Thu 1/12/2012 9:14 AM

Response to Request for Information: Public Access to Digital Data Resulting from Federal Funded Scientific Research

Dr. Karen Cole, Director, and library staff
January 12, 2011
kcole@kumc.edu
Archie Dykes Library of the Health Sciences, University of Kansas Medical Center
Kansas City, Kansas

(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 policy should require that data and calculations resulting from federally funded research be deposited in open, curated digital archives and released to the public. We suggest that policies follow the model of the Bermuda Principles.

In particular:

- Automatic release of small amounts of data (24 hours)
- Immediate publication of finished collections of data
- Free availability in the Public Domain, clarifying that no licenses are required in order to get access to the data, make use of it, create derivative works, redistribute and reorganize the data.

Rights of use for data should be clearly stated using common, successful mechanisms such as The Open Data Commons licenses: http://opendatacommons.org/licenses/ and The Creative Commons Zero Waiver: http://creativecommons.org/publicdomain/zero/1.0/

Moreover, the data and calculations should be accompanied by provenance and descriptive metadata. The metadata should also reference resultant publications.

Publicly funded data restricted behind a publisher's paywall should not be an option. This type of restrictive practice prevents scientific discovery and progress.

(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?

It is important to note that data sharing and archiving is already carried out in many different ways without intellectual property or patent conflicts. Nevertheless, there is general agreement

by publishers, researchers, and legal jurisdictions that data cannot and should not be copyrighted. Federal law should recommend or require that all data, calculations, and analysis of data be waived of copyright or licensed to allow reuse and modification. Creative Commons Zero, Science Commons and Open Data Commons License provide examples of such approaches.

Patent rights, including IP on materials/reagants, and privacy rights are different issues, and need not be waived along with a waiver of copyright (Dryad, http://wiki.datadryad.org/wiki/Terms_of_Reuse).

Publishers, e.g. journal publishers, should be allowed the choice of offering authors the ability to set embargoes on the release of data in accordance with embargoes on manuscript publication. Data repositories should establish mechanisms and workflows to support this.

Publishers should be allowed to expect that data repositories offer secure access to the data and calculations for editors and reviewers during the manuscript review process. Publishers and authors should be allowed to expect that data repositories suppress information about related manuscripts until the article has been published.

(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?

Embargo periods may differ depending on the discipline and the rate of discovery within the discipline. Less "volatile" disciplines should be allowed longer embargo periods if necessary. Embargo periods may be necessary to protect intellectual property, pending patents, the researcher or institution's interests, the funder's interests, or the public good. These exceptions will likely be more prevalent in some disciplines than in others.

Support for submission integration between journals and data repositories should be flexible enough to accommodate publishing, editing, and review workflows.

Metadata profiles must be flexible enough for different disciplines while still being interoperable. Dryad's implementation of Dublin Core Metadata Initiative Abstract Model (http://dublincore.org/documents/abstract-model/) is one example.

File formats for data and calculations will be different among disciplines. Federal agencies should require open, non-proprietary file formats whenever possible since they are more likely to be readable in the future. For example, a plain text file has a longer life than a propriety word processing format, and a file of comma or tab-delimited values has a longer life than a proprietary spreadsheet file format. ASCII text should be preferred over color, images, and other embedded objects which are difficult to migrate.

(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?

Each agency, each research project funded by an agency, working groups within the various disciplines, and the institutions responsible for long-term stewardship of data should work together to develop needs for retention, preservation, and long-term stewardship for their cases.

Agencies should make available funding and support for institutions providing long-term stewardship and dissemination.

(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 contribute to data management plan implementation by providing education, software, curation, metadata, storage, replication, and distribution.

Publishers, research communities, libraries, repositories, and institutions can contribute by working together to develop mutually beneficial workflows and submission integration between journals and data repositories.

We are a university library who has a close working relationship with our biomedical research community and with our university's information technology department. We support and contribute to software development in support of research (e.g. DSpace and BibApp). We create and manage author metadata, works metadata and content. We provide services to publicize research. We educate researchers about open access and copyright.

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

Funding mechanisms should provide for storage and dissemination of data. They should also provide resources and incentives for long-term preservation of data. Agencies must make available funding and support for institutions providing long-term stewardship and dissemination. This funding must occur at the agency level and not be parsed out in individual grants

(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?

Agencies must provide researchers with unambiguous methods for linking funding agencies, data, and resulting works.

(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?

Adopt clear licensing requirements for data that are easy for researchers to comply with.

Provide tools that reduce the burden of licensing, ease compliance, reduce duplication, and open the data for use and re-use.

Fund educational initiatives that promote the use of data and computational thinking within primary, secondary, and higher learning institutions.

We also refer you to Semantic Web: Revolutionizing Knowledge Discovery In the Life Sciences (Baker and Cheung, 2007. ISBN-13 9780387484365) which provides insightful, modern, and practical analysis of types of innovation possible and necessary when the right data, tools, and standards are available.

(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?

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data?

Linked Data and Semantic Web standards such as OWL and RDF.

Established discipline-specific or community-developed data standards. 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.

We also refer you to Semantic Web: Revolutionizing Knowledge Discovery In the Life Sciences (Baker and Cheung, 2007. ISBN-13 9780387484365) which provides insightful, modern, and practical analysis of types of innovation possible and necessary when the right data, tools, and standards are available.

(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?

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

Adopt simple minimal standards to allow for inter-operability and search that have buy-in and participation from a wide range of stakeholders. Optional discipline based standards could then evolve within those scholar communities.

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

Unique and persistent identifiers for publications and associated data are critical for linking the two. Identifiers should conform to modern namespaced URN/URI schemes. Examples of current successful schemes include DOIs, Handles, and PURLs. These schemes and standards are currently widely by publishers, content repositories, and digital libraries. The schemes have proven their utility in modern semantic web and linked data applications.

Access to identifiers and registries should be publicly available so that identifiers may easily be dereferenced and resolved to content endpoints via linking standards such as OpenURL.

Publications and data should be as atomistic as is economically feasible so that clear references and inferences can be made by human as well as by machine. For example, an OWL Reasoner and a human alike should be able to traverse the path from an assertion made within a publication to some element within a dataset or step within a calculation. In general, policies and practice should follow and build upon modern Linked Data practice.

In addition, please identify any other items the Working Group might consider for Federal policies related to public access to peer-reviewed scholarly publications resulting from federally supported research. Please attach any documents that support your comments to the questions.

Karen Cole
Director of Dykes Library
University of Kansas Medical Center