Thu 12/22/2011 2:17 PM
Comments on RFI: Public Access to Digital Data

To whom it may concern:

I'd like to submit the following comments regarding the Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research. I write representing only myself, as a professional librarian working to assist researchers with data management and data management planning. My views do not necessarily represent those of my employer.

(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?

- Requiring data management plans is a good start. More specific guidance and stronger encouragement to openly share at least a portion of a project's data would help.
- Funders SHOULD support infrastructure for the data resulting from the research they fund, especially the NSF, which provides less of this support than other federal agencies (see the the NSB 2005 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.") Not adequately supporting domain-based services forces researchers in domains without appropriate data centers to adopt ad hoc approaches that may not serve well in the long run. Even those in a domain with a data center may find their data is not a 'fit' - for example, CUAHSI's HIS only accepts georeferenced data, yet researchers may do lab-based hydrologic research. Existing data centers, particularly those that receive federal funds, should be encouraged to accept and curate a very broad range of data in their disciplines.
- That said (funders should support infrastructure), there will likely always be cases where some data sets do not fit into existing infrastructure, the policies and management of existing infrastructure are not aligned with a data producer's needs, or a researcher or institution has valid reasons for wishing to manage the data "closer to home" (i.e. in their home institution). This should be permissible.
- Funders should make public the specific criteria by which data management plans in grant proposals are evaluated.

(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?

See my comments on (1) - funded, domain-based data centers with an ongoing mandate would serve well to ensure long-term access to date. In addition, the long-term value of a particular digital data set may not be known until well after its creation. Establish processes to identify and "promote" data that are recognized as worthy of preservation, and do not burden the researcher with the cost of subsequent preservation when value of their data sets becomes evident long after their creation. The NSB's 2005 report (referenced in (1)) describes three categories of data collections, research, resources, and reference, based primarily on the breadth of the user community. The report that notes that collections may evolve and change categories; it's also possible individual data sets might move from one category to another. We don't have a good way to make this happen.

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

There are several roles that research libraries can play in this arena:

- Connect researchers with campus and external services.
- Manage data resources locally, when appropriate.
- Collaborate on the development and operation of domain-based infrastructure.
- Promote and support the use of existing infrastructure. Assist researchers in identifying appropriate infrastructure and data repositories, advise on best practices for preparing data and metadata for deposit, etc.
- Participate in the development of standards.
- Maintain a current awareness of institutional and funders' policies and assist researchers in meeting them.
- If sufficient expertise is available on staff, train graduate students on new requirements and best practices for meeting them.

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

Specify periods of data retention in data sharing policies (where that isn't already the case) so researchers have clear guidelines to inform budget planning.

Best regards,

Gail Steinhart
Research Data & Environmental Sciences Librarian, Albert R. Mann Library (mailing address)
Fellow, Digital Scholarship & Preservation Services, Cornell University Library
Cornell University
Ithaca, NY 14853