Thu 1/12/2012 10:59 PM
Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

Note – A previous draft was sent to this e-mail and should be disregarded. This is the correct version.

Name/Email

Leslie D. McIntosh, PhD, MPH

Jon Corson-Rikert
Director, Mann Library IT Services

Affiliation/Organization

Washington University in St. Louis (LMc)

Cornell University (JCR)

City, State

St. Louis, MO

Ithaca, NY

General Comment

Giving researchers credit and recognition for work they do with data in the form of career advancements and awards will do the most to assure public access to and preservation of digital data. While Federal policies can further the requirements for digital data preservation, discoverability, and access, I believe it will be data scientists themselves that will move this effort forward.

Comment 1

Federal policies can best foster reuse of data by encouraging the adoption of open, interoperable standards for data exchange, notably Linked Open Data (http://linkeddata.org). Linked data reduces the barriers to sharing data and encourages the adoption of ontologies to more clearly express what the data represent.

Comment 2

Reporting requirements for grants could be modified in coordination with data management plan guidelines to require reporting a permanent web address for data, the nature of any access restrictions, and all significant contributors to the data.

Comment 3

Federal agency policies on the management of data should encourage the development of standards for documenting the content of digital datasets using the classifications and terminology of each discipline, ideally through non-proprietary ontologies and controlled vocabularies developed by members of the discipline itself. If datasets within any discipline are documented with explicit references to established ontologies and authoritative databases, the effort to accommodate differences among disciplines can be addressed at the discipline rather than the individual dataset level.

Comment 4

Archivists have developed policies and procedures to assess, and periodically re-assess, the significance of the artifacts they preserve, in large part because it may not be possible to predict the uniqueness, utility, or quality of artifacts until some time has passed. Agency policies should factor in re-assessment of the costs and benefits of continuing to preserve and/or migrate forward digital data, using experts representing the original domain and the most likely domains for data reuse.

Comment 5

Stakeholders can best contribute to the implementation of data management plans by promoting their adoption and encouraging the evolution of review standards through experience with both the costs and benefits of different levels of access and different investments in preservation.

Comment 6

The biggest challenge with current funding mechanisms is the time shift between the period of the award and the need for preservation. Data management plans could require an escrow process to set aside grant funds and assure access to them through a rolling funding model that would use current research funds to pay for the continued preservation of datasets still deemed worthy of preservation.

Comment 7

Datasets can be peer-reviewed much like journal articles. Standards for the review process could be developed; however, there are very few guidelines given when reviewing manuscripts, so the

process could be a very similar.

Datasets can be peer-reviewed much like journal articles. Upon submitting a dataset, it could be made available in a repository with information indicating the level at which the data have been reviewed (e.g. none, manually curated, used by independent source to replicate other findings).

Comment 8

Marketplaces for data and related services have been emerging on their own; federal funds could help support basic infrastructure costs for data registries and to establish workable policies and at least minimal subsidies for preservation of digital data when accompanied by viable business plans.

Comment 9

For individuals and groups to be given appropriate attribution and credit for data used, the data must be identifiable and discoverable, and metadata sufficient for ready discovery must be created for datasets and disseminated in a citable fashion. Once the basic mechanisms for discovery of datasets are in place on the Web, then the data owners can be cited in the same fashion that grants are now cited in publications. Journals should require authors to cite data sources and the authors/curators of the data in order for a manuscript to be published.

Comment 10

Use of any non-proprietary controlled vocabulary and explicit references to accepted international standard scientific and other disciplinary-focused databases will go a long way.

The ANDS/VIVO ontology (http://blogs.unimelb.edu.au/vivoands/2011/07/06/the-vivo-ands-ontology/) is a lightweight extension to the VIVO ontology (http://vivoweb.org/ontology/core) used to submit metadata about research datasets in university collections in the format required for the Australian National Data Service (http://www.ands.org.au/resource/rif-cs.html).

Comment 11

The Open Biological and Biomedical Ontologies (http://obofoundry.org) are examples of standards that have been developed through an open, collaborative process across several disciplines in the life sciences. Open processes that maintain quality standards and encourage iterative improvement through consensus have a higher likelihood of adoption and ongoing maintenance.

Comment 12

A promising way to promote effective international coordination of digital data standards would be to fund a tool that allows for open adoption and development for data, similar to what VIVO (http://vivoweb.org) has done for linking researchers.

Additionally, federal government grants should encourage US Citizens to travel abroad to professional meetings using government grants. This allows personal connections to be made, which facilitate future collaborative work.

Comment 13

Library and government repositories can encourage and, when appropriate, require submission of datasets associated with publications through modification of the publication submission and review process.

Leslie D. McIntosh, PhD, MPH
Center for Biomedical Informatics | Washington University School of Medicine