To: OFFICE OF SCIENCE AND TECHNOLOGY POLICY

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

Date: 1/6/2012

From: Johns Hopkins University Scholarly Communications Group
Robin N Sinn, Chair, Research Services Librarian, Sheridan Libraries
Julie Adamo, National Library of Medicine Associate Fellow, Welch Medical Library
Mark Cyzyk, Scholarly Communication Architect, Sheridan Libraries
David Reynolds, Manager of Scholarly Digital Initiatives, Sheridan Libraries
Gabrielle Dean, Research Services Librarian, Sheridan Libraries
Claire Twose, Associate Director, Public Health and Basic Science Informationist Services, Welch Medical Library
Sue Woodson, Associate Director, Digital Library Services, Welch Medical Library

---

1a. 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?

Most importantly, agencies can develop policies that support free, immediate, and complete access to publications resulting from federally-funded research. Supporting the complete reuse of articles promotes the development of new services and products. This type of commercialization is far more probable in a fully open and accessible environment where content is fully reusable. Complete access makes cutting-edge ideas and research accessible to a much greater number and variety of users. Organizations, companies, and individuals who are not part of the traditional research enterprise have more opportunity to develop unforeseen applications out of publicly funded research. This increased commercialization supports economic growth, by encouraging private investment to capitalize on a government resource. Furthermore, unrestricted access promotes social growth and development by allowing all people, regardless of institutional affiliations or social status, to learn from the research that has been funded by their tax dollars.

1b.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?

Open Access increases the citation count of an article and promotes follow-on research, supporting growth, advancement, and diversity of knowledge (Eysenbach, 2006). This open accessibility of information accelerates advancement in knowledge and research by making it easier to incorporate new findings into research and writing, and opening doors to follow-on research to an unrestricted pool of users (See testimony from David Lipman, MD, Director of the National Center for Biotechnology Information: http://www.hhs.gov/asl/testify/2010/07/t20100729c.html). Open Access facilitates interdisciplinary research by making information available to users outside of traditional knowledge silos. It also facilitates use by non-profit research and advocacy organizations who traditionally have had limited accessibility to research.

Additionally, Open Access enables machine reading and computer-driven research, data mining, text mining and other methods that were impossible prior to the digital environment. This is another example of how Open Access provides opportunities for commercial development through the enabling of computer-driven research. This type of research is better supported and facilitated by an unrestricted environment.

Eysenbach G (2006) Citation Advantage of Open Access Articles. PLoS Biol 4(5): e157.

1c.What are the relative costs and benefits of such policies?

Research conducted through SPARC shows that an estimated five-fold increase in return on investment will result from opening access to all U.S. publicly funded scientific articles, while an Open Access policy similar to the NIH policy will generate benefits approximately eight times larger than the costs.

There is considerable infrastructure stemming from databases such as PubMed Central from the National Library of Medicine, that are already in existence, that can be further developed to support expanded Open Access policies. Building on existing infrastructure will minimize costs through reducing duplication of effort.

PubMed Central currently receives over 500,000 users per day, a statistic that embodies the great demand and potential for freely available research. A vast user base also serves as a means to increase transparency of, and accountability for, research investments, building upon the capacity of federal agencies to account for the outcomes of the research that they fund. There is also potential for positive impacts on policy, through improved access to information that guides policy decisions.

1d. What type of access to these publications is required to maximize U.S. economic growth and improve the productivity of the American scientific enterprise?

Complete, immediate, and free Open Access, with rights for re-use, will maximize the value of research, and return to taxpayers. Restricting access will limit the extent and types of follow-on research that can be conducted. Possibilities for computer-based research are maximized in an unrestricted environment. Placing restrictions on access also creates an unequal grounding for diverse groups of researchers, allowing those with full access to have greater opportunity than those with restricted access. Furthermore, full reuse prevents duplication of research efforts, making it possible to extract value on initial investments far into the future.

2a. 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?

2b. 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?

A public access policy for scholarly publications resulting from federally funded scientific research could easily be crafted to enhance access for many sectors of the public without infringing on the intellectual property interests of rights holders. Current copyright laws would enable rights holders to control initial dissemination of the research, re-use that work, and to produce derivatives work. A public access policy would not undermine any of these rights, but it would enable more researchers and the general public to put the results of the research to good use. The "facts" of the research are not protected by copyright now, only their expression. Wider and quicker dissemination of these facts would likely result in more scientific and commercial applications for the research. There is no evidence that pre-print services such as ArXiv have harmed commercial publishers. Scientists still publish their papers in peer-reviewed journals, but their work is also made available to researchers who cannot subscribe to such journals. The Johns Hopkins library has not cancelled any journal subscriptions as a result of the availability of pre-print services, and we do not know of other libraries that have done so. A policy that utilized the Creative Commons CC-BY license would enhance access to publicly funded research while still protecting copyright holders. It would also provide researchers the right to reuse research data in new and interesting ways. This could be combined with a reasonable embargo period to protect the value of the original publication to publishers.

3a. 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?

3b. 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 pros with respect to a centralized service include the following: Centralization of the service presumes centralization of the administrative functions of the service. Having a service managed in a single, centralized facility makes for a simpler service. Centralization also enables direct control of the service and easier provision of tools for search, analytic tools, etc. In short, a centralized service is administratively a simpler service. However, the downside here is that the burden of the service as a whole will likewise be focused and centralized.

The pros with respect to a decentralized service include the following: A decentralized service can likewise be administered in a decentralized manner, pushing administrative tasks out to participating partners. This relieves the administrative burden on a central entity. However, the downside here is that the functions surrounding the service, search, analytic tools, etc., become more difficult to implement. Interoperability among and between participating nodes in the service becomes crucially important. And this interoperability will inevitably require some sort of administrative and technical "glue" to hold it together. The administrative glue here will need to consist of a well-thought-out set of rules and policies to which each and every participating partner in the decentralized service must conform. Likewise, from a technical perspective, there must be a small set of standards gluing the system together and enabling the various nodes in the overall system to interoperate. A good example of such technical "glue" is something called the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH is a way of exposing metadata provided by a network service such that it can be "harvested" and ingested into a centralized index enabling search functionality across the entire set of harvested materials. So the data and publications themselves, the content, reside in a decentralized network of nodes; yet this content can be centrally searched. The glue of OAI-PMH makes this possible.

A hybrid model of centralization is also possible. One scenario here would be for the content to be decentralized, the metadata (data about the content) to be centralized, and a parallel system aimed at long-term, trusted preservation of both content and metadata be set up. One such preservation system, LOCKSS (Lots of Copies Keep Stuff Safe), does just this. LOCKSS is a federation of nodes holding encrypted copies of content -- multiple, duplicate, encrypted copies of content. The LOCKSS network itself actively monitors the health of each copy of content such that, if one copy degrades, it is automatically overwritten with a known-good copy housed elsewhere in the LOCKSS network. In this way, content on the publicly-available nodes in the system can be thought of as working copies; metadata about that content in a centralized service can provide access to such content; and the copies of the content held in the preservation system can be deemed archival copies, copies to be used in case a working copy on the public network degrades to the point of being unusable.

At the least, it seems prudent for federal agencies to maintain preservation copies of content. This could be accomplished, as illustrated above, as simply as setting up a LOCKSS network into which a copy of content is deposited.

Whether a centralized or decentralized approach is used, the federal government will generate and enforce the rules of the system: open access to content and the long-term availability of the content. A decentralized approach will need to stress interoperability and will also encourage innovative partnerships with other entities.

The policies that are developed must give federal agencies the right to store and make freely available publicly funded articles. If the content is to be stored across multiple sites, public or private, the agencies will use contracts to outline the technical and legal requirements for the participation of private entities in the provision of services. Long-term stewardship will be part of that contract.

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?

In order to build a national system of open repositories that feature interoperability, robust metadata, excellent search functionality, ingestion, and consistent policies and procedures, public-private partnerships are necessary. The federal government, publishers, libraries, archives, universities, and commercial businesses will need to work together to provide the infrastructure and services needed to make this project successful.

Openness is one aspect that must be stressed. The content in the repositories must be open to the public; code should be available to both commercial and non-profit entities for service development.

One possible model is DRIVER (Digital Repository Infrastructure Vision for European Research), a set of distributed repositories that share a mission to support European research and public access.

Another model is PubMed and PubMed Central. Both are open. PubGet and Quosa have built products that utilize the information in both PubMed and PubMed Central.

5a. 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?

5b. 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?

Proper metadata is vitally important for maximizing the utility of federally funded research. Metadata is more than a way to facilitate discovery of research articles—it must also enable researchers to use the publications and underlying data in new and interesting ways. Services such as data mining and visualization could greatly increase the value of the research. In order to perform such actions, the metadata must be in a common, machine-actionable format such as XML. Whatever standard or standards are used must go beyond minimal data such as Dublin Core, and include richer descriptions and relationships. The NLM DTD Journal Archiving and Interchange Tag Set Library shows much promise as a minimal standard. The use of discipline-specific controlled vocabularies for elements such as subjects will enable a high level of interoperability.

Whichever standard or standards are chosen, it is essential that they should be allowed to evolve. Emerging standards should constantly be examined for relevance. As long as interoperability is considered in their adoption, there is no reason to stick with a static set of metadata standards forever.

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?

Consistency of the requirements engendered by a mandate for broad public access and long-term stewardship of research articles will be fundamentally important to the success of this program. Uniform requirements across granting agencies will enable the most gain for the public, for research, and for business, while also leveraging the work done by researchers, publishers, libraries, institutions, and entrepreneurs.

- Researchers, institutions, and publishers will follow one set of uniform requirements to comply with the law. This eliminates complexity and duplication of effort, increasing compliance.
- Businesses and entrepreneurs who wish to build applications and services using the information in the repositories will be able to work with a large system working under one set of rules. This means that their work will appeal to larger markets, thus creating more incentive for them to do the work in the first place.
- The public, as well as librarians, researchers and educators, will find information more quickly and reliably since search engines won't have wildly different types of repositories to work with.
- The granting agencies will have the opportunity to work together on this system and create something truly efficient. One system for multiple agencies would allow them to share the staffing, training, and funds to support that system. This will keep costs down and encourage high compliance rates.
- Higher compliance rates will provide the public, libraries, institutions, researchers, educators, and entrepreneurs with more information to work with and on. It's a positive feedback cycle that encourages growth and participation.

The policies created to support the public access policy should incorporate certain beneficial characteristics.

- Use of existing and developing IT protocols will increase interoperability and save time. SWORD is an example of an extant protocol that can be incorporated into the new policies, saving time and increasing compliance. ORCID is in development and will individually identify authors and link them to their published work. Optimizing the information for searching by Google and Google Scholar will improve the discovery of this information.
- New ideas and products should be encouraged. IT changes constantly, so the repositories created by this program should be structured for easy adaptation of new technologies. Ongoing programming work will need to be part of the infrastructure.
- Agencies and institutions can create policies and infrastructure that allow them to integrate the articles arising from grants with their grant management programs. This will improve tracking and reporting for both groups. Institutions and agencies will be able to assess grants, PIs, and research teams using the information made available from this integration. The public will also be able to see what their tax dollars are supporting.
- This integration can also help individual researchers keep their CVs, bibliographies, and PI profiles up-to-date, thus improving efficiency for all involved in the grant process.
- These open repositories will provide many educational opportunities. Librarians will be able to teach their patrons better ways to search the scholarly literature. New ways to measure research productivity will emerge and need to be taught as well.

Consistency, openness, interoperability, and integration will make this a very successful program.

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?

Other types of publication arising from federal funding should be publicly accessible. Those other publication types may need a different set of policies than journal articles. Authors are generally paid for books and book chapters, while journal article authors are not paid. This is a fundamental difference that requires separate policies for open access.

Conference proceedings should also be considered separately from journal publishing. There are enough differences between conferences and journals that it would be too difficult to fold conference proceedings into the current discussion. Some conferences only publish abstracts, some publish papers of the presentations made. Some treat the proceedings as a journal, others as a book series. Often, conference proceedings aren't published until 2 or 3 years after the event.

8a. 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.

8b. Are there evidence-based arguments that can be made that the delay period should be different for specific disciplines or types of publications?

Since scientific research moves so rapidly, its value to researchers is greatest soon after publication. Therefore, it is important to make the results of publicly funded research widely available as soon as possible. Such prestigious publishers as the American Institute of Physics, the American Mathematical Society, and the New England Journal of Medicine allow authors to make their published research available through institutional repositories or other free-access e-print servers either immediately upon publication or after a six-month embargo period. They have reported no reduction in number of subscriptions as a result of this policy. In fact, no publishers have presented evidence that a six-month or less embargo period on public access to research has harmed them. As mentioned in a previous question, the Johns Hopkins Libraries have not cancelled any journal subscriptions as a result of the availability of journal articles in open access repositories. There are many factors that contribute to journal cancellations, but those relate to overall budget constraints and competitive pricing of rival journals.