Thu 1/12/2012 9:30 AM
Response RFI: On Public Access To Digital Data Resulting From Federally Funded Scientific Research FR Doc. 2011-28621

Effective reuse of data resulting from scientific research is a multi-faceted challenge that goes beyond simply archiving and distributing data files. To effective build upon the prior work of their peers, researchers must be able to both find and interpret relevant data:
voluminous data archives will be of little value without specific clear and informative metadata and tools that leverage that metadata to help identify data of interest.

The generation of this metadata presents significant challenges. Although the success and evolution of metadata formats like the MIAME model cited in the RFI provides examples of best practices, these models suffer from several deficiencies that limit their impact. So-called "minimal information" models like MIAME are, by definition, limited in their expressiveness. Effective data sharing might require metadata that goes significantly beyond the baseline "minimal" description. However, the generation of more fully-descriptive metadata is a time-intensive task that is often not well-supported by existing data management tools. This difficulty is compounded by a lack of incentives: data annotation is most often the responsibility of the data generator, who may see this task as a cumbersome overhead requirement with little direct value-added.

A combination of data models, annotation tools, and search tools that leverage those annotations is needed to address these shortcomings. To be successfully and widely adopted, these tools must be designed to be well-integrated with existing tools and work practices.

Responses to Specific questions:

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?

Scientific agencies should take several steps to encourage public access and preservation:

* Provide specific requirements for preservation and publishing of public access data
* Identify specific data models and tools to be used for various data types
* Promote the development of more extensive and usable tools for annotating and finding research data
* Support the development of tools that promote best practices for archiving and managing data
* Require specific data sharing plans and dedicated resources in appropriate funding rewards

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?

Continuing embargo policies that provide researchers with the opportunity to use data for publication and to apply for patents should be promoted and adopted to the need of specific communities, particularly with respect to delays relative to publication, patent, or other trigger time points.

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?

Engagement with appropriate research organizations, discipline-specific workshops, and additional RFIs can be used to understand the needs of specific communities and to plan accordingly.

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

Preservation and archiving costs could be considered as "overhead" costs that would go "above the line" and therefore be included above and beyond current funding limits.

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?

Standard guidance for reporting data reuse, and community recognition might help with attribution and credit. For examples, effective reuse of data -either in reusing data from others or having one's own data reused by others - might be considered as a positive factors during grant reviews.

11. 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.

Scientific ontologies such as those curated by the OBO Foundry and the NCBO provide well-defined semantic models that encourage interoperability.
The use of these ontologies to publish scientific data as linked open data should be encouraged.

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

Unique identifiers for both publications and data sets, along with tools for using those identifiers, might be used in combination with linked open data on both publications and datasets, to support this linkage.

Harry Hochheiser
University of Pittsburgh
Department of Biomedical Informatics