Attn:
Office of Science and Technology Policy
725 17th Street, Washington, DC 20501

RE:
OSTP RFI: Public Access to Digital Data Resulting From Federally Funded Scientific Research

Massachusetts Institute of Technology's comments on Federal Register Document 2011-28621

Claude R. Canizares, Vice President for Research and Associate Provost;
Ann J. Wolpert, Director, MIT Libraries /
Massachusetts Institute of Technology
Cambridge, MA

The Massachusetts Institute of Technology (MIT) appreciates the opportunity to comment on approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally funded scientific research. The comments below, in concert with our comments specific to Federal Register Document 2011-28623 (OSTP RFI: Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research), affirm MIT's belief that public access to unclassified research and the data collected as part of research funded by Federal science and technology agencies is a topic of substantial significance to this institution because MIT's mission includes a commitment to generate, disseminate, and preserve knowledge. This commitment carries particular weight when the new knowledge generated at MIT flows from federally funded research. We address each question from this RFI in turn:

(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?

Comment 1: As articulated in MIT's response to the OSTP RFI on Public Access to Federally Funded Research last January, MIT believes that a key first step in providing access to the research results of federally funded research is to expand the goals of NIH's public access policy to other federal funding agencies. Providing access to the data without the context of the corresponding, peer-reviewed research results would be short-sighted and create an unnecessary barrier to other researchers and interested parties in being able to interpret and re-purpose the data in productive ways. In addition, the research data resulting from federally funded research should be subject to a data management and sharing policy, similar to either NIH's current policy, http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html, or the NSF's, http://www.nsf.gov/eng/general/dmp.jsp. To generate the most benefit it will be important to avoid unnecessary proliferation of varying requirements. Critical to the success of the NIH's policy has been the infrastructure provided by NIH through the National Library of Medicine's National Center for Biotechnology Information (NCBI), http://www.ncbi.nlm.nih.gov/. It's also important to note that there are other efforts currently supported through collaborations between federal agencies and research institutions for access to other important disciplinary domains in the sciences, e.g. http://cdp.ucar.edu/, and social sciences, http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/fifty/factsheet.html. Leveraging the infrastructure created by the NCBI and others to include other research domains rather than suggesting that each federal agency funding research create their own infrastructure should be seriously considered.

(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?

Comment 2: Existing copyright laws already provide a framework for protecting the IP interests of all. Deploying open licensing and/or waiver protocols for data made available under public access policies would expand the benefits of openness to allow for more innovation, by allowing for data mining and creating new works from existing works. However, because the IP landscape is so complicated, significant educational efforts are needed and will need to continue to ensure these stakeholders understand these complex issues particularly as they relate to data. Useful background on this topic can be found at http://sciencecommons.org/projects/publishing/open-access-data-protocol/, http://creativecommons.org/about/cc0, http://www.dcc.ac.uk/webfm_send/332, and http://www.nature.com/nature/journal/v461/n7261/full/461171a.html.

Of particular concern is a scenario where data is transferred to publishers who then assert copyright due to value added services, e.g., extended and/or normalized descriptive information. While services of this type should not be limited, it will be important to insure that the original data resulting from unclassified federally funded research is still publicly available. It is important to note that MIT, like most research institutions, has explicit rules regarding compliance with HIPAA, http://web.mit.edu/committees/couhes/procedures_healthcare.shtml.

Whatever policies are developed, consistency of requirements is the key element that will allow federal agencies to maximize the benefits of their public access policies. Based on our experience supporting the NIH Public Access Policy and the MIT Faculty Open Access Policy, compliance will rise directly with convenience to the author. For this reason, common procedures, requirements, and processes should be established across all funding agencies whenever possible.

Another key factor in protecting the IP interests of stakeholders is adopting an agreed upon standard for citing data. This will enable the easy reuse and verification of data, allow the impact of data to be tracked, and create a scholarly structure that recognizes and rewards data producers, http://datacite.org/whatisdatacite.

(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?

Comment 3: Working with disparate stakeholders to develop a minimum set of core metadata for all datasets along with an API for standards based data exchange will help ensure a level of interoperability and discovery across all disciplines. Also, these inherent differences mean that there is a need for flexibility in funding amounts for data curation, and a commitment by agencies to provide the necessary funds for the data curation. Again, consistency of requirements as much as possible is the key element that will allow federal agencies to maximize the benefits of their public access policies.

(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?

Comment 4: Flexibility in cost models for data curation will be necessary. The importance of having a data management plan that is adequately resourced prior to the beginning of the research is paramount to prevent unnecessary costs and the possibility of data loss during the life cycle of the research. In addition, there needs to be a clear understanding that the long-term stewardship of the research data, when appropriate, is a much deeper commitment than that of costs incurred during the life of the research project. For such cases the intent will be for the data to be preserved and disseminated well beyond the life of the particular project.

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

Comment 5: The current landscape is complicated by the heterogeneity of practice across disciplines, a lack of common standards, rapidly developing (and changing) tools for data management, and confusion regarding roles and responsibilities. Developing standard practices for data attribution and citation will be critical. Collaboration among all stakeholders will also be necessary to minimize costs and maximize data sharing. Raising awareness for all those involved in the enterprise is necessary. Most research libraries and institutions are now involved in advising researchers when needed on best practices for data management, and many are working to develop tools to support the long-term preservation and dissemination of research data within the parameters of intellectual property concerns. Examining how existing federal infrastructure, e.g. NCBI, can be leveraged to support long-term stewardship and dissemination of different types of data resulting from federally funded research will also be important to prevent a scenario where proprietary solutions are developed which might be counterproductive to the goals of public access over the long term.

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

Comment 6: Recognition that the costs associated with preserving and making digital data accessible beyond the life time of the research project is vital, and providing options – whether it is providing funds within the research project to cover this long-term cost and/or providing infrastructure to manage the preservation and accessibility to the data, e.g. NCBI – is an absolute requirement if this effort is to be successful.

(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?

Federal granting agencies will need to develop roles, responsibilities, and procedures to review grants awarded to ensure compliance. While requiring that individual responsibility for the management of the research data be assigned and made publicly known might be considered, it will be paramount to develop approaches that both make it easy for researchers to comply and add value to their research efforts. The successful implementation of data management plans from previous awards might also be considered when examining new grant proposals.

(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?

Comment 8: Encourage data creators to use the Creative Commons CC0 license whenever possible, http://creativecommons.org/about/cc0. Promote "success stories" that demonstrate the successful use of secondary data in advancing research and productivity. Establishing a new granting program to develop innovative tools for mining scientific research data should be considered.

(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?

Comment 9: Adopt standard practices for data attribution and citation, and require that these practices be required for funding and publication. Relevant initiatives currently underway are DataCite, http://datacite.org/, and ORCID, http://orcid.org/.

(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.

Comment 10: The Data Documentation Initiative, http://www.ddialliance.org/, is an example of an initiative that works collaboratively to develop and adopt standards across different disciplines and stakeholders. These standards enable machine usability of the data, as well as facilitate data documentation throughout the life cycle. Balancing discipline specific needs against the desire to have easy interoperability and repurposing of data will be challenging, and may require the adoption of a simplified core set of metadata standards for all data types. Also key will be the implementation an API for standards based data exchange.

(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?

Comment 11: Relevant examples are the Internet Society and its IETF, http://www.ietf.org/, the W3G, http://www.w3.org/standards/, the Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/pmh/, and the DOI, http://www.doi.org/index.html. Characteristics that have made these successful are that they are completely open and transparent, and that they typically require a proof of concept.

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

Comment 12: Work proactively with national and international standard bodies.

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

Comment 13: Persistent identifiers for data sets are critical. Also important will be enabling relationships between different versions of data sets to be made visible, something similar to the CrossMark service being developed by CrossRef, http://www.crossref.org/crossmark/index.html. Again, the DataCite initiative, http://www.datacite.org/, is a useful forum and resource to further explore these issues.

Ann J Wolpert
Director
MIT Libraries
http://libraries.mit.edu