10 January 2012

White House Office of Science and Technology Policy Request for Information on public access to peer-reviewed scholarly publications resulting from federally-funded research.

RESPONSE from: The Oberlin Group of Libraries

We appreciate the opportunity to provide comments on this important issue. The Oberlin Group of Libraries is a consortium representing 80 libraries of selective liberal arts colleges (see http://www.oberlingroup.org/). Our institutions are critical training grounds for young scientists, innovators, and entrepreneurs; we help fill the pipeline for graduate-trained scientists and inventors. According to the National Science Foundation, more than half of the top 50 baccalaureate institutions that produced Science & Engineering doctoral recipients from 1997 to 2006 were baccalaureate colleges (after normalizing for number of bachelor's degrees awarded 9 years earlier). Our institutions – the so-called "Oberlin 50" – have been cited as having particular importance in producing doctoral candidates (see http://www.nsf.gov/statistics/infbrief/nsf08311/). Our curricula are based on inquiry in laboratories, field stations, and other primary research materials – including the peer-reviewed reports of federally funded research under discussion here – and not primarily on textbooks. Our faculty publish sponsored research in collaboration with their undergraduate students. However, our schools do not receive the level of research funding available to larger institutions. Our faculty and students require access to cutting-edge research reports and data to assure they remain current with research in their fields and learn current science as science is actually practiced.

In these comments, we recommend that peer-reviewed published articles reporting research funded by any U.S. government agency be required to be freely available on the Web, in full, to all readers, no later than 12 months after publication, with full rights of re-use for text mining, data mining, computing, and creation of derivative works, without commercial restriction.

All of the references in our answers are openly accessible on the Internet.

Respectfully submitted on behalf of The Oberlin Group by:

Amy E. Badertscher
Director of Library Services
Kenyon College
badertschera@kenyon.edu

Richard Fyffe
Samuel R. and Marie-Louise Rosenthal Librarian of the College
Grinnell College
fyffe@grinnell.edu

Jonathan Miller
Library Director
Rollins College
jxmiller@rollins.edu

(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?

Immediate free access to federally funded research will maximize both educational and commercial potential. These papers should be searchable through open repositories (if there is more than one federal repository, then the policy should mandate interoperability; see our answers to question 5). Access should be free of charge and should include a broad range of re-use rights so that users can build on and innovate from the research that they find.

Providing open access to federally funded scientific and scholarly research reports – including full rights to re-use or mine these reports – allows more users to stay abreast of cutting-edge ideas, access these ideas quickly, and generate new uses and applications from this research, speeding the launch of new services and products into the marketplace and energizing the economy. Moreover, open access levels the educational playing field for teachers and students alike, helping teachers at both the high school and college levels stay current with their fields and giving students a more direct look at science as it is practiced, complementing the syntheses presented by textbooks. Open access also levels the economic playing field, giving new and established enterprises equal opportunity to compete.

We urge full open access as the norm: free, immediate access with full rights of re-use. Restrictions on access or on use simply reduce the return on taxpayer investment – whether that return is in the

education of new researchers or the entry of new products and services into the marketplace. We see the current NIH mandate as a good first step in this direction, one that matters to our faculty and students who routinely use PubMed Central and Google Scholar as research tools, in addition to benefiting research scientists, businesses, and private citizens. A good example of open access usage can be found in the testimony before Congress in 2011 given by the Director of the National Center for Biotechnology Information, who noted that in the previous year "99% of the articles in PubMed Central were downloaded at least once, and 28% were downloaded more than 100 times," and that 17% of the users are from companies and 40% from personal Internet accounts (not universities or government) (see http://www.hhs.gov/asl/testify/2010/07/t20100729c.html).

Moreover, the NIH mandate appears to have spurred development of new services like the Public Library of Science journal suite and BioMed Central journals. As a for-profit publisher of open-access journals, BioMed Central makes a particularly interesting case study. Research papers in BMC journals are freely available via the Web, and copyright is retained by authors (for federally funded research papers we recommend instead that authors be required to release full rights for re-use). BMC sells value-added products and services that take advantage of these open-access publications, but those products do not impede access to the research papers themselves for non-customers. BMC's business model was sufficiently successful that Springer, a major STM publisher, acquired it in 2008. We think that to unlock the whole value of federally funded research, the NIH mandate needs to be extended to all federal funding agencies and expanded to release full rights of re-use.

Finally, we note that an openly accessible database of research papers can itself become the target of innovative commercial or not-for-profit services that analyze, select, or present the results. Increasingly, scientific research depends on computer-based mining of text, data, and other digital information – but the mining can be only as productive as the lode is rich: full open access ensures both the availability of the right information to these analyses and a properly competitive position among the researchers.

(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 purpose of copyright is to "promote the progress of science and useful arts". In considering intellectual property protection, it is important to be clear about the different kinds of intellectual contribution made by the various participants in the research and publication chain, and to reward them according to their value (we are specifically addressing copyright here, and not patents or trademarks). Scientists and other scholarly authors are the primary creators of the intellectual property in research articles, and these researchers are rewarded through the institutions that support them rather than through direct payment or royalties on sales. The interests of these authors are best advanced through wide dissemination of their work, as attested by the numerous studies that show that open access to scholarly articles increases the pace and number of citations by other researchers (see, *inter alia*, Eysenbach, G. 2006. Citation advantage of open access articles. *PLoS Biology,* 4(5), 692-698. Available as an open-access article at http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0040157 (accessed 16 Dec. 2011)).

In this respect, scholarly authorship is special, and copyright is not the best policy tool to create incentive for innovation. Ensuring wide distribution, the authors' right of attribution, accurate quotation, and the authors' ability to measure and report on the quantity and type of use of the work – rather than protection against copying – best secure their interests.

The contributions of publishers are important, too, but those interests should not overshadow the interests of their authors. We recognize that limited periods of embargo – during which use-rights might be limited to fair use – may be necessary to reward publishers for their investments. As we note in our answer to question #1, we urge that after that embargo period, full rights of re-use be granted.

(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?

Repository services for managing and maintaining long-term public access to peer-reviewed scholarly publications from federally funded research require these basic elements: open access, technical interoperability, and long-term stewardship. A centralized repository may be able to enforce these expectations more easily than a decentralized set of repositories. On the other hand, there will be some differences in the practices of the different disciplines funded by the various federal agencies, and a small set of decentralized repositories (each perhaps with its own advisory group) might be more successful in accommodating these differences. Whether centralized or interlinked, this repository structure must be open to commercial search engines like Google Scholar, since even advanced researchers often start there for an information search.

Any repository in this system must also support access and use conditions that allow all interested parties (human and machine readers alike) to read the work, re-use it in ways that respect attribution, and create new services and products on top of this publicly funded information.

The repositories must be able to ensure that the information contained in them will remain fully useable over the long term (decades, not years). Commercial entities are subject to the vicissitudes of the marketplace, and even well-established publishers have not traditionally acted as archivists for their own work (instead, they have relied on libraries to archive their publications over decades or even centuries – Oxford University Press, for example, was established in 1586). Even non-commercial third-party projects are not keeping pace with the quantity of research needing long-term preservation. A recent presentation at the Fall 2011 meeting of the Coalition for Networked Information reported that even the well-established Portico and LOCKSS preservation initiatives preserve only 15-20% of the journals held by Cornell and Columbia University Libraries ("Preservation Status of e-Resources: A Potential Crisis in Electronic Journal Preservation," http://www.cni.org/topics/digital-preservation/preservation-status-of-eresources; accessed 16 December 2011). The funding agencies

themselves should collect the papers and data they sponsor and provide unrestricted access for the educational, research, and commercial sectors to utilize. Federal custody of research papers has proven to be cost-effective: the Director of the National Center for Biotechnology Information testified before Congress in April of 2011 that PubMed Central costs less than 1/100th of one percent of NIH's operating budget (http://www.hhs.gov/asl/testify/2010/07/t20100729c.html).

If third-party repositories that met conditions for public accessibility, use rights, interoperability, and long-term preservation of articles were to be approved as the dissemination vehicles for federally-funded research papers, it would still be necessary for the federal government to maintain ultimate custody of these resources, and any third-party contract would have to acknowledge the government's permanent stewardship responsibility.

(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?

The existing stakeholders – higher education, not-for-profit agencies, industry, and government – can all make important contributions to this emerging repository structure as designers, advisers, administrators, and hosts. Members of the Oberlin Group of Libraries believe that research universities represent the best candidates for partnership with federal agencies in ensuring access and preservation for publicly funded research papers (and data). Many of the researchers funded by federal agencies are university faculty who understand and trust their institutions; universities already have sophisticated technology infrastructures; and university libraries have long and successful traditions of working with federal agencies to preserve and disseminate the government's own publications.

We urge, however, that if non-governmental agencies (including universities) serve as hosts to the database of research papers then federal agencies must maintain an open mirror site to ensure ongoing accessibility for the public. Among models of partnerships, we suggest that particular attention be given to ArXiv, the e-print server for Physics, Mathematics, Computer Science, and related subjects, which was started at Los Alamos National Laboratory and later moved to Cornell University (http://arxiv.org/), and HathiTrust, an partnership of major research institutions and libraries working to ensure that the cultural record is preserved and accessible long into the future (http://www.hathitrust.org/about).

Any partnership should be predicated on clearly articulated standards for access, interoperability, and preservation.

(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?

Descriptive metadata following the Dublin Core standard and the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) should be the core standards for repositories and their contents; they will ensure the greatest interoperability with existing search and discovery systems, including commercial search engines like Google and Google Scholar and commercial indexing and discovery services like Serial Solutions' Summon. NISO and the Library of Congress – each with deep experience in developing and implementing such standards – should be enlisted to ensure that as linked data standards such as the Resource Description Framework (RDF) mature, these standards are reflected in the repository architectures that provide access to federally funded research papers (and data).

In addition, the metadata must carry information about the rights of re-use associated with research papers (and data). These metadata must be both human-readable and machine-readable, to ensure that the papers can be mined for the greatest benefit. We urge that Creative Commons licenses – specifically, the Attribution-ShareAlike 2.0 Generic (CC BY-SA 2.0) license – be embedded in metadata since CC licenses are machine-readable. Maximizing the accessibility of the research corpus and the metadata to machine processing will enhance research and educational use, optimize return on taxpayer investment, and lessen the compliance burden on researchers.

Finally, we recommend that any repository fulfilling a public-access mandate follow the COUNTER standard (Counting Online Usage of Networked Electronic Resources; http://www.projectcounter.org/), to ensure that authors, agencies, and the public see download and usage statistics generated in a consistent way (there are many ways to count downloads and accesses – libraries and publishers worked together to develop the COUNTER standard to lessen confusion for publishers and libraries alike). This is a basic step in ensuring accountability.

(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?

Public access provides good value for taxpayers in economic return. We noted in our answer to Question 3 the small percentage of NIH's operating budget required to operate PubMed Central. John Houghton and Peter Sheehan's 2006 working paper "The Economic Impact of Enhanced Access to Research Findings" estimates that "With the United State's [sic] GERD [Governmental Expenditures on Research & Development] at USD 312.5 billion and assuming social returns to R&D of 50%, a 5% increase in access and efficiency would have been worth *USD 16 billion*" (John Houghton and Peter Sheehan, "The Economic Impact of Enhanced Access to Research Findings," CSES Working Paper no. 23, Centre for Strategic Economic Studies, July 2006. Emphasis supplied. Available open access at http://www.cfses.com/documents/wp23.pdf; accessed 16 December 2011).

The Oberlin Group of Libraries believes that open-access requirements can be implemented without creating a burden for the stakeholders. The most important factor in keeping that burden low will be consistency of policy and procedures across the funding agencies – different requirements and submission procedures will surely increase complexity and confusion. It will also be important to implement repositories and submission systems that take advantage of automated protocols like SWORDS (http://swordapp.org/about/) which facilitate deposit of articles into multiple repositories, schedule the release of embargoed material, etc.

(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?

Journal articles, along with research data (to which the articles should be linked), represent the highest priorities for open access. There would be benefits to open-access release of other kinds of material that result from federally funded research, but the benefits would be smaller and should not distract attention from the primary goal of opening access to journal articles and their research data.

The Oberlin Group of Libraries recommend giving next priority to educational materials, especially materials targeted to the K-12 sector. It would be desirable for book chapters to be released openly as well, but these – unlike journal articles and research data – typically pay royalties to their authors and therefore require a different kind of business model. For such materials, longer embargoes or shorter terms of copyright (with dedication to the public domain at the end of the term) might be in order.

(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?

Immediate access is the best way to leverage taxpayer investment in research for educational, scientific, and commercial progress. Anything short of this withholds value from taxpayers and the larger economy. Even so, we recognize that journal subscriptions are an important source of revenue for publishers and that an embargo period might be necessary to protect them against loss. A period between 6 and 12 months has emerged as a world-wide norm for such an embargo. It's the standard used by NIH (12 months), the Wellcome Trust (6 months), and other major funders (see, for instance, http://roarmap.eprints.org/), and it has been adopted by hundreds of commercial and not-for-profit journals (see the list at (http://highwire.stanford.edu/lists/freeart.dtl).

As librarians, however, we also want to urge that any argument based on anticipated subscription cancellations by libraries be analyzed and tested carefully. Evidence needs to be presented that immediate access would actually cause economic harm. Libraries cancel journals for many different reasons. Among the key reasons are local budget reductions; price and price history (high annual percentage increases are flagged for review and cancellation in many libraries); emergence of new journals that have higher priority for the local academic program; and changes in the local academic or research program. In our collective experience, libraries rarely if ever cancel journal subscriptions based on any single reason, including open-access availability with or without an embargo.