Thu 1/12/2012 6:22 PM
RFI:Public Access to Digital Data Resulting from Federally Funded Scientific Research

[Assigned ID #]
[Assigned Entry date]
Name/Email
Morteza Gharib, Vice Provost / vpr@caltech.edu
Affiliation/Organization
California Institute of Technology
City, State
Pasadena, CA 91125

Caltech is a PhD university employing 922 principal investigators whose research funding comes largely from 6-10 different federal agencies. In addition, Caltech is committed to education and recognizes its profound obligation toward public dissemination of its research results ideally unfettered by the demands of commercial profit so that learning and discovery, two major pillars of the enterprise, will thrive. The global network provides the means to ensure maximum access for uptake of new knowledge via electronic distribution of publicly funded research results. Therefore, Caltech urges and supports action to require prompt public access to results of all government funded research.

In response to the request for information from your office released on November 3, 2011 on the topic of public access to peer-reviewed scholarly publications resulting from federally funded research we offer the following comments.

922 includes professorial faculty, research faculty and postdoctoral scholars.
The Dept. of Defense is counted as one agency.

Comment 1
The National Science Foundation and National Institutes of Health require data management plans for all current grants. This is an important first step in creating incentive to improve access to and reusability of data generated by federally funded research. Broadening the data management requirement across the full range of unclassified, federally funded scientific research will significantly improve the efficiency and productivity of the American scientific enterprise.

Standardized data deposit policies across the spectrum of funding agencies will go a long way toward ensuring ease, and therefore consistency, of deposit. Where possible, discipline-specific data repositories (e.g.; ICSPR, IRIS, CDIAC, PDB) should be utilized to create content-rich destinations for data seekers, both human and machine. These databanks will facilitate discovery, access and preservation activities, moving past the widely varying interpretations of data access policies of individual primary investigators.

Comment 2
Existing publisher copyright policies and attendant business models for peer reviewed publications are ill suited to data. Publishers focus on dissemination and licensing for all of the content to which they hold rights. Preservation and universal access are not core values or business functions, especially in light of the vagaries of business mergers, consolidations, divestitures, and bankruptcies.

Data, as facts, are not subject to copyright. The American Chemical Society, in 2010, ceased to claim exclusive rights to supplementary material published in its journals. Dryad provides data repository services for 100 journals.

Creative Commons CC-BY and CC-0 licenses represent excellent consistent and standard starting point from which to build licenses for data, acknowledging potential copyright issues, while maximizing the prospects for both people and machines to build services that maximize discovery and access.

A rational argument for embargoes can be made to acknowledge the unique efforts of the primary investigator or original scientific team. The efforts of primary investigators should be acknowledged through a term of exclusive access to data (i.e.; throughout the publication process) allowing time for data to be thoroughly processed and analyzed. Primary investigators and their teams should be given the first opportunity to make discoveries and produce publications. It is important to note that a right of first publication does not preclude the deposit of the data into a certified data repository even during an embargo period, particularly to initiate archiving activities.

Comment 3
Different scientific disciplines offer a broad spectrum of requirements for data management practices and policies. There are some basic conditions for archiving and preservation that apply regardless of scientific discipline. Datasets require significant documentation (e.g.; equipment and equipment settings, provenance, data processing) to make comparisons and combinations with other datasets valid. Identifiers, fixity information, Persistent URLs are a few of the critical pieces of data infrastructure supporting archiving and preservation that should be applied to all scientific datasets.

Each scientific community is best qualified to address its specific data management needs. However that perspective tends to be narrowly conceived and minimally applied. Not all domain scientists may be explicitly aware of data management issues and needs. Basic requirements from funding agencies offers the opportunity to gain the necessary attention from the researchers. A major point is that they are held accountable to the public access concept and focus on scientifically defensible criteria for any deviation from full and complete access for humans and machines.

Comment 4
Different types of data and the needs of specific scientific communities will introduce different relative costs and benefits. It may be useful to consider needs as they relate to baseline services that apply across all scientific disciplines (e.g., archiving) and secondary services (e.g., discovery and specialized query capabilities). Agency policies should accommodate the relative emphasis between these two categories of services in different disciplines, as it relates to the distribution of costs and benefits across the full array of stakeholders. Federal agencies might wish to fund libraries and museums to develop the data archiving capacities, yet expect those libraries and museums with their parent organizations to bear the long-term costs given their cultural memory missions over the long haul. Agencies could provide seed funding for preservation while provide ongoing funding to a scientific community to develop secondary services.

Comment 5
Data management practices within scientific communities are currently diverse. This is not necessarily a bad situation. Some scientific disciplines already benefit from ongoing, centralized data repositories (e.g.; ICSPR, IRIS, CDIAC, PDB). It would be negligent not to build upon this already existing infrastructure. Of significant importance is the need for these data repositories to demonstrate their ability to preserve data functionality over time, not just assure the community that preservation is being done. Archives, libraries and museums have an extensive track record with these functions and could serve the essential purpose of developing or implementing frameworks that thoroughly test and certify assertions of preservation. However the parent institutions need to participate in the strategic re-ordering of resources to meet the new needs. With a clear set of requirements, it will become possible

to identify how various stakeholders can implement data management plans, noting that these roles will vary by scientific discipline or community.

Comment 6
The real costs of preserving and making digital data accessible are legitimate and important costs of the scholarly infrastructure necessary to support research. Grant proposals will need to include funding for data curation for preservation similar to the usual practice of providing funds for publishing. That being said, funding of cyberinfrastructure to create community-based data repositories, for scientific disciplines where the cost benefit ratio supports the notion, should not be ignored. Some funding in research grants is necessary for the preparation for data sharing, but the bulk of the cost of preservation, discovery, and dissemination services may reside in the operation and maintenance of discipline-specific data repositories.

Comment 7
Workflows of library-based or community-based data archives are implementable and effective platforms to ensure compliance. Plus such community based services are more likely to offer the researchers a consistent and reliable set of rules over time. Persistent identifiers (ORCID, doi, Persistent URL) and appropriate licenses (CC-By, CC-0) represent critical mechanisms through which compliance and verification can be automated thereby reducing costs. NIH currently requires PMCID reporting for all articles published under a grant in progress reports and final grant reports. A similar reporting mechanism for data deposition would flesh out the NIH data management requirement and could be generalized to other agencies, enabling verification mechanisms for data management requirements.

There are two activities that require the researcher's interaction with a third party: proposal submission and publication submission. Proposal submission is the place to insert data management requirements, as evidenced by the NIH and NSF data management plan requirements. Publication submission is a key point at which investigators have validated the datasets which are relevant and trustworthy for sharing via publication and deposition in data repositories. By embedding appropriate data management planning with proposals, data deposition with publication, the use of appropriate licenses (CC-By, CC-0) and cyberinfrastructure requirements into these workflows, the prospects for efficient compliance and verification, through standard grant reporting mechanisms, are heightened considerably. Researchers will resist any additional burden. However their institutions or their communities must develop capacity to support and implement data management plans, so those "burdens" can be shifted to support infrastructure (libraries, museums, disciplinary data repositories) that view such activity as part of their core mission.

Comment 8
Federal agencies could stimulate the development of discipline-based and institutional data archives that support discovery, download, and preservation. A uniform mandate across agencies to make data freely available through such archives, under appropriate licenses (CC-By, CC-0), would encourage the growth of such archives and, by extension, the development of APIs to allow individuals and machines to develop new capabilities and services. This type of open system would facilitate new business opportunities even by smaller businesses. The licensing arrangements (CC-By, CC-0) would be critical to ensure that no single entity or group has an exclusive right to generate such new business opportunities.

Comment 9
One of the most important components is author and institutional identifiers (e.g., ORCID) that would support developing attribution and credit processes. Machine-based access argues for CC-0 licensing. The seismology community already acknowledges use of datasets deposited in IRIS. While the original

primary investigators are not necessarily credited, the community benefits from the ability to compare and compile datasets from a wide variety of seismic arrays.

Comment 10
Barcode of Life Data Systems (BOLD) is a community-based data repository supporting evolving tools and data standards that aid in the collection, management, analysis, and use of DNA barcodes. While there are many community-driven data standards for scientific data, most of them deal with interoperability or sharing rather than archiving or preservation. There are too many discipline specific efforts to list. Basic requirements for more x-disciplinary sharing need to be developed.

Comment 11
There are examples of successful data standards development efforts within various domains (e.g.; Structural Biology Knowledgebase (SBKB) of the Protein Structure Initiative (PSI), FITS (Flexible Image Transport System) in astronomy, FGDC (Federal Geographic Data Committee) geospatial metadata standards). In each of these cases, there are undoubtedly several, perhaps common, characteristics or reasons for the success of the effort (or alternatively reasons why such efforts did not succeed in other cases).

Comment 12
While there exist groups that work in this area (e.g.; the National Academies' Board on Research Data and Information (BRDI), CODATA (Committee on Data for Science and Technology)), it would be helpful for Federal agencies to support community-based efforts that connect nodes of data infrastructure development activities, both disciplinary and institution-based repositories. For example, the European-based EUDAT project has already reached out to projects within the U.S. regarding a Data Access and Interoperability Task Force (DAITF) along the lines of the Internet Engineering Task Force. NIST could be helpful in this context supporting the development of a "data grid" that would operate in a similar manner to the power grid.

Comment 13
There is widespread consensus within the research community that it is essential to link publications and underlying or associated data. Traditionally, the linkage, if it existed, has been through supplementary material in journals. The Society for Neuroscience, for example, no longer accepts supplementary materials for distribution with articles in *Journal of Neuroscience*. The peer-reviewed publication is viewed as the final "snapshot" of the research process and outcome. Dryad has created partnerships with 100 journals, from a wide variety of publishers, to host, preserve, and provide long-term access to the data underlying formal publications. A key consideration from a policy, practices and standards standpoint is a requirement to use persistent, unique identifiers (ORCID, doi, Persistent URLs/handles) for publications, data, authors, and any entity of interest. These identifiers not only bolster the linking and attribution of publications and data, but also help foster the re-use and development of new services by people and machines. While there are multiple identifier schemes, at this point, perhaps the most important policy decision would be to require using persistent identifiers instead of relying upon existing mechanisms such as individual investigators' website URLs, a step which *The Astrophysical Journal* is in the process of undertaking. Agencies may have to specifically require that a national identifier scheme be used. Many researchers do not understand the difference between a url and a supported identifier that resolves to the currently active url.