of access controls, researchers who legitimately need to know these details could get what they need, while those who lack proper credentials would be denied access. |

| | |
|---|---|
| *(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?* | Although inherent differences between disciplines do pose a problem for those seeking to establish equitable data management expectations in the grant funding context, there are some clear watershed areas for distinguishing between groups. One of the most important litmus tests involves research data that needs to contain personally identifiable information (PII) and also data that has sensitive implications across a broad spectrum of security issues. Fortunately, these two areas of PII and security tend to be governed by other rules that can take precedence over federal grant funding guidelines. A second area of concern would be the "haves vs. have nots" in terms of adequate repositories, which my response to question #1 above attempts to address. Until and unless there is an appropriate place for a researcher's data, it does not seem fair to ask her to meet the same requirements for deposit as those who already have adequate places to park those files. |
| *(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?* | Long-term usefulness cannot be immediately known or quantified, but the historical lesson that libraries teach us is that information kept just-in-case does in fact tend to come in handy within a sufficiently inclusive timeframe. It would be short-sighted to jettison reasonably retainable data now just because we make some capricious determination that it will be of no use down the road. Tossing information out guarantees that it can be of no help in the future. Also, crosswalking data sources to make them searchable for cross-disciplinary purposes will greatly enhance their potential usefulness as a benefit for all. |
| *(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?* | The top priority in contributing to the successful implementation of data management plans is the establishment of an adequate repository infrastructure, especially the core metadata ecosystem that makes ingest, management, discovery, access, sharing, and preservation all feasible. |
| *(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?* | Cost-shared efforts for archiving the data produced in grant projects deserve to be weighted more heavily than a 1:1 dollar value. Related service costs should be given higher value and consideration than those traditionally featured in that grant proposal budget column, at least in these early stages when we are attempting to establish adequate workflows. To be more explicit, a project that follows its discipline's established metadata schema has less preservation work to do than a project for which metadata schema development is lacking. Any schema development done within a project, then, deserves to be incentivized. |
| *(7) What approaches could agencies take to* | By standardizing repositories and automating their functions, management issues like metrics, verification, and compliance checking would become vastly easier. We need |

| | |
|---|---|
| *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?* | to raise expectations in these areas and put the mechanisms in the right places to accomplish these goals. |
| *(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?* | After funding proper repositories, next logical steps would be data mining and presentation projects. The pent-up wealth of information could give rise to new fields and specializations that dig into the fabric of the information assembled and cull from there patterns that in turn spawn a demand for eyes and hands that can present the new findings in visually stimulating and meaningful ways, not to mention then applying the knowledge then revealed to the real world to make our lives, our cities, and our societies better. |
| *(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?* | Access control mechanisms could enable tracking that conveys full credit and attribution. Components of such controls would include better universal identifiers for people, institutions, publications, and parts thereof. There are significant researcher privacy concerns here, but certainly many would be open to an opt-in identification model if it also made their citation work easier through automation. For the rest, there is no perfect solution to plagiarism and theft, but at least it may be easier to deny access to known past offenders if adequate controls are in place. |
| **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.* | Any standards that the respective communities develop are the right ones. It is always the hands-on users who should make that determination. This is not to say that we cannot do a better job of aligning variants within a discipline or of making cross-disciplinary standards more compatible with each other, but the specialists should always decide about the particular data points captured as a "business rule," as code developers and analysts would say. Alignment and compatibility is something that metadata librarians would be able to help with and should be involved in. |

| | |
|---|---|
| *(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?* | I would point to MARC for bibliographic information in libraries, EAD for collection finding aids in archives, DDI for social sciences data sets, and FGDC for GIS data as effective standards that have created efficiencies and opportunities for sharing. The main characteristic of their development processes is that they all achieved community acceptance above a certain threshold, which in turn made the efficiency pieces happen. |
| *(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?* | Federal agencies could promote coordination by creating funding opportunities for metadata development. Targeting disciplines that lack proper common metadata schema, agencies could offer to fund a conference to discuss community needs, with deliverables that would include draft data points toward a schema. It would not be difficult to find metadata librarians, analysts, and information architects to polish that draft into a serviceable metadata approach, and these professionals could also keep an eye out for cross-disciplinary functionality. |
| *(13) What policies, practices, and standards are needed to support linking between publications and associated data?* | The key piece for linking support lies in persistent identifier solutions. Such identifiers assume other crucial infrastructure is in place, such as proper stable repositories, so these are the most important priorities for the time being in this area of endeavor, as discussed in my responses throughout above. There is a huge role for the professional organizations to play in establishing community norms around metadata schema, discipline-specific repositories, embargoes, and access controls. |