Thu 1/12/2012 4:41 PM

Comment on RFI: Public Access to Digital Data Resulting From Federally Funded Scientific Research

Mary Ochs / mao4@cornell.edu

President / United States Agricultural Information Network

Ithaca, New York

I am writing on behalf of the United States Agricultural Information Network (USAIN) to respectfully respond to the Request for Information for recommendations related to public access to digital data resulting from federally funded scientific research.

USAIN (usain.org) is an organization of over 150 agricultural information professionals that provides a forum for discussion of agricultural issues, takes a leadership role in the formation of a national information policy as related to agriculture, makes recommendations to the National Agricultural Library (NAL) on agricultural information matters, and promotes collaboration and communication among its members. USAIN has testified before Congress, played an advisory role in the National Agricultural Text Digitizing Project, written a national agricultural literature preservation plan, served on blue ribbon panels to review NAL services, and participated in the selection process for new NAL Directors. Our members are skilled librarians and information specialists with knowledge of the modern theories, principles, practices, techniques, and policy issues pertinent to the current practice of librarianship and information science. Many of our members work at Land Grant institutions with extensive federally-funded research programs and are experienced in acquiring, organizing, and preserving scientific and agricultural data. The USAIN Executive Council is privileged to provide the following input related to this important topic of public access to information.

Comment 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?

Access to digital data is critical for scientists, students, innovators, entrepreneurs, and other interested citizens. This vital part of the process of the discovery of knowledge ultimately leads to the creation of new products, job opportunities, economic growth, research achievements, and the strengthening of

our society's knowledge base. Digital data created at the public's expense, but left unavailable or referenced only via subscription-based journal articles may not be equitably available to all who might benefit from its content**.** A system requiring a data management plan for data output of federally-funded research would necessitate a level of data planning. Ideally it would be beneficial to require researchers to also provide data to a subject-based or institutional repository so the data would be more readily available.

On one hand it is exciting to see the growing interest in digital data dissemination, but it is bittersweet given the recent budget woes and the mandated termination of the National Biological Information Infrastructure (NBII). The main Web site, www.nbii.gov, will be taken offline on January 15, 2012, along with all of its associated node sites. "January 15, 2012, will see the end of a long-term project to empower users of biological resources data and information. The National Biological Information Infrastructure, or NBII, was begun in 1994 within what was then the National Biological Service (NBS) of the Department of the Interior. Its purpose and mission were to ensure that scientists, resource managers, decision makers, and concerned citizens could go to a single place on the Web and find biological resources data and information from vetted sources—whether in government, academia, non-governmental organizations, or the private sector." See the announcement in USGA @ccess http://www.usgs.gov/core_science_systems/Access/p1111-1.html.

Comment 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?

Nothing in a public-access system for data should threaten the protection of intellectual property. In fact, greater access to research information and data will ensure greater visibility and recognition of an author's intellectual achievements. For sensitive or proprietary data or data that data owners don't want to make public for whatever reason, the system can provide access controls, such as password protection.

Comment 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?

Just as there are different journal articles and citation styles in various disciplines, there are different expectations for sharing discoveries and reporting results. There may be differences in the disciplines

but stakeholders and research communities should be encouraged to establish standards that enable sharing and interoperability across disciplines. This will aid in the discoverability of data and data sets by libraries and research portals. Agencies should allow for and encourage subject-based repositories and build on the models of successful discipline-based data models such as ISCPR, GenBank, Dryad, and others.

Comment 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?

There is no magic guarantee that when a data set is created, there is an immediate recognition of the benefits of the long-term stewardship and dissemination. The same could be said for books and journals. At least with journals there are metrics for journal use and citation patterns. Similar metrics may emerge over time for data. Stakeholders, data creators and librarians/data managers can play a collaborative role in determining the standards for preservation. It is understood that large amounts of raw data may not be useful or understandable, so standards must be set to describe the data in its most usable form.

Current models of stewardship and dissemination that merit consideration include the approach provided by the NIH-mandated deposit of peer-reviewed research articles in PubMed Central. Their deposit requirement and sufficient program funding have made this repository successful. A comparable repository for USDA-funded research could be managed by the National Agricultural Library (NAL) as an expansion of the existing NAL Digital Collections (NALDC). The advantages of a centralized repository include better control of the deposit process, author compliance, and consistent metadata applications. Funding agencies managing a smaller grant portfolio may have a more difficult time supporting a separate repository, so centralization would benefit these agencies. Centralization also minimizes issues of interoperability, consistency and redundancy. Many universities maintain an institutional repository and could help facilitate required deposits within the institutional site or a centralized repository. Even with clearly articulated standards, achieving full interoperability across many repositories may be a challenging goal. Although the examples above relate to the management of articles, a similar system could be created for various types of data.

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

Stakeholders can best contribute to the implementation of the data management plans by being involved in the creation and use of the plans. Professional societies can play a role by providing leadership in defining data structures and types pertinent to the scholarship of their disciplines. Data creators may need encouragement to realize how their datasets may be of value to others. Funding agencies and institutions should promote and reward exemplary projects and best practices.

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

Allow costs of data storage, management, preservation, etc to be included in grant applications and funding. Funding support for national libraries and data repositories should also be provided.

Short-term funding would get projects such as iPlant off the ground, but limited funding for the beginning of projects would not be sufficient to sustain a project long-term. "iPlant is a community of researchers, educators, and students working to enrich all plant sciences through the development of cyberinfrastructure - the physical computing resources, collaborative environment, virtual machine resources, and interoperable analysis software and data services– that are essential components of modern biology" http://www.iplantcollaborative.org/about. It would be critical to have ongoing funding towards digital data management and accessibility included in the budget of national libraries (NAL, LOC) and agencies such as USDA.

Comment 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?

The national libraries, which are mandated with providing stewardship of printed scholarship and given suitable resources, should play a similar role in managing data and electronic information.

Comment 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?

No comment.

Comment 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?

There should be policies and development of persistent identifiers that would allow the tracking of provenance, ensure data integrity, and contribute to successful citing of data and attribution to the authors/creators of the data. Creative Commons licensing may provide additional support for this effort.

Comment 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.

The national libraries have the requisite skills, experience and mandate to define and implement the standards that must be put in place to create an interoperable data repository system. The minimum metadata elements for describing bibliographic information are currently well-defined by the Dublin Core metadata standard. These elements can be readily derived from publisher data and incorporated as part of the deposit. Adherence to this standard, as well as the OAI-PMH standard for metadata harvesting, will facilitate the sharing of data from multiple repositories and lead to discovery by the public. Metadata standards are critical for describing publications and data within a repository, but institutions are also faced with the added challenge of increasing access to those resources. Resources must be highly discoverable and understood within a larger context of scientific data and research. For that to happen, several things must occur: 1) the advanced support of author disambiguation initiatives, such as ORCID, which "aims to solve the author/contributor name ambiguity problem in scholarly communications;" 2) a general mandate requiring federally funded authors to identify their funding source when submitting publications to a repository; and 3) the development and support of Semantic Web technologies that allow for the re-purposing, reuse, and analysis of publication and other data. By design, Semantic Web technologies are machine-readable; continuing to encourage the development and accessibility of these technologies would allow for flexible re-purposing of data, regardless of the model - centralized, decentralized, or mixed-model - chosen by Federal agencies.

Comment 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?

The National Information Standards Organization (NISO) provides an example of a successfully improving procedures and streamlining standard development processes. This has resulted in a reduction of time spent releasing consensus documents and sped up the process for launching new initiatives. Library of Congress' Cataloging in Publication Program (CIP) offers a detailed explanation of the process on their website, including pre- and post-publication. National Institute of Standards and Technology (NIST) has also been instrumental in producing effective standards for the U.S. Their process involves working with cooperative programs and partnering with 1,600 manufacturing specialists and staff at locations around the country. Details regarding these organizations and their role in standards development can be found at their respective websites. (http://www.niso.org/; http://www.loc.gov/publish/cip/; http://www.nist.gov/)

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

Federal agencies could work closely with an organization such as the World Wide Web Consortium (W3C), an international community in which member organizations, staff and the public collaborate on the development of Web standards. Another possibility is The InterNational Committee for Information Technology Standards (INCITS), a forum for information technology developers, producers and users for the creation and maintenance of IT standards. A third option is International Organization for Standardization (ISO), the world's largest developer and publisher of International Standards. ISO is a network of the national standards institutes of 163 countries.

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

There should be policies and development of persistent identifiers that would allow the tracking of provenance, ensure data integrity, and contribute to successful citing of data and attribution to the authors/creators of the data.