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

RE: OSTP RFI: Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research

Massachusetts Institute of Technology's comments on Federal Register Document 2011-28623

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 the merits of Public Access to Peer-Reviewed Scholarly Publications Resulting from Federally Funded Research (and have also submitted comments in response to Federal Register Document 2011-28621 on Public Access to Digital Data Resulting From Federally Funded Research). Public access to archived publications resulting from 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) Are there steps agencies could take to grow existing and new markets related to the access and analysis of peer-reviewed publications that result from federally funded research? How can policies for archiving publications and making them publically accessible be used to grow the economy and improve the productivity of the scientific enterprise? What are the relative costs and benefits of such policies? What type of access to these publications is required to maximize U.S. economic growth and improve the productivity of the American scientific enterprise?**

Comment 1. Open access to scientific knowledge in the form of research articles encourages economic growth through generating opportunities to create products and services. In order to maximize the growth of existing markets and develop new markets associated with peer-reviewed publications from federally funded research, the entire corpus of articles that have derived from publicly funded research should be made available for public use and reuse. This means not only being able to read the articles, but being able to create new works from them, and being able to use mechanized tools to analyze them or derive data from them.

While making the articles openly available for reading would speed science, making the articles available for reuse would also enable the development of new services and products. In fact, Victoria University Economist John Houghton's 2010 study concludes that if all the major federal agencies were to make the research they sponsor openly accessible, over 30-year period this would yield $1B of benefits to the US, approximately 5 times the cost to provide the archiving. (http://www.arl.org/sparc/bm~doc/vufrpaa.pdf*)*

A study published by the National Bureau of Economic Research supports this analysis with real-world testing of how closed research reduces innovation. In examining genes sequenced by the private company Celera, which were not publicly available, vs. those sequenced through public effort and made publicly available, MIT Professor of Economics Heidi Williams found that closed access reduced "subsequent scientific research and product development outcomes on the order of 30 percent." (http://www.nber.org/papers/w16213.pdf). This study uniquely focused on whether "differences in scientific publications translate into differences in the availability of commercial products" and found "persistent negative effects" on the accumulation of new scientific knowledge from closed research. These negative effects continued even after the limits on access were removed.

Open access to research articles increases scientific productivity by increasing access and citations (see Hajjem et al 2005 as summarized in http://www.keyperspectives.co.uk/openaccessarchive/Journalpublications/American_Scientist_article.pdf), leading to quicker take-up of new ideas. In part, this increase in access and citation is driven by new and unexpected readers that take advantage of openly available research. At MIT, for example, where the faculty's open access policy has been in place since 2009, despite the small size of the collection (4500 articles) readers routinely and eagerly read and download papers that describe current MIT faculty research results. In contrast, paid-access journals have increasingly constrained readership because of high and rapidly escalating access prices. (At MIT expenditures for journal subscriptions increased 400% from 1986 to 2009, during which time the number of journals to which we subscribed actually declined, and overall inflation increased by just 96%). The closed, subscription-based model shrinks audiences, locking out not only traditional users who are priced out of the market but unanticipated new users as well, such as those not associated with large research universities.

While public access to publicly funded research will stimulate the economy, not all public access policies are created equal in terms of their ability to improve productivity and encourage growth:

Public access policies that call for immediate open access are more powerful than those that allow for delays.

Public access policies that include open reuse rights (not simply 'read-only' access) allow the mining of information and encourage the creation of new tools and the use of new tools, thus providing access to key scientific knowledge more quickly, and offering faster application of that knowledge. When there is full open access, including reuse rights, machines can become readers, fostering new layers of connection and innovation that are not possible through human-based processing. We see this at MIT where faculty in many disciplines (e.g. economics, political science, computer science) have found innovative ways to process and analyze information using software tools, providing new insights and new ways of approaching research questions across disciplines, speeding connections and discovery and driving innovation.

- Public access policies that are mandatory will achieve the highest levels of participation, and therefore have the biggest impact on economic growth and scientific innovation. We have seen evidence for this with the NIH policy, where participation went from 4% to 75% following a move to mandatory deposit (according to NIH's Neil Thakur, December 2011).

- Public access policies like the proposed Federal Research Public Access Act (FRPAA), which would extend NIH-like mandates consistently to all federal agencies granting more than $100 Million in research annually, would expand all of the benefits of public access dramatically. Victoria University Economist John Houghton's research concludes that opening access to all US publicly funded scientific articles would result in an increase in return on investment of at least 5 times. His research shows that "[t]he overall impacts of openly archiving all FRPAA agencies' funded R&D article outputs" would have "likely US national benefits of around 8 times the costs." (Houghton www.arl.org/sparc/bm~doc/vufrpaa.pdf) Supporting this theoretical study is the actual data from the NIH public access policy, which has been shown to be cost effective, absorbing only about .0001 percent of the overall NIH budget.

To generate the most benefit from such public access policies, it will be important to establish standards, avoid unnecessary proliferation of deposit systems and requirements, benefit from the experience of other repository services, and work with existing infrastructure, such as that developed by the NIH in operating PubMedCentral.

**(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders involved with the publication and dissemination of peer-reviewed scholarly publications resulting from federally funded scientific research? Conversely, are there policies that should not be adopted with respect to public access to peer-reviewed scholarly publications so as not to undermine any intellectual property rights of publishers, scientists, Federal agencies, and other stakeholders?**

Comment 2. The existing copyright law does not need to be changed in order to accommodate broader public access policies, such as a FRPAA-style mandate for all large federal agencies. For example, Creative Commons licenses (http://www.creativecommons.org) allow for the exercise of existing copyrights and are valid and enforceable under existing copyright law, and can be used under public access policies to extend the benefit of openness, by signaling appropriate uses that can be made under copyright law without the delays and complexities of permission-seeking.

CC licenses would improve upon the NIH model, which allows for "fair use" of the articles but does not unambiguously signal permission to create modified versions (derivative works), which is important to fuel innovation. Deploying Creative Commons (CC) licenses for articles made available under public access policies would expand the benefits of openness to allow for more innovation, by allowing for text/data mining and creating new works from existing works.

To address potential publisher concerns about making publicly funded research immediately available under a CC license, there could be a period during which research articles are available under US copyright law's fair use provisions (as are the articles collected under the existing NIH policy), after which they could be opened up under an appropriate CC license.

**3) What are the pros and cons of centralized and decentralized approaches to managing public access to peer-reviewed scholarly publications that result from federally funded research in terms of interoperability, search, development of analytic tools, and other scientific and commercial opportunities? Are there reasons why a Federal agency (or agencies) should maintain custody of all published content, and are there ways that the government can ensure long-term stewardship if content is distributed across multiple private sources?**

Comment 3. There is no adequate substitute for the Federal government providing long-term archives for the research it funds, though Federal archives could be effective whether highly centralized or less centralized. Current private models, on the other hand, have not proved adequate. A 2011 MIT study found that only 22% of MIT's subscribed journal titles are included in one of the top four e-journal archiving projects (Portico, JSTOR, HathiTrust, and CLOCKSS) (MIT only: https://wikis.mit.edu/confluence/x/enS6B). Another major research library (Cornell) confirms these findings, estimating that less than 15% of their journal holdings are archived by either Portico or LOCKSS (http://2cul.org/node/22).

A decentralized government approach to archiving could work if archives were interoperable. Partnerships between private and public entities might be an effective way to ensure that sufficient repositories are available for all federal agencies, as long as clear standards for access and use were adhered to.

Whether centralized or decentralized, all archives of publicly funded research must be set up to manage and actively serve content over the long term. Decades of digital archiving experience have shown that storing papers out of view for future audiences, will not serve the public over the long term: active, ongoing access and use are critical to successfully maintaining an archive.

**(4) Are there models or new ideas for public-private partnerships that take advantage of existing publisher archives and encourage innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research?**

Comment 4. The Digital Repository Infrastructure Vision for European Research offers a new model of confederated repositories whose goal is to "create a cohesive, robust and flexible, pan-European infrastructure for digital repositories." (see: http://www.driver-repository.eu/)

This kind of network of interoperable repositories could act as a model for US federal agencies. Given American universities' experience with existing archives, including for example MIT's development of the open source repository software DSpace (in partnership with Hewlett-Packard), universities are well positioned to act as partners in developing this kind of infrastructure in the US.

The existence of the PubMed Central International network, including UK PubMed Cental

(launched by the British Library in 2007) and Canadian PubMed Central (a partnership of Canadian governmental agencies and the US NLM) also suggests the kind of collaborative infrastructure that could be built. Another example is the Digital Public Library of America, announced in 2010. It is still in its early stages, but grew out of a public-private partnership that included universities, government, public libraries, and foundations. Its goal is to become an "open, distributed network of comprehensive online resources that would draw on the nation's living heritage from libraries, universities, archives, and museums in order to educate, inform and empower everyone in the current and future generations."
An important consideration in looking at these examples and possible models is that any archiving system must include some redundancy of access, so that one location is not the only source for an article. Relying on a single proprietary system, for example, creates unacceptable vulnerability. We saw evidence of this clearly decades ago when the nonprofit JSTOR was building its archive of journals. In many cases, JSTOR turned not to the original publishers, but to libraries (including MIT's) to provide copies of journals to complete the scanning of back issues, since publishers had not been maintaining complete archives of what they had published and could not supply the full back runs.

**(5) What steps can be taken by Federal agencies, publishers, and/or scholarly and professional societies to encourage interoperable search, discovery, and analysis capacity across disciplines and archives? What are the minimum core metadata for scholarly publications that must be made available to the public to allow such capabilities? How should Federal agencies make certain that such minimum core metadata associated with peer-reviewed publications resulting from federally funded scientific research are publicly available to ensure that these publications can be easily found and linked to Federal science funding?**

Comment 5. Metadata standards are critical to allow for interoperability. Existing standards can support the development of minimum core metadata, including the Dublin Core and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, see: http://www.openarchives.org/pmh/).

Critically important is the development of controlled identifiers for authors, such as the work being done by the ORCID project (see: http://orcid.org/). Standard name identifiers should create much-needed efficiencies in author name disambiguation for research institutions and research funders, for whom connecting particular individuals unambiguously to specific research papers is a critical function that underlies all other tools and services related to use of the research literature.

To maximize the potential for analyzing how publicly funded research is being used, adherence to usage tracking and analytics standards, such as the COUNTER project (http://www.projectcounter.org/about.html) and NISO's SUSHI project for automated harvesting of usage statistics (http://www.niso.org/workrooms/sushi/), is essential.

Also significant is the need to require and support metadata that clearly identifies the research sponsors and the grant IDs associated with a paper. Relevant to this need is recent work by

CrossRef (http://www.crossref.org/) to add grant identifying information to their collaborative reference linking service.

Metadata can facilitate the analysis and reuse of publications, speeding the acquisition of scientific knowledge, but must be machine readable as well as interoperable to achieve this goal most fully. Use of metadata standards along with an application programming interface (API) that allows for standards-based data exchange will promote the most efficiency in analysis and reuse (see for example JSON http://www.json.org/).
Metadata describing article version is also an important tool to support effective public access. Readers need to know whether they are looking at a version that incorporates peer-review changes and whether there are updated versions of an article. Standards for identifying and linking versions are needed, building on recommendations from the NISO Journal Article Versions working group (http://www.niso.org/publications/rp/RP-8-2008.pdf) and those of the JISC VERSIONS project and Version Identification Framework (http://www.jisc.ac.uk/whatwedo/programmes/reppres/vif.aspx). Tools to expose version information are also needed, with CrossRef's new version-identifying CrossMark tool (http://www.crossref.org/crossmark/index.html) providing an important step in that direction.

Metadata is needed not just to describe or provide information. Metadata is an essential tool to enable services related to the publications described. For example, at MIT we are using metadata from CrossRef to enable services in some small but significant ways in our implementation of the faculty's Open Access Policy. We save considerable time and improve accuracy by automatically obtaining bibliographic data on our scholarly articles based on a DOI match.
Using specialized metadata that captures the meaning of content will support implementation of collaborator and expert identification systems like the open-source profiling system, VIVO (http://vivoweb.org/). Such "semantic metadata" offers rich, meaning-based terms to index scholarly output. By use of standard vocabularies, inter-institutional networks of information become possible, with improved discovery.
Structured metadata like this will be increasingly important for scholarly articles as they move into new formats that don't allow for mining of full-text, such as video and image formats. For example, JoVE, the Journal of Visualized Experiments, is the first scholarly, peer-reviewed "video journal," a format that for which structured metadata will help with discovery and reuse.
Rights information is another area where standardized metadata is necessary to grease the wheels of reuse, indicating to human and machine readers whether and to what extent a work can be modified and built on. By including specific metadata such as Creative Commons or other licensing metadata, we should enable not just discovery of content, but its reuse.
As we move forward, metadata (such as citable URIs) will be the key to connecting publications with associated research data, allowing for newly efficient access to data and providing the opportunity for the development of automated tools to find, mine, and use the data. This critical connectivity is missing from our current landscape (see for example results of the PARSE.Insight survey at: http://www.dlib.org/dlib/january11/smit/01smit.html) and will provide a foundation for the next level of efficiency and growth in scientific knowledge management.

**(6) How can Federal agencies that fund science maximize the benefit of public access policies to U.S. taxpayers, and their investment in the peer-reviewed literature, while minimizing burden and costs for stakeholders, including awardee institutions, scientists, publishers, Federal agencies, and libraries?**

Comment 6. 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.
At MIT we have seen direct benefits from enabling tools for automated deposit. We have been using the SWORD protocol (Simple Web-Service Offering Repository Deposit, see: http://swordapp.org/about/) to automatically deposit papers from a major commercial publisher into our local repository. This process has afforded dramatic time savings and has tangibly confirmed that enabling automated repository and inter-repository deposit is a critical component of creating efficiencies for authors and publishers, driving compliance rates and therefore impact.
Looking ahead, public access policies should consider leveraging existing protocols such as SWORD, and engaging with open source tools and open architecture solutions. The goals should be to promote the benefits and availability of publicly funded research, make such research readily and persistently accessible, and provide platforms for use and reuse.

**(7) Besides scholarly journal articles, should other types of peer-reviewed publications resulting from federally funded research, such as book chapters and conference proceedings, be covered by these public access policies?**

Comment 7. Public Access policies should reasonably include all outputs that result from publicly funded research including articles, book chapters, conference proceedings, multimedia, and future forms of scholarly communication not yet normalized.

The approach taken at MIT is that if a research article has been produced for the purpose of contributing to research and scholarship and is made available for publication at no cost and with the assumption it will be shared with other researchers, the faculty's Open Access Policy applies. The MIT Faculty Open Access Policy covers all "scholarly articles," defined as "articles that describe the fruits of [faculty] research…that they give to the world for the sake of inquiry and knowledge without expectation of payment. Such articles are typically presented in peer-reviewed scholarly journals and conference proceedings." The same approach would appropriately be applied to federal agency public access policies, in order to incorporate all relevant publicly funded scholarly articles under public access policies.

If other forms of scholarly output, such as scholarly monographs or text books, incorporate knowledge acquired by the author pursuant to federal funding , policy conditions may need to account for author effort, timeliness, and market differences.

**(8) What is the appropriate embargo period after publication before the public is granted free access to the full content of peer-reviewed scholarly publications resulting from federally funded research? Please**
**describe the empirical basis for the recommended embargo period. Analyses that weigh public and private benefits and account for external market factors, such as competition, price changes, library budgets, and other factors, will be particularly useful. Are there evidence-based arguments that can be made that the delay period should be different for specific disciplines or types of publications?**

Comment 8. Immediate access benefits the public most fully. Experience with funder-mandated embargos of 6 months or less (as with the Howard Hughes Medical Institute's policy) suggests that publishers can sustain their publishing activities under these conditions. We note that the Open Access Policy adopted by MIT's faculty views immediate access as the appropriate goal, and that the policy does not support delays of any length in sharing the author's final manuscript version.

The experience of arXiv (http://www.arxiv.org) has demonstrated that open access to scholarly articles in high-energy physics and related disciplines prior to formal publication does not affect journal subscriptions. The American Physical Society has repeatedly and publicly confirmed that they have no data associating subscription cancellations with open access to the high-energy physics literature. Embargoes of 12 months or less have been adopted by hundreds of journals, such as those supported by Highwire Press (http://highwire.stanford.edu/lists/freeart.dtl), and by research funders worldwide (see: http://roarmap.eprints.org).

Similarly, the Royal Society has learned that open access to back content does not result in significant revenue loss. The Royal Society has determined that less than one half of one percent of their overall publishing revenue is attributable to content older than 12 months. There is likewise no evidence that any publisher has been harmed by the NIH Public Access Policy, even during a significant economic downturn.

The full array of market conditions that may influence journal cancellations must be included in any analysis that purports to show a correlation between journal cancellations and open access to a portion of journal articles. Factors that are determinative at MIT include changing faculty research directions, budget support for the library, journal prices and compounding journal price increases, cost per use, terms and conditions of use, and increased competition from new journals.

Closing comments:

MIT wishes to thank the Office of Science Technology and Policy for this opportunity to comment. MIT's mission, like that of other US research universities, is to generate, disseminate, and preserve knowledge -- -- and through its research to serve as an engine for economic growth.

MIT's research has the potential to translate directly into growth in the regional and national economy. A study by the Kauffman Foundation finds that companies started by MIT alumni generate revenues that are comparable to the 17th-largest economy in the world (See:

http://web.mit.edu/newsoffice/2009/kauffman-study-0217.html). At the time of the study (released in 2009), 6,900 MIT Alumni companies with worldwide sales of approximately $164 billion were found to be located in Massachusetts, representing 26 percent of the sales of all Massachusetts companies. Those in California generated another $134 billion in worldwide sales. This study documents the dramatic and “critical role universities play not only in fostering innovation and entrepreneurial growth, but in stimulating the much-needed recovery in regional and global economies.” Open access to research results has the potential to further enhance the economic impact of federally funded research.

Other countries are increasingly aware of the national advantage and improved return on investment created by expanded access to their publicly funded research. These advantages have been documented in studies of potential improved return on investment in the UK, the Netherlands, Australia, and Denmark as well as the United States (see http://www.cfses.com/documents/wp23.pdf ; http://www.jisc.ac.uk/publications/documents/economicpublishingmodelsfinalreport.aspx; and *http://www.arl.org/sparc/bm~doc/vufrpaa.pdf*)

The ability of research universities to continue to contribute to the welfare of the nation and the interests of the states and local communities in which we reside is fundamentally connected to the open availability of the research results produced by MIT and by the country’s large and small research universities. We appreciate the Office of Science and Technology and Policy’s interest in this connection between public access to research and public benefit, and we thank the Office of Science and Technology and Policy for this opportunity to comment on a critical aspect of the research cycle.

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