Thu 1/12/2012 5:21 PM

response to RFI

Response to Request for Information "Public Access to Digital Data Resulting From Federally Funded Research", November 2011

January 12, 2012

Clifford Lynch
Executive Director
Coalition for Networked Information

Cliff@cni.org

I am pleased to have the opportunity to submit comments to this request for information on "Public Access to Digital Data Resulting From Federally Funded Research" on behalf of the Coalition for Networked Information (CNI). CNI is a membership organization consisting of some 200 organizations, primarily but far from exclusively universities, who share a common commitment to advancing the intelligent use of information technology and digital content in support of scholarship. You can find more information on CNI at www.cni.org.

I want to be clear that while these comments are certainly informed by discussions with CNI's member organizations, they should not be viewed as representing the position of any specific member of CNI.

There are a tremendous number of questions in the request for information, and I cannot comment on all of them here; I also know that you will be getting many other well informed and thoughtful responses, including some that I have already seen in draft. But I want to begin my comments with an overarching strategic point: we are relatively early in a great transition in scholarship. We have now explicitly recognized the large scale emergence of information technology enabled and data intensive scholarly practice, and have begun to make systematic accomodation of this transition in our funding, policy, infrastructure and scholarly communication mechanisms. We need to carefully monitor what is actually happening as this transition moves forward. We need to examine the effects and outcomes of policy interventions on a continuing basis, and to be prepared to adjust these policies as experience dictates. We still have a great deal to learn about what data is of greatest lasting value, and what data is most likely to be reused in ways that produce new and important scholarship. This is a process, not a one time event. Our ability to manage this transition will be greatly improved by the availability of funding to help underwrite data collection, research and analysis, and also by policies that promote transparency and the public availability of data that can support research, analysis and evaluation. The level of investment needed here is miniscule relative to the scale of the scholarly and research enterprises.

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?

Funding agency requirements for data management are an excellent start, and should be extended from NIH and NSF to all federal funders. Funders also need to continue to reinforce policies that require investigators to share data publically; this would include guidance not just to investigators, but also, importantly, to proposal reviewers. As an overall strategy, most data curation needs to be institutionalized: responsibility needs to transition away from the investigator and to institutions (either universities or disciplinary repositories) early in the data lifecycle, with these institutions being the primary long-term contacts for data access.

Both one-time (startup) and even more importantly sustained federal investment in disciplinary data interchange standards, and in data repositories (both institutional and disciplinary) and related infrastructure (such as disciplinary and cross-disciplinary discovery tools for datasets placed in repositories) are essential parts of the federal contribution.

A particularly problematic set of issues - legal, technical, policy, and ethical -- that may well be ripe for federal leadership are those inhibiting access and reuse of data that involves human subjects and personally identifiable information. Here we see disconnects and conflicts between long-standing but continually evolving practices and policies designed to protect human health, privacy and dignity, and the new opportunities for large scale computational reuse, recombination and analysis. These conflicts are becoming major barriers to progress, notably but far from exclusively in the health sciences. Note that these issues arise on both national and international levels.

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?

I am not a lawyer, but I think that the way this question is framed helps to illuminate one of the barriers to greater data sharing and reuse. Research data, as I understand it, basically isn't subject to copyright in the United States; while access controls and contracts can clearly be used to limit the ways in which it is shared and used in specific cases, and may be appropriate tools for supporting short term exclusive use embargos or similar arrangements, I don't think that there are traditional intellectual property issues here - data, and particularly data from federally funded research, should be regarded as part of a knowledge commons that belongs to all of us. Clarifying and codifying this, for all of the players, including researchers themselves, is extremely important, as is differentiating legal rights and obligations from moral or ethical ones (such as the moral obligation to acknowledge sources of data that are reused in subsequent research).

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?

It is clear that disciplinary differences must be recognized, honored and accommodated. And obviously different kinds of data require different policy frameworks (see my earlier comments on human subjects, for example). Yet we should also recognize that some disciplinary differences are largely disciplinary traditions, and some of these traditions are, in my view, ripe for re-assessment. And disciplinary practices and traditions that are inherently inconsistent with ideas about an open knowledge commons cannot be excused simply on the basis that they are disciplinary practices and traditions. Further: as we move into an era where interdisciplinary and multidisciplinary research is increasingly commonplace and increasingly necessary, greater consistency (or at least interoperability) across disciplinary practices will be more and more desirable - particularly in terms of describing and managing data resources and facilitating reuse of such resources.

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?

I think that this is a very poorly understood area and one that merits ongoing examination and research. There are classes of data that can be re-created, and here one can look at the cost of preserving versus the cost of re-creating. There are also ethical issues involved in data from experiments involving human beings, and, at least arguably, animals. Most observational data, once gone, cannot be replaced. We also have great problems quantifying the likely benefits of preserving various kinds of data. Some of the best thinking currently revolves around thinking in terms of ten or twenty year re-evaluation cycles for data stewardship, where at the end of a given cycle stewardship might transition from one responsible party to another in an orderly way. Cultural memory organizations have considerable expertise to contribute in managing this process.

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

Greater clarity about the extent to which data stewardship costs can be included in grant budgets would be very helpful. We also badly need funding mechanisms to address the existing base of data that has already been created.

But I want to stress that this is not simply a matter of funding mechanisms - I think that federal funding agencies need to be clearer, as a matter of fundamental policy, that they share in the ongoing fiduciary responsibility for stewardship of data that is created as a result of their research funding.

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?

There are several approaches that should be applied in parallel. Investigators can be encouraged to ensure good data stewardship by asking about data specifically as part of the reported results of previous federal research funding supplied with new grant applications. To the extent that operational responsibility for stewardship and access are shifted to institutional actors (institutional and disciplinary repositories) and away from individual investigators, it should be much easier for major funding agencies to measure and verify compliance on a programmatic basis.

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?

Clearly, discovery systems that can look broadly across the repository infrastructure are a key investment here (the search work going on as part of the NSF-funded DataONE datanet grant looks to be a very promising contribution). Beyond this, one could imagine SBIR-type programs to target small business investment to exploit available research data. Speaking personally, I would love to see some sort of prize competition making awards annually for the most creative and highest impact reuse of publically available data in commercial, educational, and research categories.

What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported? Work here is well underway in initiatives like DataCite. The federal funding agencies can help a great deal by asking investigators about datasets that they have created and made publically accessible, and about the impact that the availability of these datasets have had. They can also help by encouraging research proposals that make creative reuse of existing datasets, and by encouraging review panels to consider whether proposals are making effective use of existing data resources.

Standards

I want to first recognize that these are essential in effective data sharing, reuse, and stewardship, and that lack of appropriate standards is a real barrier. Standards evolve, and there is a continuing need to develop or update standards to reflect new scholarly practices. This is a high-leverage area that I think has suffered from chronic lack of funding - there's no clear source to fund the development of standards that are needed to support most data intensive scholarship, or to maintain those standards, or to finance the necessary software development or maintenance/upgrades to make the standards part of the community's tools and workflows. (One rare and noteworthy exception to this has been the INTEROP program that has been part of NSF's Office of Cyberinfrastructure for the last few years.) There's also a lot of experience in standards development, but yet all too often specific scholarly communities establishing standardsf for the first time seem to be re-inventing the wheel. Standards development and support are an area where I believe modest investments by funders would have high payoffs.