Thu 1/12/2012 4:27 PM

Response to OSTP RFI on Public Access to Digital Data Resulting From Federally Funded Scientific Research

Access to primary research data is important for the advancement of the scientific enterprise. It facilitates the validation of existing observations and provides the raw materials to build on those observations. These benefits, however, must be balanced against the burden to researchers of providing such access. The time and money required to provide access to research data are time and money lost to research.

The National Science Foundation recently enacted a blanket policy to require sharing of all data generated using their funds. But it is impractical and not essential to require researchers to share every piece of data they acquire. Sharing policies will have to be more specific about the types of data that will be useful to share and whether preliminary data sets are included. The NIH policy takes steps in this direction. In addition, funding agencies will have to provide the mechanisms for sharing large data sets. Individual scientists cannot be expected to provide these mechanisms. Sharing policies will also be ineffective unless they are enforced.

In biomedical research, there are numerous instances in which the members of the research community have determined that a particular type of data would be useful and necessary to share. These include gene sequences, protein structural data, and gene and protein expression profiles. In these cases, the community united to standardize the structure of the data and its associated metadata, and federal agencies created centralized repositories (or funded their creation) to facilitate deposition, promote discoverability, and ensure the longevity of the data. Funding agencies will have to maintain an ongoing dialog with the research community to decide what other types of data will benefit from standardized repositories, and they will have to fund the creation of those repositories.

In addition to standardized data and metadata structures, two other elements are essential for the success of any repository: customer service, to make it as easy as possible for researchers to deposit data in the correct format; and curation, to ensure that data are properly formatted and tagged, and to monitor any crowd-sourced tagging through Wiki applications. Proper tagging of data with accession numbers and/or digital object identifiers will help to ensure the longevity of links to those data.

Access to research data is important, but data without context are not useful to third parties. Why were the data obtained (to answer what question)? How were they obtained? What was the interpretation of the data by the person who obtained them? These questions closely mirror the introduction, methods, and results/discussion structure of primary research articles, and the most obvious way to provide context to data is through a scientific publication. Thus, the primary burden of enforcing deposition of data into repositories has fallen on journals.

Biomedical research journals require that certain types of data such as nucleotide sequences, protein structural information, or protein/gene expression profiles be deposited in repositories hosted and curated by (or at least funded by) funding agencies. This enforcement process, which evolved with the development of these repositories, functions for specific types of data underlying published research articles.

In rare cases, a publisher may develop its own repository to fill a gap in providing access to a type of data that is prevalent in a particular journal. One of the journals published by The Rockefeller University Press, *The Journal of Cell Biology (JCB)*, publishes a large amount of microscopy image data. In the absence of a standardized, international repository for this type of data, the journal developed the JCB DataViewer – a browser-based application for viewing original, multidimensional, microscopy image data.

The imaging community is coming to some consensus about the data and metadata structures necessary for sharing and archiving microscopy image data, but most of these data reside on desktop computers in proprietary file formats (PFFs) that cannot be shared. The JCB DataViewer uses an interpreter to convert those PFFs into a standardized format and display them over the internet. It can also host the complete data sets from high-content imaging screens. Deposition of image data by authors into the JCB DataViewer is encouraged but remains voluntary.

The JCB DataViewer is currently used for original image data supporting articles published only in the *JCB*. We hope that it will serve as a prototype for the development of a larger repository for images published in any journal. But it is not reasonable or sustainable for an individual publisher to undertake such an expansion. This must be done by national or international funding agencies.

Funding agencies have relied on journals for enforcement, and, indeed journals are in a strong position to place requirements on authors before publishing a paper. However, journal publishing is competitive, and journal editors may be reluctant to afflict potential authors with additional demands that they may consider burdensome for fear they will submit their papers to another journal with less stringent requirements. Most journals will establish an access policy to a particular type of data only once a standard for sharing that data type has been set and an expectation of compliance has been established in the research community. But even then, there will be variability in the stringency of enforcement.

For newer standards (for example, high-content image screens), there will be great variability in requirements by journals until an expectation of compliance has been established. Given these variabilities, the funding agencies should monitor published data for compliance with sharing policies, and they should not rely solely on the journals. Enforcement of sharing polices for data that have not resulted in a publication will fall completely on the funding agencies. They will have to decide what types of data to monitor (they can use editorial policies of biomedical research journals as examples) as part of the grant application/renewal process, and they will have to create a monitoring step in that process.

It will be vital to provide context to unpublished data by ensuring that sufficient metadata are associated with the data for a third party to understand their origins (and to recognize that they are unpublished, and thus the methodology has not been vetted through peer review). Funding agencies will also have to develop policies about the timing of data release to the public. For data underlying a published research article, it is easy to set such a policy – the date of publication. For unpublished data, sufficient time will have to be provided to license data that may have commercial value. Funding agencies will have to monitor licensing terms to ensure that reuse by non-profit institutions is allowed.

Blanket policies regarding sharing of primary research data sound impressive and progressive, but they are neither practical nor enforceable. Funding agencies need to be specific about the types of data that they expect to be made public. Relevant scientists must be engaged to develop these standards, and funding agencies must provide the mechanisms for applying them.

Mike Rossner, Ph.D
Executive Director
The Rockefeller University Press

These comments are the opinion of the author and do not necessarily reflect the position of The Rockefeller University.