Request for Information: Public Access to Peer-Reviewed Scholarly
Publications Resulting From Federally Funded Research

Response: University of Illinois at Urbana-Champaign

[Assigned ID]
[Assigned Entry date]

Paula Kaufman
Juanita J. and Robert E. Simpson Dean of Libraries and University Librarian
University Library
University of Illinois at Urbana-Champaign
Urbana, IL 61801

We thank the White House Office of Science and Technology Policy (OSTP) for the opportunity to comment on policies related to public access to peer-reviewed scholarly publications resulting from federally funded research. We support the development of policies that allow immediate, free, and open access to and reuse of such scholarly publications without commercial restriction. We strongly believe that such policies are in the best interest of the US public and would advance the scientific enterprise by allowing unimpeded access and use of the best and most current research findings. In addition, such policies would allow economic growth through the application of such findings in private, commercial enterprises. Such policies are feasible and realistic. We respond below to the specific questions in the RFI.

Comment 1:

Are there steps that agencies could take to grow existing and new markets related to the access and analysis of peer-reviewed publications that result from federally funded scientific 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?

We believe that scholarly publications–in particular peer reviewed scholarly articles and conference publications—resulting from federally funded research should be made publicly, freely accessible without commercial restriction in such a way that allows the public, the scientific community, and the commercial entities to make full use of them whenever possible. Full use means not only access to download and read a publication, but also the ability to text and data mine, to make use of the information through computational methodologies, and to create derivative works with attribution. Enabling open access of these publications allows innovative individuals, organizations, and companies immediate access to research findings and trends to inform new research, services, and applications.

Enabling full reuse facilitates the ability of innovative individuals, organizations, and companies to build a variety of applications and services on top of this corpus—discovery and data mining systems, improved search algorithms, semantic web applications, to name a few—that will continue to extract value for this initial investment for decades to come. We do not really know who will be interested in a research publication either now or in the future. Restricting access and reuse rights essentially cuts off potential previously unexpected public uses and commercialization of federally funded research.

We offer one other small example of such a public use. The Mid-America Earthquake Center at the University of Illinois makes its research reports openly available in the campus institutional repository. One of these, a report on the potential impact of a severe earthquake at the New Madrid Seismic Zone that runs through Illinois, Indiana, Missouri, Arkansas, and other Midwestern states (http://hdl.handle.net/2142/14810), was used by a group of ham radio enthusiasts in Arkansas to develop emergency plans and maneuvers. While not a peer-reviewed publication, this is exactly the type of public use that open access to such publications can encourage and that restrictions inhibit. As an example, Cambridge University Press' analysis of Cambridge Journals Online indicates that of the 60 million views of journal abstracts, 20% were from users unaffiliated with a subscribing institution; only 0.01% of those views led to the purchase of an article. (*Chronicle of Higher Education*, Nov 30. 2011)

There is evidence that restricting openness and reuse rights also inhibits new research within the scientific enterprise. Murray et al (2009) analyze citations and use of material before and after access and IP rights of mouse genetic data was released from commercial control in 1998 through a National Institutes of Health (NIH) brokered arrangement. They come to the conclusion that:

> …particularly in research settings characterized by high levels of freedom [i.e. researchers can set own agendas], openness not only increases the overall flow of research output, it should also be closely associated with the establishment and exploration of entirely new research lines. Moreover, while openness should affect both basic and applied research, the impact on basic research will…dominate when researchers in the pre-openness period face high fixed costs of initiating a new line of research. In contrast, the increase in applied research will dominate when significant basic research has already been conducted. (3)

Murray et al also note that "the bulk of the new citations arise from articles published by new researchers or institutions. In other words, most of the incremental citations to a given mouse-article come from researchers working at institutions that had not cited that mouse-article prior to the NIH agreement." (4) A similar analysis was done by Williams (2010) on the impact of restrictive intellectual property rights gene-level data by the private firm Celera on subsequent scientific research and product development. Her analysis suggests that the restrictive IP "led to economically and statistically significant reductions in subsequent scientific research and product development outcomes." (2)

Evidence on the costs and benefits of enacting a public access policy can be found in the Houghton et al 2010 study of the economic and social returns on investment in open access to publicly funded research outputs (the study was commissioned by SPARC to understand the impact of the proposed Federal Research Public Access Act (FRPAA)). Houghton et al estimate that if FRPAA was enacted, the overall

potential benefits to the United States would be about 8 times the cost of implementing a public access policy at about $5 billion (using the cost structure associated with the NIH mandate) (Houghton et al, 2010, 8). The authors note that these benefits are diffuse, that is they would be found throughout the scientific community and economy, and would accrue over time.

References:

Houghton, John, Bruce Rasmussen, and Peter Sheehan. 2010. *Economic and Social Returns on Investment in Open Archiving Publicly Funded Research Outputs*. Report to SPARC. Centre for Strategic Economic Studies. Victoria University. Victoria, BC. Retrieved from http://www.arl.org/sparc/bm~doc/vufrpaa.pdf

Howard, Jennifer. Nov. 30, 2011. "Cambridge University Press Would Like to Rent You an Article." *The Chronicle of Higher Education*. Washington, D.C.: The Chronicle of Higher Education. Retrieved from http://chronicle.com/blogs/wiredcampus/cambridge-u-press-would-like-to-rent-you-an-article/34500

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. 2009. *Of Mice and Academics: Examining the Effect of Openness on Innovation*. NBER Working Paper 14819. Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14819.pdf.

Williams, Heidi L. 2010. Intellectual Property Rights and Innovation: Evidence from the Human Genome. NBER Working Paper 16213. Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w16213.pdf.

Comment 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?*

The NIH policy as it currently stands is a good model to examine in this context: it allows open and public access to federally funded research while still allowing the publishing community to profit and benefit from the publication of this research. This mandate asks that the researchers confirm during the publication process that they can deposit their research in PubMedCentral; this is a matter of ensuring that such language is included in the copyright transfer agreement or license agreement, and most publishers now have standard boilerplate language inserted. What the NIH policy does not do is allow wider reuse of this material (including data mining).

In general, the primary goals of the publication and dissemination of peer-reviewed scholarly publications is to reach as wide an audience as possible and to have the greatest impact as possible. Most researchers within academic institutions are 'paid' for publication through indirect means: promotion and tenure, grants, invited talks, recognition within their scholarly societies, etc. The greater the impact of their research (as measured through citations, attributions, amount of research dollars received, invitations to collaborate on research projects, etc.) the better 'paid' they are. It is to the advantage of scientists and researchers to take allow maximum levels of access **and** reuse, provided that they are attributed appropriately. We believe that the intellectual property interests of scientists, Federal agencies, and the taxpaying public (who also wish to see an impact made given the research dollars invested) are best protected by not overly restricting access and reuse of scholarly publications given appropriate attribution. Retention of copyright by the authors of the paper and use of a license such as the Creative Commons Attribution license[1]–as the Public Library of Science does in all of its publications—would satisfy this interest in the majority of cases.

While publishers certainly add value to peer-reviewed scholarly publications through their management of the peer review process (though, we note, not the actual conducting of the peer review), editorial oversight, marketing, and branding, it is not necessary for researchers to transfer full copyrights to them, as is done in most cases. Publishers should retain rights to copy and distribute, as well as the right to be named as the place of first publication (i.e., citation). Stronger protection of intellectual property rights for publishers has and would continue to result in restricted and impeded access to federally funded research; this simply does not make sense if the goal is to increase the impact, use, and potential commercialization of federally funded research.

Comment 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?*

In our experience (gained from extensive research on the effective aggregation of metadata and content from disparate sources as well as the development of federated search technology to bring together content[2]) the centralized vs. decentralized approach included the following:

- Pro: Centralization facilitates consistent enforcement of standards and compliance (the last assuming that such management would be at the agency level);

---

[1] http://creativecommons.org/licenses/by/3.0/
[2] See http://imlsdcc.grainger.illinois.edu/about.asp and http://search.grainger.uiuc.edu/searchaid/easy_search_summary.html for publications and summaries of research.

- Pro: Centralization allows a 'one stop shop' for search;
- Pro: Centralization allows for a standard, well documented programming interface (or API) that would facilitate the development of analytic tools;
- Pro: Decentralization allows institutions and publishers to develop systems that meet their needs as well as a public access policy;
- Con: Decentralization does not facilitate enforcement of standards, compliance, or standard APIs to facilitate tool development;
- Con: Decentralization would necessitate the need to aggregate the metadata somewhere or provide a standard way to search across the decentralized system;
- Con: Centralization can have a high overhead and cost given the necessary size and scale of management;
- Con: Centralization generally does not allow the flexibility to meet different needs of depositors, publishers, and searchers;
- Pro and Con: Decentralization can spread the costs of a public access policy across the scholarly publishing ecosystem.

We are sure that there are others to consider, but these are some of the major issues in our experience.

That said we are agnostic whether there should be centralized or decentralized approaches to managing public access to peer-reviewed scholarly publications that result from federally funded research. In a decentralized approach, however, we urge you to consider the role not only of publishers, but of the research library community in helping to support such a decentralized approach. In the digital sphere, academic research libraries have now years of experience in managing large repositories, in digital preservation, and have strong relationships with researchers (particularly on their home campuses); in the physical sphere, we have centuries of such experience. Organizations such as the Coalition for Networked Information (CNI), the Association of Research Libraries (ARL), Portico, the Hathi Trust, and the Digital Library Federation (DLF) should be consulted for their advice on roles that research libraries could play in a decentralized model. We do not believe that commercial, for-profit publishers that ultimately answer to their shareholders should be solely or even mostly responsible for the long-term stewardship of the results of federally funded research. Academic research libraries are the memory institutions of science, and we believe could be an important component of a decentralized model if that is the direction that is chosen.

Comment 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?*

Repeating our last paragraph in Comment 3: we urge you to consider the role of the research library community. In the digital sphere, academic research libraries have now years of experience in managing

large repositories and in digital preservation, and have strong relationships with researchers (particularly on their home campuses); in the physical sphere, we have centuries of such experience. Organizations such as the Coalition for Networked Information (CNI), the Association of Research Libraries (ARL), Portico, the Hathi Trust, and the Digital Library Federation (DLF) should be consulted. Examples such as Portico, the Interuniversity Consortium for Political and Social Research (ICPSR), the Hathi Trust, and arXiv are evidence that there are partnerships possible under a variety of models to both encourage innovation and to ensure long term stewardship. We do not believe that commercial, for-profit publishers (that ultimately answer to their shareholders) should be the single point of access to the results of federally funded research.

Comment 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?*

Making peer reviewed publications resulting from federally funded research openly available is the first step to encouraging interoperable search, discovery, and analysis capacity across disciplines and archives. The second step is to require the use of such broadly accepted standards such as the National Library of Medicine (NLM) journal archiving and interchange tagging sets; this will enable the creation of standard services across sets of archives, publisher files, and other repositories. As mentioned in comment 3, within the Library and the Graduate School of Library and Information Science we have extensive experience in applied research on the aggregation of metadata and content from disparate sources as well as the development of federated search technology to bring together content.[3] Put simply, the more consistent the metadata and content types, the easier it is to build effective search and discovery tools on top of the metadata and content.

In addition to search tools, we believe there should be investment in tools that allow better discovery through browsing and linking. This would require use of controlled vocabularies, standard identifiers for specific entities (certainly people, institutions, and funding agencies), use of grant IDs, and metadata that would allow the expression of relationships between items (research papers and data sets), items and entities (research paper and author), and entities (author and institution). There currently exist an array of broadly accepted and emerging standards that could be used; again, consistency in their use is the key to easily building effective discovery tools.

---

[3] See http://imlsdcc.grainger.illinois.edu/about.asp and http://search.grainger.uiuc.edu/searchaid/easy_search_summary.html for publications and summaries of research.

In order to encourage analysis capacity specifically (which we understand to mean the ability to text and data mine, markup, and make use of information within a text through computationally based means, as well as to download and access the item) ideally three pieces must be in place:

1) rights and permissions information must be expressed explicitly within the metadata;
2) the full text of articles must be in a standard format that is friendly to such machine processing; and
3) this format should be available via application programming interfaces (API) in a standard and well documented way.

For point 1, for publications to be fully useful, they should be licensed, as argued above, in such a way that allows such use without explicit permission being sought (for example, using a Creative Commons Attribution Only license). For point 2, the NLM journal archiving and interchange tagging sets (http://dtd.nlm.nih.gov/) are well accepted standards in the library and publisher community.

The core metadata set must not simply be descriptive of the published piece, but as we suggest above, must enable specific actions–discovery of similar works, datasets associated with the work, understanding of what can be done with the work–and should also be machine readable and processable. We suggest that the National Library of Medicine, the Library of Congress, and the National Information Standards Organization (NISO) be consulted about the creation of a minimum metadata set that could be made available to the public. These organizations have long experience with working with standards as well as a deep understanding of what would be required to enable discovery and use.

Comment 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?*

In order for Federal agencies to maximize benefits to taxpayers while minimizing burden and cost for stakeholders, there must be consistency of policies and requirements. Uniform policies and requirements allow researchers, awardee institutions, and other stakeholders to set into place standard workflows for depositing publications resulting from federal research funding, rather than a workflow for each agency (which would be particularly confusing for those researchers who hold grants from multiple federal agencies).

In addition, whatever the technical implementation of the public access policy is, the emphasis should be on an open technical infrastructure that allows the development of applications and toolsets that would help both to maximize benefits and minimize costs (particularly after the initial set up costs). Examples of such tools include:

- *Minimize cost* - Use of openly documented application programming interfaces (APIs) and standard protocols that facilitate automatic deposit of publications to multiple repositories;
- *Minimize cost* - Tools that would enable the automatic recording of deposits in internal and external grant management systems, minimizing the tracking of these deposits for reporting purposes;
- *Minimize cost and maximize benefits* - Tools that would allow awardee institutions to more accurately measure research output, as well as to promote research;
- *Minimize cost and maximize benefits* - Tools that would allow researchers to automatically create profiles of their output and research interests, as well as research that might interest them, and networks of other researchers doing similar or complementary work. In some ways this type of work is being developed through VIVO (http://vivoweb.org/) which makes heavy use of the data provided through PubMedCentral; and
- *Maximize benefits* - Applications or portals that could highlight and contextualize material for use in teaching and learning at both high school and college levels.

These are just tools and applications that are easily imagined; as noted above it can be difficult to guess what other potential tools could be developed given consistent policies and requirements and an open technical infrastructure.

Comment 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?*

There are two issues to consider for this question.

1) Publication practices are quite different from discipline to discipline; thus limiting a public access policy to scholarly journal articles may have the unintended consequence of limiting access to federally funded research for one discipline but not another. For example, from our own analysis, a significant amount of the output from a computer scientist is in conference proceedings (http://connections.ideals.illinois.edu/groups/15-Computer_Science), while for an animal scientist the most significant output is in journals (see http://connections.ideals.illinois.edu/groups/24-Animal_Sciences). In the physical and life sciences, in general, there is relatively low output in the form of books or book chapters. For the majority of the social sciences, the output is mainly in the form of journal articles, with book chapters second (see http://connections.ideals.illinois.edu/groups/%2014 for an example of a Communication Department). In the Humanities there appears to be a fairly even mix between journal articles and book chapters and monographs (see http://connections.ideals.illinois.edu/groups/45-History for a History Department).

2) The economics of the book market is quite different than that of the journal and conference proceedings market. For example, authors may be paid for book chapters, and generally are paid for books, including textbooks. Authors are not paid for journal articles, conference papers, or conference posters. We believe that requiring that these be made readily accessible to the public would require a broader investigation of the impact on this market, and a better understanding of the cost and benefits of such an approach.

Based on our understanding of the publication patterns of researchers likely to receive federal funding for their research, we believe that providing open access to **scholarly journal articles and conference papers and posters** would allow for the greatest impact of federally funded research. While we would ideally like to see book chapters and other peer reviewed and edited material included under the public access policy, this would require a broader investigation.

Comment 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?*

Ideally, in order for the public, researchers, and commercial entities to be able to fully utilize the information contained in scholarly journal articles and conference papers and posters resulting from federally funded research, they should have immediate access to these resources. Houghton et al estimate that "a six-month embargo reduces the returns [from increased R&D] by around $120 million." (Houghton, Rasmussen, & Sheehan, 2010, p. 8) In a similar study but focused on the United Kingdom, Houghton et al found that a one year embargo reduced returns by 2% or £120 million. (Houghton Particularly for fast paced disciplines, such as much of the life sciences, long embargo periods would be particularly detrimental to access to the current state of the art.

That embargoes are necessary for those journal publishers that rely on subscription income, particularly in the sciences, is the common argument for instituting an embargo. We have found little evidence to support this argument beyond a report from the Publishing Research Consortium, a group representing publishers and societies, that indicates some librarians may be likely to cancel journals if all of the content was openly available. (Beckett and Inger, 2007). However, we note that this represents attitudes rather than actual data on the impact of embargoes (or lack of them) on journal subscriptions. This lack of data is likely because most funder mandates in Canada, the United Kingdom, and Europe as well as that from the National Institutes of Health (NIH) do have embargo periods attached to them for a period of 6 to 12 months.

Given this, we recognize that in order to balance the public's interest in having unrestricted access to scholarly publications resulting from federal research funding and that of commercial publishers and scholarly societies to see subscription income from journals and conference proceedings, embargoes may be a necessary part of public access policies (though we note that embargoes benefit only those publishers who use the subscription model). We believe that embargoes should be kept closer in line with other funder mandates that are limited to six months such as the Canadian Institutes of Health Research, the Australian Research Council, the Research Councils U.K., and the Wellcome Trust, rather than the 12 months that is the current policy of the NIH.[4] Some of the major scientific journals such as Nature and Science have only a six month embargo rather than 12 months. In the end, the length of an embargo—between 0 and 12 months—should be left to the author(s) to negotiate; it is the authors who best understand what the impact of an embargo might have on the dissemination and use of their research results.

We do not believe that a public access policy should institute different levels of embargos or delays based on disciplines. While there certainly exist disciplinary differences in the rate of publication, use of published literature, and primacy of the most current research results, in practice such differing embargos would be extremely difficult to administer; in particular, in areas of heavily interdisciplinary research or in sub-disciplines which are focused on applied research, embargos might be applied that may not be appropriate. A common embargo policy across all disciplines between 0 and 12 months that can vary based on negotiations of the author(s) or the policies of a publisher should be sufficient.

References:

Beckettt, Chris and Simon Inger. 2007. *Self-Archiving and Journal Subscriptions: Co-existence or Competition? An International Survey of Librarians' Preferences*. London, UK: Publishing Research Corporation. Retrieved from http://www.publishingresearch.org.uk/documents/Self-archiving_summary2.pdf.

Houghton, John, Bruce Rasmussen, and Peter Sheehan. 2010. *Economic and Social Returns on Investment in Open Archiving Publicly Funded Research Outputs*. Report to SPARC. Centre for Strategic Economic Studies. Victoria University. Victoria, BC. Retrieved from http://www.arl.org/sparc/bm~doc/vufrpaa.pdf

Houghton, John, Bruce Rasmussen, Peter Sheehan, Charles Oppenheim, Anne Morris, Claire Creaser, Helen Greenwood, Mark Summers, and Adrian Gourlay. 2009. *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits*. *A Report to the Joint Information Systems Committee (JISC)*. Victoria University and Longborough University. Retrieved from http://hdl.handle.net/2134/4137.

---

[4] ROARMAP: Registry of Open Access Repositories Mandatory Archiving Policies (http://roarmap.eprints.org/) provides the most comprehensive list of current funder public access policies.