Wed 1/11/2012 12:24 PM
Response to RFI: "Public Access to Digital Data..."

Dear National Science and Technology Council's Interagency Working Group on Digital Data :

I am responding to your RFI as a US federally-funded structural biologist and a member of the International Union of Crystallography Diffraction Data Deposition Working Group and the International Union of Crystallography Commission on Biological Macromolecules.

The International Union of Crystallography Diffraction Data Deposition Working Group is actively working on the issue of archiving the raw data (diffraction images) in crystallography (in addition to summary data and crystallographic models; see http://forums.iucr.org/viewforum.php?f=21).

I give some comments on your requested information items below.

Sincerely,
Tom Terwilliger
Los Alamos National Laboratory

(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?

Long-term Federal support for:
A. International databases that preserve this data (example: The Protein Data Bank) B. Development of international agreements on data and metadata formats C. Development of software and procedures to make it rapid and easy for researchers to deposit their data and to retrieve data from these databases

(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?
—-

(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?

As much as possible, the policies should focus on the desired outcome , not on the mechanism of achieving it.
Policies should be minimized and general. Instead , the focus should be on support for achieving the goals (making deposition and extraction of information easy for the researchers)

(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?

In the structural biology community there was a most lively discussion on this point on the CCP4 bulletin board (see
https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1110&L=CCP4BB&F=&S=&P=323904
for example). The discussion showed that even within a community there is a great difference of opinion on what needs to be archived. It also showed that these costs and benefits can be discussed in a thoughtful way within a community. The CCP4bb discussion revealed some variations of opinion but >50% voted to archive raw data rather than processed data. I expect more will support raw data archiving via local repositories.
It is less clear how this can be done between communities, however.

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

The international structural biology community has shown that a research community can work with publishers to develop standardized formats for presentation of data (Rapid structural reports were developed for Protein Science, Acta Crystallographica Section F, and Journal of Structural and Functional Genomics, among others to respond to the need for short reports on macromolecular structures.) This is being extended to working together on coordinating deposition of data with publication.
Research institutions and universities can contribute by providing local repositories for data archiving and retrieval. As some raw datasets can be very large, transferring them via the network can be slow so that local data storage may be most effective.

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

Repositories (such as the Protein DataBank) that store data in perpetuity and that show a clear benefit for the community should (continue to be) funded at a high level.
Research institutions and universities that store data locally and make it available should also receive funding to help cover the cost of doing this.

(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?

With clear policies and with procedures to walk researchers through any compliance, the burden can be minimized. However it simply will take some effort on the part of researchers to provide their data in a way that others can use.

(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?
--

(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?

A basic requirement is that a simple and clear path to those who produced
the data should always be available. The user of a piece of data should
always have an easy way to know who produced it.

Secondarily it would be helpful to have general policies that indicate
that credit should be given. For example a granting agency could require
its grantees to check off a box saying that they will give attribution to primary data producers in their publications that use this primary data.
Third, journals that use Supplementary Materials sections should be expected to include these sections in indexing and citation analyses.

Standards for Interoperability, Re-Use and Re-Purposing

(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.

In structural biology, the mmCIF (macromolecular Crystallographic Information File) is the community-driven standardized data format. This format is used by the PDB to represent and transfer data and it is likely to soon be the standard for communication between researchers as well.
Crucially, the mmCIF file contains metadata about the experiment carried
out. It uses a standardized extendable dictionary of terms. This data
structure and associated software is sufficiently developed so that many aspects of the experiment can be re-analyzed automatically , though it is not yet possible to automatically fully interpret an experiment.

(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 Protein Data Bank (www.pdb.org) developed standards along with the International Union of Crystallography (IUCR).
The key characteristics were (1) international involvement (IUCR) in development of standards (2) a funded central repository (the PDB), (3) a very active structural biology community using the central repository

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

The NIH and NSF should continue strong financial support the PDB, which is a highly effective organization that carries out international coordination of digital data standards for structural biology. Existing nternational committees working on this issue should be brought into the discussion. For example the International Union of Crystallography and The International Council for Science Ad -hoc Strategic Coordinating Committee on Information and Data on these issues (see http://forums.iucr.org/viewtopic.php?f=21&t=63#p175)

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

This is a highly important goal. One practical implementation is the general use of the Digital Object Identifier (http://www.doi.org) as a reference for data items and publications. The structural biology community is considering this for archiving of large data items (TB size) so that they can be stored locally.