Provost

Northwestern University
633 Clark Street
Evanston, Illinois 60208-1101

Phone 847-491-5117
Fax 847-467-1630
nu-provost@northwestern.edu

NORTHWESTERN
UNIVERSITY

January 12, 2012

The Honorable John P. Holdren
Assistant to the President for Science and Technology and
Director, Office of Science and Technology
New Executive Office Building
725 – 17th Street, NW
Washington, DC 20502

Comments in response to Office of Science and Technology Policy Request for Information: Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research
Federal Register Doc No 2011-28623
http://www.gpo.gov/fdsys/pkg/FR-2011-11-04/html/2011-28623.htm

Dear Dr. Holdren:

Northwestern University is a private research institution with 16,377 students and approximately 3,000 full time faculty. In 2010-11, Northwestern researchers attracted total awards and grants of approximately $511.7 million. Northwestern's libraries hold more than 5 million volumes, 4.6 million microforms, and provide access to 110,341 current periodicals and serials. In addition, the library system boasts more than 700 databases and 6,000 electronic journals. 56% of the libraries' $14 million collection budget is devoted to these e-resources.

Northwestern is recognized both nationally and internationally for the quality of its educational programs at all levels. *U.S. News & World Report* consistently ranks the University's undergraduate programs among the best in the country.

Among graduate programs, the Kellogg School of Management regularly ranks among the top five business schools in the country for both its traditional curriculum and its executive master's program. *U.S. News & World Report* rankings placed Northwestern's School of Law 11th, and the Feinberg School of Medicine in the top 20.

(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 publicly 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?

Making peer-reviewed scientific publications freely available after publication, with minimal restrictions on use, will accelerate scientific discovery and expand opportunities for entrepreneurs to develop new services and products. Lowering or removing barriers to access to new research results

will increase opportunities to identify new partnerships with industry, complementing the goals of university patent and technology transfer processes, and the goals of federal programs like the Small Business Innovations Research/Small Business Technology Transfer (SBIR/STTR) http://www.sbir.gov/ and the recently announced initiative to speed commercialization of university research (National Advisory Council on Innovation and Entrepreneurship, Department of Commerce, 2011). Closer academia-industry links and shorter cycles between research, dissemination of results, and commercialization accelerate the public's return on its investment, creating new markets, new jobs, and new tax revenue for local, state and federal governments.

It has been conservatively estimated that expanding an NIH-type post-publication open access policy to other federally funded research will result in improvements in research efficiency and accessibility, and yield for the American taxpayer a return approximately 8 times larger than the initial research investment (Houghton, 2010, p. 8). These projections align (and yet, they pale in comparison) with the measurable economic impacts of the Human Genome Project. That massive public project, whose results were made immediately available for both public and commercial use, yielded a return on investment of approximately $141 per $1 of public funding (Battelle Technology Partnership Practice, 2011, p. 6). In contrast, it is estimated that the IP restrictions temporarily placed on genes sequenced by Celera in its competing project have had a lasting negative impact on subsequent research and innovation. Genes first sequenced by Celera have fewer scientific publications and are less likely to be used in genetic tests (Williams, 2010, p. 2).

Likewise, providing immediate free access to research articles removes barriers for researchers, who can more quickly and effectively incorporate up-to-the moment findings into new research, accelerating scientific productivity. Even in university environments, researchers still report some difficulty gaining access to all of the scholarly material they need to conduct research, and these effects will be more severe for smaller businesses and worse yet for the general public. Open access publications, available through models ranging from fully open access journals to self-archived publications in university, disciplinary or funder repositories like PubMed Central, are downloaded more and cited more frequently than publications for which a subscription is required. Citation rates are significantly higher for immediate open access articles even when controlling for factors such as mandated vs. self-selective archiving, journal impact factor, and number of references cited (Gargouri et al., 2010, p. 8). Some studies have shown increased citation rates as high as 600% for open access publications (Swan, 2010, p. 17), though this varies significantly by discipline, and ranges from 40% to 90% are more common. Across social science, science and humanities disciplines, providing open access to published literature, particularly if the access is granted immediately after publication, will increase the impact of research. Most importantly, respected open access initiatives have succeeded in providing this broad access while maintaining and sustaining a robust peer-review process and continuing to provide many valuable services such as editorial enhancement, error checking, citation mining, and indexing and linking services.

Large databases of freely accessible scientific literature can also spur development of new knowledge exploration tools that aid researchers facing the daunting task of finding relevant publications amongst the hundreds of thousands of new articles published each year. Software like IN-SPIRE™ http://in-spire.pnnl.gov/ and the Action Science Explorer (Ferrante & Zgorski, 2011) and projects such as the Large Knowledge Collider (LarKC) http://www.larkc.eu/ give scientists powerful new tools for finding connections between previously unconnected research, using machine learning, automated reasoning, and network science to make new inferences and suggest new pathways for research. Tools such as BioXM(Maier et al., 2011) combine assertions drawn from published

literature with data about genes and other objects to yield new insights. These powerful computational tools depend on access to both metadata and full text for published articles, and constructing the new data sets and indexes on which they operate requires that the articles be free of downstream use restrictions, including prohibitions against commercial use.

Another example of machine-aided exploration may be found in the small but vibrant community developing around research networking (RN) tools. Both open source (VIVO http://vivoweb.org/ , Harvard Profiles http://profiles.catalyst.harvard.edu/) and commercial tools (SciVal Experts http://www.info.scival.com/experts , Thomson Reuters InCites http://researchanalytics.thomsonreuters.com/incites/) demonstrate the power of constructing author and concept network visualizations atop metadata and full text of research publications. These tools give universities and funders more accurate pictures of research output and ease the burden of publication tracking and reporting, but can also facilitate new collaborations and suggest new directions for exploration. However, RN tools will be limited by the quality and breadth of their inputs. In implementing a research networking tool at Northwestern, we have found that commercial database providers can be reluctant to make metadata or full text available for these non-consumptive uses, particularly if a commercial competitor developed the RN tool. The promise of research networking tools and other machine-aided inference systems will be severely constrained without access to large, freely reusable collections of research publications.

The National Institutes of Health (NIH) experience implementing a public-access policy and a large central database of results clearly show that this is a cost-effective approach to supporting open access to research. The article system's annual maintenance costs are approximately $3.5 - $4 million dollars, or roughly 1/100th of 1% of the NIH's $30 billion budget (Lipman, 2010).

(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?

To gain the greatest societal benefit from publicly funded research, preserve the ability to redistribute publications for teaching and other purposes, prepare derivative works, and promote development of new discovery tools and products, a public access policy should seek to make publications as broadly usable as possible, as close as possible to publication time, with few to no restrictions on reuse. Licenses or policies that permit publishers to restrict uses of open access copies to single reader use only must be avoided, or at the very least a phased approach considered that will permit some restrictions on use during an embargo period, but release the works for full reuse afterwards. Any open access, whether green, gold, gratis or libre (Suber, 2008), is better than none, but the Creative Commons CC-BY Attribution license is most conducive to use by readers and machine reading systems.

Policies that permit publishers to compel authors to sign over their copyrights must also be avoided. Nonexclusive licenses to publishers should become the norm, rather than a surrender of the author's copyright. There is sufficient leeway in composition of such licenses to permit publishers to recoup costs associated with provision of publication services without restricting self-archiving or productive

and socially beneficial derivative uses of scientific publications. Some publishers have developed paid options ("SHERPA/RoMEO - Publishers with Paid Options for Open Access," n.d.) for selective open access, often a hybrid model that mixes free and paid access content in the same publisher-hosted journal site. In some cases, as with the recently revised Taylor & Francis iOpenAccess service ("Taylor & Francis Author Services - iOpenAccess & NIH policy," n.d.), articles are portable, and may be posted to any institutional or disciplinary repository, but carry with them additional terms and conditions to prohibit, as in the Taylor & Francis example, certain uses including commercial uses. It is understandable that publishers are leery of repackaging and reselling, but blanket prohibitions on commercial use, particularly when authors have paid several thousand dollars—as high as $3000 per article in the case of T&F—for open access, may be unnecessarily restrictive, and cripples innovative, value-added and highly productive uses as well as simple reselling.

(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?

The federal government has already demonstrated through the NCBI systems, including PubMed Central, that it is capable of efficiently mounting a large scale, trustworthy and robust repository for both publications and data. Furthermore, it is appropriate for the U.S. government, as a major funder for scientific research, to also accept the responsibility for permanent stewardship of these important assets, to preserve them, and to continue to provide broad public access. Centralizing management of publications achieves economies of scale and eliminates the need for federated search tools, metadata or full text harvesting services, and other linking or mirroring systems to tie distributed archives together. Consistency and uniformity for publishers and authors will be the result. A disadvantage of a centralized approach may be that it minimizes the role of disciplinary and institutional repositories, and reduces capacity to provide specialized services and description tailored to the data and publication types specific to certain domains. A decentralized approach may also facilitate better access to research that is not funded by the U.S. federal government, but is available on an open access basis.

(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 research networking and knowledge extraction tools discussed above present very compelling cases for the potential of effective public-private partnerships. Many commercial database providers contract with open access publishers to include their full text in added value products that greatly enhance university researchers' information discovery experiences. These offerings can and should continue to flourish in an open access environment, and can provide publishers and scholarly societies with additional revenue streams, greatly expanding the number of resources they can index, mine, and provide access to, and significantly enhancing their value. Publishers could also act as contracted service providers to provide open access repository services, provided they are able to meet conditions for trustworthiness, accessibility, reuse, and openness.

However, better models for partnership may exist between funders and universities, particularly with libraries that have amassed significant experience with digital repositories over the past decade. Northwestern has an internal digital repository system based on the Fedora Commons software, but is also a founding partner in the HathiTrust shared digital repository system. HathiTrust has satisfied a Trustworthy Repository Audit and Certification (TRAC) assessment and currently houses some 10 million digitized volumes. The partnership plans to expand support for other content types, and to pilot digital publishing services through the HTPub project http://www.hathitrust.org/htpub . This development and others like it, such as the California Digital Library's Merritt repository and eScholarship system, could dovetail with plans to expand federal open access requirements and accelerate scientific publication archiving programs. Should the U.S. government decide not to expand with NCBI-like central repositories, a promising model is partnerships with large university digital repositories or large multi-institutional repositories such as HathiTrust. Likewise, disciplinary repositories such as arXiv and the Social Science Research Network (SSRN) have succeeded in developing scalable, reliable solutions to open access archiving and could be logical partners in a distributed or shared/mirrored archive model. A network of distributed repositories, like the European DRIVER project http://www.driver-repository.eu/, would build on existing investments and disciplinary customizations.

(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?

The current NCBI databases and the NLM DTDs are an admirable model and an excellent starting point for an expanded network of open access repositories. Standardization of the form and content of metadata and full text (JATS XML, for example) will be critical to successful interchange, will promote use by machine readers, and can facilitate automated deposit and other unmediated or minimally mediated activities. Existing metadata standards such as Dublin Core and machine exchange via APIs or OAI-PMH are well established and accepted components in data linking and exchange. Custom metadata schema for most domains can be crosswalked to Dublin Core in the absence of common element sets for cross-searching, which is easily extended through application profiles or qualifiers. In addition to predictable forms for descriptive metadata, standard approaches must be devised to express rights and provenance beyond authorship (version, lifecycle events, etc.) in a machine-readable and machine-interoperable manner. PREMIS is an accepted standard for provenance and rights metadata in the library domain, and may be suited for extension for these purposes. The SWAN provenance, authoring and versioning ontology specification may also be a useful model. Emerging standards such as ORCID, for disambiguating author identities, and I-2 for consistent identification of institutions will also be important and must be supported, and collaborations with NISO, the Library of Congress and other groups involved in standards development and maintenance will be invaluable to consensus building.

The systems must fully enter the linked open data ecosystem, and must be capable of supporting semantic description and enhancement, either natively in the database or by exposing sub-article information through durable URIs so that inferences drawn from published research can be banked separately and linked to the evidence in the underlying articles. Ideally, funder databases can be

expanded to include direct storage of RDF-based statements formally asserting relationships between concepts or objects. Robust support for RDF and concept linking can enable formalization of statements, sometimes referred to as 'nanopublications,' as recognizable contributions to the scientific discourse (Mons & Velterop, 2009).

(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?

Adopting uniform requirements across all funding agencies will greatly simplify the burden on universities and their researchers. Scientists are likely to have grants from multiple agencies, and a single set of deposit requirements reduces complexity and simplifies compliance, reporting and monitoring.

(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?

Expanding public access policies to include other products of federally funded research may help to meet the goals of broad and timely sharing of results and address the lag between discovery and formal publication. Educational materials, including conference proceedings, technical reports, books and book chapters, should all be eligible for coverage under a policy, but the specifics of the policies may differ from those developed for journal articles. Conference presentations and technical reports may lie on one end of a spectrum, where rapid deposit is a reasonable expectation, but the economics of book publishing are more complex, and longer delays may be necessary to recoup author payments or other publication costs. Access and reproduction could also be significantly more complicated and cumbersome with books or proceedings where it could often be the case that some chapters or sections are Federally funded but not all. The full book or proceeding might not be able to be distributed as an integral product with consistent pricing or rights management. Public access policies should evolve in keeping with the norms and practices of academic disciplines and scholarly societies, preservation of high quality peer review, and the types and forms of publications natural to academic discourse. There may be opportunities to encourage sharing of other types of research results, e.g. negative results, which can also increase research efficiency. In all cases, policies should apply to results that the investigators have decided to disclose through publication, presentation or other means, thus avoiding potential conflicts with technology transfer processes as well as risks to national security or patient privacy. A public access policy should not develop new, alternative forms of publication, such as final project reports, as a substitute for the forms of scholarly communication that already exist and that serve the goals of research dissemination.

(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?

We are not aware of any data or studies showing that the NIH-permitted embargo period of 0-12 months has harmed publishers in any way, and ample evidence exists that the sooner an article is open access, the greater its research impact (Swan, 2010). Publishers of all kinds have been able to sustain the high quality peer review critical to a reputable scholarly communication system without suffering economic harm. Although embargo periods prior to the use of the publisher PDF vary ("SHERPA/RoMEO - Publishers allowing the deposition of their published version/PDF in Institutional Repositories," n.d.), there is no clear pattern along disciplinary lines, and indeed, many publishers (225 according to the SHERPA/RoMEO lists) are willing to allow immediate self-deposit of this version. Many publishers, such as the American Chemical Society, are now willing to deposit the final version of an article on behalf of the author. The Houghton study into likely economic impacts of a broader federal open access policy states: "These estimates assume a six-month embargo period between publication and open accessibility. If there were no embargo, we estimate that incremental returns might be closer to $1.75 billion. Hence, a six-month embargo reduces the returns by around $120 million (NPV) (Houghton, 2010, p. 8)."

Thank you for this opportunity to comment.

Sincerely,

Daniel Linzer
Provost

Works cited

Battelle Technology Partnership Practice. (2011). *Economic Impact of the Human Genome Project*. Battelle Memorial Institute. Retrieved from http://www.battelle.org/publications/humangenomeproject.pdf

Ferrante, E., & Zgorski, L.-J. (2011, December 7). A New Visualization Method Makes Research More Organized and Efficient. *National Science Foundation (NSF) Discoveries*. Retrieved December 18, 2011, from http://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=122509&org=NSF

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PLoS ONE*, *5*(10), e13636. doi:10.1371/journal.pone.0013636

Houghton, J. (2010). *Economic and Social Returns on Investment in Open Archiving Publicly Funded Research Outputs, report to SPARC*. Centre for Strategic Economic Studies Victoria University.

Lipman, D. J. (2010). *Testimony, Public Access to Federally-Funded Research*. Washington, D.C. Retrieved from http://www.hhs.gov/asl/testify/2010/07/t20100729c.html

Maier, D., Kalus, W., Wolff, M., Kalko, S. G., Roca, J., Marin de Mas, I., Turan, N., et al. (2011). Knowledge management for systems biology a general and visually driven framework applied to translational medicine. *BMC Systems Biology*, *5*, 38. doi:10.1186/1752-0509-5-38

Mons, B., & Velterop, J. (2009). Nano-Publication in the e-science era. Presented at the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009). Retrieved from http://www.nbic.nl/uploads/media/Nano-Publication_BarendMons-JanVelterop.pdf

National Advisory Council on Innovation and Entrepreneurship, Department of Commerce. (2011, April). Letter to Secretary Locke: Recommendations to facilitate university-based technology commercialization. Retrieved from http://www.eda.gov/PDF/NACIE_Letter-University_Commercialization.pdf

SHERPA/RoMEO - Publishers allowing the deposition of their published version/PDF in Institutional Repositories. (n.d.). Retrieved December 20, 2011, from http://www.sherpa.ac.uk/romeo/PDFandIR.php?la=en

SHERPA/RoMEO - Publishers with Paid Options for Open Access. (n.d.). Retrieved December 20, 2011, from http://www.sherpa.ac.uk/romeo/PaidOA.html

Suber, P. (2008, August 2). Green/gold OA and gratis/libre OA. *Open Access News*. Retrieved December 20, 2011, from http://www.earlham.edu/~peters/fos/2008/08/greengold-oa-and-gratislibre-oa.html

Swan, A. (2010). The Open Access citation advantage: Studies and results to date. Retrieved from http://eprints.ecs.soton.ac.uk/18516/

Taylor & Francis Author Services - iOpenAccess & NIH policy. (n.d.). Retrieved December 20, 2011, from http://journalauthors.tandf.co.uk/beyondpublication/iopenaccess.asp

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