Name/Email
Response to "Request for Information: Public Access to Peer-Revi
Publications Resulting From Federally Funded Research, Novemb
Prepared for and the official RFI response of the University of Ore
JQ Johnson
Director, Scholarly Communications & Instructional Support
Professor, University of Oregon Libraries
University of Oregon
email: jqj@uoregon.edu
[corresponding author]
Dean Walton
Science Librarian and UO Libraries Biology subject specialist
University of Oregon
email: dpwalton@uoregon.edu
Deborah A. Carver
Philip H. Knight Dean of Libraries
University of Oregon
email: dcarver@uoregon.edu

Affiliation/Organization
University of Oregon Libraries

City, State
Eugene, OR

Comments

Prefatory note: The following comments begin from the perspective that the most natural and preferred federal policy for the next few years includes extension of a system similar to the NIH public access mandate to multiple federal agencies to promote public access, and that the major questions involve variants on that scheme. Our overall belief is that the NIH public access mandate was an excellent start in 2008, but that as the world has changed it requires extension and refinement to further meet the overarching public access goals and take advantage of changes in the scholarly publishing industry.

Comment 1

(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?

The NIH mandated deposit policy has had a major positive impact on creating new markets and models for peer reviewed publication within the biomedical community, and needs expansion to additional agencies. For example, it appears to have been the major driver for the development of new publishing enterprises such as PLoS and BioMed Central. Overall, the rapid growth in the number of "gold" open access journals, particularly in the life sciences, indicates that the NIH policy is having a positive effect in creating new markets, and hence is one that should be emulated by other federal agencies. As the purchase of BMC by Springer last year indicates, these new enterprises are clearly seen as potentially valuable by more traditional publishers, and in fact the traditional publishers are rapidly moving to invest in their own open access journals under brands such as "SpringerOpen," "Wiley Open," "Sage Open," the Taylor & Francis "Author Rights Initiative," etc. Some traditional publishers are beginning to realize that public access to selected articles in their (hybrid) journals is essentially free advertising, whereas others appear to have identified open access publishing as a potentially profitable business model. Most importantly, note that these new commercial ventures are currently growing exponentially, with no evidence that growth is anywhere near the flattened top of the logistic (Laakso et al., 2011); we can expect substantial further expansion in this market as long as agency policies do not choke off this growth.

Although creating new markets and business models for peer reviewed publications is an important goal, it is part of a larger goal of using publicly funded research to spur increased commercialization (Houghton & Sheehan, 2006). This goal will be best served by creating opportunities for innovation in the use of scientific information, rather than focusing narrowly on peer reviewed publication. In particular, it is vital that the full text of articles, as well as other products of the funded research such as specimens and datasets, be made freely accessible without commercial restrictions. Having access to this corpus will allow new innovative uses such as text mining, synthesis and abstracting services, visualization tools that allow researchers to identify new interdisciplinary connections and reviewers to assess scholarly impact in new ways, etc. It spurs economic growth not just in the publishing industry but throughout the economy. Ultimately, the economic benefits accrue because open access drives scientific and technological innovation by increasing readership (e.g., as measured by increased citations to OA publications (Wagner, 2010)), decreasing time to impact, and promoting diversity in follow-up research by making research widely available outside of narrow disciplinary areas.

Estimates of direct R&D benefits to opening access to all articles derived from U.S. publicly funded research suggest a 5X return on investment (Houghton, Rasmussen, & Sheehan, 2010).

It is gratifying to see the degree to which public access policies have spurred innovation and creativity in what appeared a decade ago to be a very conservative academic publishing industry that was not responding rapidly to the opportunities afforded by the world wide web. Examples of innovation that have been driven by the NIH deposit policy abound. For example, it seems doubtful that PLoS One, with its innovative model of peer review and journal funding, would have been successful absent funding agency mandates. And PLoS One has certainly been successful – it is expected to publish more than 14,500 peer reviewed articles during 2011, making it the largest peer-reviewed journal in the world.

One may also point to Google Scholar as a successful commercial (advertising-funded) service that depends on the existence of a large corpus of open access journal articles including PubMed Central.

More directly, the existence of a corpus of freely accessible publications is beginning to result in new tools where "readers" are computers rather than individuals. One typical example from the technology sector is recommendation engines such as those used by Google or Amazon (to offer improved search results based on mining an individual user's previous search patterns and data on global popularity and click-thrus from particular searches). Another example is publishing tools for detecting plagiarism or duplicate publication such as iThenticate (iParadigms, 2011), almost all of which mine open access databases including PubMed as one of their resources. Applications that mine publication databases and produce meta-analyses or visualization of the pattern of scientific results from multiple studies have been in existence for substantially more than a decade (Kostoff & DeMarco, 2001), and indeed the field even has its own OA journals *(*e.g., *BMC Bioinformatics*), but the tools are maturing with the availability of more open publication data. Some well known examples include BioCreative, CoPub, and PubGene (Krallinger, Valencia, & Hirschman, 2008). Another promising new example is the new openSNP system (Greshake, Zimmer, Rausch, & Bayer, 2011), which in addition to collecting open data on genotypes correlates that data with relevant references in open access publications.

Concrete examples of additional steps that could be taken to improve access and analysis of peer reviewed publications include:

- Expanding the corpus of open access articles available in central and standardized repositories to a much wider range of disciplines and research areas, by collecting and making available with standardized interfaces copies of journal articles that would otherwise not be open access and extending NIH-style mandates to additional funding agencies;

- Clarifying that public access is based on a limited and nonexclusive rights transfer via prospective license from funded authors as part of federal research contracts, prior to any transfer of additional rights that an author may agree to as part of a publication agreement and hence not based on any theory of imminent domain or taking of publisher rights;

- Providing consistent programmatic interfaces and metadata that foster text mining of multiple simultaneous text collections;
- Providing more convenient mechanisms that as a supplement to existing mechanisms allow authors to deposit articles in institutional repositories, with automated harvesting using SWORD, OAI or similar protocols into centralized, federally maintained, archives (Harnad, 2008);
- Assuring widespread public and scholarly access (to preprints, to peer reviewed and accepted manuscripts, and to formal – often publisher maintained – copies of record) to maximize serendipitous and interdisciplinary discovery;
- Encouraging the development of "gold" open access publishing through appropriate funding, both via direct and automatic grant funding to extramural researchers to pay article processing charges, and via grant programs to stimulate the creation of new open access peer reviewed publications or to explore technologies, new business models, and new models of peer review that could further reduce the overall cost of publication and dissemination;
- Providing incentives to researchers to review manuscripts submitted to venues that make articles available open access;
- Investing in research into software and tools for text mining and visualization; collaborations between other agencies and those such as NSF and DARPA that have been traditional funders of machine learning and computer science research seem particularly fruitful;
- Investing as part of the federal grant funding and evaluation process in the internal use of text mining tools for identifying promising areas for future research, for example by identifying newly "hot" topics and research areas that cross disciplinary boundaries;
- Developing and promulgating standards for metadata that facilitate discovery and broad reuse, plus standards for programmatic access to publication data both for text mining and for mirroring of collections.

Specific costs and monetary benefits of providing access to publications depend heavily on the details of the implementation. We assume a mixed strategy where publishers may maintain a copy of record and where in many cases universities provide additional copies as part of their institutional repositories with tools that allow mirroring of articles and archives for preservation and specialized needs, but where the federal government, with its interest in guaranteeing access, provides access to additional copies and is able to assure consistent accountability, metadata, and access standards across multiple federal agencies. This NIH-style model has proven quite cost effective: NIH reports costs of $3.5 to $4.0 million per year (about

1/10,000 of the total NIH budget) to provide access (Lipman, 2011). Generalizing it to other agencies would leverage existing infrastructure and minimize costs and provide consistency in user interfaces. Note, though, that centralized archiving in the physics and mathematics communities (arXiv) is perhaps a factor of 3 even less expensive, so there is room for improvement. On the benefit side, calculating only the direct benefits of a FRPAA-style policy on improved U.S. R&D, Houghton et al. calculated that net present value of benefits over a 30 year period would be in the range of $1.6B to $1.75B, a factor of 4 to as high as 24 benefit-cost ratio depending on assumptions for the efficiency of providing the centralized access (Houghton, et al., 2010).

It is important to note here (for further discussion see Comment 2) that at a minimum the benefits of public access require unrestricted read access to a comprehensive collection of the texts of at least the author final versions of articles, plus the right of other researchers to use such articles in ways consistent with typical academic practice such as quoting from them. To the extent that the corpus is incomplete or embargo-limited it introduces serious problems with possible bias as some studies are ignored. To the extent that the access is limited by subscriptions or usage constraints it decreases the probability that new entrepreneurs will enter the market and may make it impossible to mirror archives for preservation or experimentation with new access approaches. Almost certainly further improvements in the use of federally funded articles will require standardization on a generally-accepted form of license that allows greater usage rights than the current NIH-mandated minimum. In the 3 years since the NIH mandate was established, Creative Commons licenses have emerged as the clearly preferred standard throughout the world, with hundreds of millions of works now released under Creative Commons licenses (Creative Commons, 2011).

It is not yet completely clear whether adequate commercialization opportunities will exist if works are made available under a license such as the minimalist Creative Commons CC-BY-ND (which does not even appear to grant any right to quote portions of an article beyond what is allowed by fair use); there are growing and compelling arguments that rich data mining and visualization require that authors also allow creation of derivative works. Certainly an application that used the corpus of Pub Med Central articles to provide improved machine translation of articles (perhaps from Mandarin to English, a tool that would be extremely valuable given the rapid growth of important scientific literature published only in Chinese) would require such rights. This suggests that although standardization on a CC-BY-ND license in 2012 might be politic, it is likely to be important in the future to make works available under a CC-BY license or equivalent to maximize the likelihood both of new applications and of the successful commercialization of derivative products.

References

[Please observe that works cited in this response are except as explicitly noted to works released as open access.]

Creative Commons. (2011). Creative Commons [web site] Retrieved 6 Dec 2011, from http://creativecommons.org

Greshake, B., Zimmer, F., Rausch, H., & Bayer, P. (2011). openSNP Retrieved 5 Dec 2011, from http://opensnp.org/

Harnad, S. (2008). *Optimize the NIH Mandate Now: Deposit Institutionally, Harvest Centrally*. Technical Report. Electronics and Computer Science. University of Southampton. Retrieved from http://eprints.ecs.soton.ac.uk/15002/

Houghton, J., Rasmussen, B., & Sheehan, P. (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, J., & Sheehan, P. (2006). *The Economic Impact of Enhanced Access to Research Findings*. CSES Working Paper No. 23. Centre for Strategic Economic Studies. Victoria University. Melbourne, AU. Retrieved from http://www.cfses.com/documents/wp23.pdf

iParadigms. (2011). Plagiarism Checker | Plagiarism Detection Software from iThenticate Retrieved 7 Dec 2011, from http://www.ithenticate.com/

Kostoff, R., & DeMarco, R. (2001). Extracting information from the literature by text mining. *Analytical Chemistry, 73*(13), 371-379. doi: 10.1021/ac012472h

Krallinger, M., Valencia, A., & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology, 9*. doi: Artn S8. DOI: 10.1186/Gb-2008-9-S2-S8

Laakso, M., Welling, P., Bukvova, H., Nyman, L., Bjork, B. C., & Hedlund, T. (2011). The Development of Open Access Journal Publishing from 1993 to 2009. *Plos One, 6*(6). doi: ARTN e20961. DOI: 10.1371/journal.pone.0020961

Lipman, D. J., M.D. (2011). *Testimony on Public Access to Federally-Funded Research* Committee on Oversight and Governmental Reform, Subcommittee on Information Policy, Census and National Archives. United States House of Representatives. Retrieved from http://www.hhs.gov/asl/testify/2010/07/t20100729c.html

Wagner, A. B. (2010). Open access citation advantage: An annotated bibliography. *Issues in Science and Technology Librarianship*(60).

Comment 2

(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 existing copyright and IP licensing framework provides adequate tools for protecting IP interests. However, since the IP interests of various stakeholders may not always be aligned it is necessary to balance competing interests. In addition, some of the defaults embodied in copyright law and traditional practice need to be avoided to maximize benefits.

One occasionally still sees arguments of the form that the government has no business mandating public access, arguing that such publications are funded by the private sector rather than by the federal government. Such arguments not only ignore the legislative mandate of COMPETES but miss the point on at least three grounds: the vast majority of the costs associated with publication are in the research that it reports and in the nominally-free services such as reviewing that are provided by the academic community. Remaining costs are indeed funded by a variety of sources most notably libraries that pay for subscriptions, but many of those libraries are not usually considered part of the "private sector." Secondly, it ignores the recent growth in robust gold open access publishing models such as that of PLoS and BioMed Central. Thirdly, such consideration is irrelevant to an IP analysis.

Equally arguably the only constitutionally mandated stakeholders with an IP interest are authors and inventors, though the constitutional mandate for copyright is justified by the goal of promoting "the Progress of Science and useful Arts," implicating the public (and public access) as an important stakeholder. The primary IP interest for present purposes is that of the authors who produce new peer reviewed publications, and their interests are primarily in assuring that others make widespread and immediate use of their work within a framework that discourages misuse such as misquotation or use without attribution. The interests of this group of stakeholders largely align with the interests of consumers of federally funded research results, including other scientists, federal agencies and Congress, university tenure committees, library archives, entrepreneurs interested in commercializing reported discoveries, and the general public. In general, these stakeholders need at a minimum the right to read and copy works both individually and in bulk, and the right to make use of the knowledge contained in such articles to produce new discoveries and new articles.

For these groups of stakeholders the various Creative Commons licenses (Creative Commons, 2011) such as CC-BY are consistent with current U.S. copyright law and provide an excellent basis for effectively meeting the needs of authors and the public. The CC family of licenses have numerous advantages for use with academic works, for example careful handling of moral rights through a "no endorsement" clause applied

to derivative works, and of digital rights management that might otherwise circumvent the author's intention to make the work publicly available. There is clear community agreement that the CC family of licenses are well crafted and of small enough number to provide standardization in license terms, a feature highly beneficial to effective public access by non-lawyers.

Our own reading of typical copyright transfer agreements that authors routinely sign when publishing on traditional journals suggests they often are not well crafted to preserve the rights authors want. For example, they do not adequately preserve moral rights, and often include by reference external publisher websites that are subject to change and sometimes internally inconsistent. It is in the public interest to assist authors in clarifying the rights they actually want to transfer to publishers, but it is worth reiterating that the NIH mandate and related policies are fundamentally rooted in the IP interests that exist before such transfers.

There is currently some discussion in the academic community (McLennan & Malenfant, 2011) about whether a very restrictive license such as CC-BY-ND provides adequate reuse rights to allow robust text mining and computation on the corpus of publications while further protecting author needs for attribution and integrity or whether a license such as CC-BY is needed for effective academic use of archived works; conversely, it may be that an even more open (though in some senses more restrictive) license such as CC-BY-SA is needed for effective widespread use of derivative works created from publications. Given that our understanding of the actual needs of our users is evolving along with the types of uses that are technically realistic, a reasonable compromise strategy would be for federal mandates in 2012 to specify a minimum access corresponding to, or preferably explicitly referencing, CC-BY-ND, but also (a) provide metadata as part of any dissemination that describes the particular license adequately to allow readers and text mining applications to determine their rights, (b) encourage authors to license their work under the most non-restrictive terms they feel comfortable with (either using CC-BY or dual licensing under CC-BY-ND and some other orthogonal license, the way many open source software projects do), and (c) reevaluate the need for more open licenses as the community obtains greater experience with data mining and computational applications.

Apropos of the need for a bare minimum of CC-BY-ND, we observe that academic readers at a minimum need rights unambiguously sufficient to allow uses consistent with current academic practices by researchers at rich universities who can afford subscriptions to journals. Routine practice, for instance, is often to make copies of licensed articles to distribute to students in a graduate seminar. Similarly, needs by the general public are not yet well understood but likely extend beyond minimal access. The first author on this comment is currently undergoing chemotherapy, and observes information usage patterns in his oncologist's office; in that office the doctors do not subscribe to most journals and do not appear even to access relevant articles on PubMed Central even if such access would affect their treatment decisions. Failure to read PubMed Central articles seems largely an issue of convenience; they would be much more likely to do so if an assistant found relevant current research and distributed electronic copies to the doctors in the practice rather than suggest that they go beyond their computing abilities to view PubMed Central directly.

Apropos of CC-BY-ND versus CC-BY, the University of Oregon is an example of a typical academic publisher, with the University Libraries publishing 3 open access scholarly peer reviewed journals. We initially required that our authors license all articles using a CC-BY-ND (no derivative works) license, but are in the process of a transition to consistent use of CC-BY licenses. We have concluded as publishers that the public interest and our own interest as a publisher is only served by granting full reuse rights to readers, and that that license matches the desires of most of our potential authors, thereby improving our chances of publishing high quality work. Commercial publishers whose business model charges subscription fees for access to scholarly works may have a somewhat different set of needs, particularly for works where their business model still requires that they acquire copyright from the authors. In order to ensure the continued value of the publisher's copy, it may be desirable that full reuse rights apply at least to the author's final manuscript version, but that publishers have available the right to impose somewhat more restrictive rules governing the final published version. IP policies must also require that any use of or reference to works include a citation that references the publisher version as the copy of record. Some publishers have argued the need for an embargo period where public access to any version of the author's work is restricted. There is very little data to support the claim that embargoes actually strengthen the financial position of the publisher or influence the willingness of libraries to subscribe to a journal, and clear data that they decrease both visibility of the articles and the benefits of open access in fostering R&D (Houghton, Rasmussen, & Sheehan, 2010). However, it may be that short embargoes can be shown to be needed. If so, a reasonable way to balance interests for new peer reviewed works is to follow the apparently-working NIH PubMed Central model and allow for a short (less than one year) author specified embargo period, with full reuse rights (e.g., CC-BY licenses) applying after the end of the embargo period. Please see Comment 8 below for further discussion.

Under the current implementation of the NIH public access rules, legal compliance responsibility falls primarily on the institutions that contract with NIH. It would be very desirable to provide standardized mechanisms to clarify and shift that responsibility to individual authors. One mechanism for doing so would be a small change to the standard contract terms that would require institutions to obtain on the government's behalf from all potentially affected employees a prospective (perpetual but nonexclusive) grant of license to works affected by the policy. Such a license would make explicit that the authors not only commit to depositing a copy of their work but grant to the federal government the right to public display or, as suggested here, the right to sublicense the work to the general public under the appropriate CC license. Such license terms would also be easy for institutions to implement as part of their standard employment contracts.

Additional specific steps that can be taken to protect copyright interests include:

- Enforcement of the standard federal requirements (National Science Foundation, 2011) for including in grant funded publications the detailed acknowledgement of the granting agency, including the award number (which is useful data in text mining and meta-analyses), plus extension of these rules to require that research publications produced by U.S. Government employees

be comparably labeled to make clear that such articles are in the public domain, and that additional metadata be provided;

- Careful documentation and systematic publication as part of publicly accessible article metadata of the IP rights retained by each party in articles;
- Education for authors about their rights and copyright responsibilities;
- A requirement that citations to works deposited in federal archives always include a full citation to the publisher copy of record, including DOI if available, perhaps with the limitation that publishers may waive this requirement;
- Prohibition on the transfer to federal repositories of copies of scholarly articles that contain technological protection mechanisms – under current copyright law defeating a technological protection mechanism is illegal, but clearly imposing such mechanisms would undermine the goals of policies such as the NIH deposit mandate.

A copyright-related policy that should not be implemented is the current PubMed Central prohibition on systematic downloading, which needlessly restricts large classes of use and imposes restrictions beyond those mandated by the copyright or license terms embodied in individual articles.

Another important class of IP interests derives from patent law. In many cases the innovations flowing from federally funded research will result in patents held by researchers, their institutions, or commercial entities. In most cases such rights do not conflict with the goals of a public access policy, though in some they may imply an additional need for a temporary embargo during filing. However, such rights do potentially limit the rights of the public to use results reported in peer reviewed publications. For example, in a case currently being considered by the U.S. Supreme Court (Mayo v. Prometheus) it is argued that a reported correlation between two variables and its usefulness in creating a cancer diagnostic limits the right of doctors to even discuss the correlation with patients (Anderson, 2011; Barnes, 2011; Lee, 2011). At a minimum this suggests that authors should be required to be transparent about any patent claims they or their institutions and assignees make associated with research they report, since such claims have an impact on how readers can use articles.

Currently, PubMed Central is in a somewhat awkward position with respect to licensing, since it contains works subject to a wide variety of usage restrictions. Some articles are embargoed and the license granted by the copyright owner to NIH does not even permit public viewing; many articles may be viewed but not copied despite the fact that typical PC viewing software often requires that the web browser make a non-transient downloaded copy of the file in order to display it to the user; other works are released under licenses that allow some copying but no further use, and still others allow a variety of uses. In most cases it is appears to be unclear to users what rights they have to individual articles. PubMed Central contributes to this confusion by pointing to the copyright statements included in individual articles,

which often do not note the additional rights that authors and publishers have granted to PMC or the public as part of the deposit process, or to external publisher websites that often provide confusing or contradictory information. Greater standardization of minimal licenses is needed, at the very least to the point where a user can reuse copies of individual articles and where the default absent explicit terms to the contrary is an attribution requirement that meets traditional academic standards for avoiding plagiarism.

References

Anderson, J. (2011, Dec 08, 2011). Summary of Mayo v. Prometheus Oral Argument. Retrieved from http://www.patentlyo.com/patent/2011/12/summary-of-mayo-v-prometheus-oral-argument.html

Barnes, R. (2011, Dec 7, 2011). Supreme Court has hard time finding an easy test for patents on medical processes, *Washington Post [online; not available open access]*. Retrieved from http://www.washingtonpost.com/politics/supreme-court-has-hard-time-finding-an-easy-test-for-patents-on-medical-processes/2011/12/07/gIQAneUldO_story.html

Creative Commons. (2011). Creative Commons [web site] Retrieved 6 Dec 2011, from http://creativecommons.org

Houghton, J., Rasmussen, B., & Sheehan, P. (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

Lee, T. (2011, Dec. 15, 2011). Medical Mind Control. *Slate Magazine [online]*.

McLennan, J., & Malenfant, K. (2011, December 2, 2011). Getting the rights right: Next SPARC-ACRL forum at ALA announced Retrieved 11 Dec 2011, from http://www.arl.org/sparc/media/11-1202.shtml

National Science Foundation. (2011). *Research Terms and Conditions, June 2011*. National Science Foundation Retrieved from http://www.nsf.gov/pubs/policydocs/rtc/termsidebyside_june11.pdf.

Comment 3

(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?

Centralized deposit of works in which the federal government has an interest has a long and successful tradition, starting with the creation of the Library of Congress and the requirement in 17 USC 407 that copyright holders deposit copies of a printed work as part of the copyright registration process. Note that as of 2010 the Library of Congress also has begun an on-demand requirement for deposit of online-only serials (Peters & Billington, 2010). Such deposit demonstrates the benefits of federally maintained centralized approaches, and addresses some issues of long term archival, but does not by itself meet the needs of public access.

It is important to distinguish among a variety of uses for corpi of peer reviewed manuscripts and to develop approaches that support the variety of uses, where different uses may require multiple deposit copies and where only some uses require centralized custody associated with the funding agency.

For long-term stewardship of federally funded research publications, the federal government is the appropriate long-term custodian. We believe that the preferred approach to maintaining public access is either a single managed repository – perhaps a somewhat expanded version of PubMed Central -- serving multiple federal agencies or is a distributed repository where multiple agencies each maintain copies of publications funded by that agency but do so in a software context that appears to the user as much as possible as being a single federated repository with uniform access and user interface characteristics.

The federal government has a continuing interest in making such works permanently available, and is the only current player with an interest in making the full corpus of federally funded works publicly available, in providing tools for using the collection as a whole, and in assuring that new services and products can be built from publicly funded information. Other private players including universities with their institutional repositories, disciplinary societies with their discipline-specific archives, and commercial and noncommercial publishers with their journal-specific collections have not in general demonstrated an ability to scale at low cost or to provide broad public access the way, for example, PubMed Central has.

Commercial entities and non-profit volunteer efforts in particular are at risk of being unable or unwilling to provide long-term (multiple-decade) access as their business models change or organizations go out of business. One need look no further than recent news stories such as the November 2011 changes in Amazon's Penguin e-book lending program (Van Camp, 2011) to realize that commercial entities have short term business incentives that may conflict with long term access. Similarly, there are lingering concerns about quality control in commercial journals (Grant, 2009).

Commercial entities also may have a conflict of interest when it comes to providing access that could allow competitors to develop new services based on an archive. Entities that have an interest in only a centralized "dark archive" in particular are not in a position to provide a viable solution since such an archive neither meets archival needs (regular access and use are vital to maintaining archival veracity) nor the public interest in access.

Centralized repositories make it easy to find materials, easy to curate, and easy to provide standard appearance, reliability, and quality. The benefit of centralization is particularly notable as we consider use of the repository by ordinary citizens who need simple access in order to find the publications that contain the information they need, and hence a repository that does not require the mediation of a librarian to use it effectively. At a minimum, the end user needs to see a single web site as the point of initial access and consistent procedures for actually viewing works of interest. Although some of the benefits of centralization can be achieved by federated search, such search is much easier to implement if multiple repositories all share the same management. In any case, federated search is a finding aid rather than an access tool, and does not solve the potential problem of multiple user interfaces and policies that can easily frustrate access even when the user has a link to the article desired.

It should be noted that centralized management and policy setting does not preclude outsourcing or cloud-sourcing if such outsourcing makes sense financially and suitable contractual arrangements can be established. For example, the federal government might contract for storage and compute resources from a commercial provider such as Amazon or Google or a consortium of publishers. However, any such outsourcing must clearly establish a set of stringent requirements for public access and a standardized interface; by far the easiest way to accomplish this would be to outsource only the back end with the expectation that the applications being run and the ingest policies were those currently in use for Pub Med Central. Even for outsourced data storage contracts must be clear that the corpus of copies is the property of the federal government, that the contractor is an agent of the federal government and is required to meet any legislatively mandated regulations that apply to the agency, and that adequate availability, security, and termination guarantees are in place. Under the last issue, for example, there must be a clear migration path for recovery of the data in the event that a provider is no longer able to meet its obligations. Outsourcing to multiple providers (for example, to individual publishers) would likely incur dramatically increased costs in providing standardized guaranteed access. Overall, our expectation is that it is unlikely that such outsourcing would prove cost effective.

However, it is also vitally important that other players be able to mirror data and that copies of publications be available across multiple public and private sources. As research in library archival has demonstrated, one of the most effective strategies for long-term preservation and access is based on principles like LOCKSS ("lots of copies keeps stuff safe") (Reich, 2008) that distribute risk across multiple servers, organizational entities, and archival approaches [it is notable, however, that this is not an alternative to a federally maintained repository; current market attempts to implement LOCKSS itself have not yet been fully successful (LOCKSS Assessment Team, 2011)]. In addition, we anticipate collections that meet specialized needs both

for access (e.g. very high bandwidth access) or content (e.g. collections that include both publications and associated data sets or derivative works). Just as with populations of biological individuals, some genetic variation within the population of archives makes it more robust in the face of environmental stresses and new demands.

For example, we do not anticipate that federal agencies should maintain custody of research products that are not federally funded, but in many cases such research will be released for open access and will be part of the ecosystem that is viewed (and manipulated as a whole) by future researchers and the public. The situation is analogous for a university institutional repository, which may have a mandate to collect and make publicly accessible all peer reviewed publications generated by its own faculty (Brody, 2011) but not related research; in other cases however, such an entity would want to collect all or a subset of the articles in a federally maintained archive (for example, all articles related to a particular research area that the university is investing in, or all articles produced with funding from a federal agency that political scientists at the university are studying). Similarly, we anticipate that publishers will usually wish to maintain the copy of record for publications that have appeared in their journals, and that it will be important that other archives (some of which may store preprints, derivative works, or copies that contain formatting changes relevant to future scholars) provide clear links to the copy of record. In cases where publishers have adopted the DOI standard, this notion of a copy of record is easily mapped to the particular version that the DOI resolves to, and is analogous to the "best edition" as used for copyright deposit purposes and defined in 17 USC 101 and detailed rulemakings (Peters & Billington, 2010, pp. 3868-3869).

Mirroring requires appropriate access licenses to the texts in source repositories, clearly specified conditions for public accessibility and long term preservation, and technical solutions such as the OAI-PMH (Lagoze & Van de Sompel, 2011) that allow bulk harvesting of content. In addition to providing a centralized repository for federally funded articles, federal agencies need to consider standards for archive interoperability and need to invest in research and development of tools for effective mirroring and archive description. One important modification that needs to be made to present policies is relaxation of the gratuitous limit in PubMed Central on bulk copying. Such copying may be restricted by licenses to specific articles, but should be freely permitted is the article's usage license permits it.

References

Brody, T. (2011). ROARMAP: Registry of Open Access Repositories Mandatory Archiving Policies Retrieved 7 Dec 2011, from http://roarmap.eprints.org/

Grant, B. (2009, 30 April 2009). Merck published fake journal. *The Scientist: Magazine of the life sciences*.

Lagoze, C., & Van de Sompel, H. (2011). Open Archives Initiative Retrieved 7 Dec 2011, from http://www.openarchives.org/

LOCKSS Assessment Team. (2011, October, 2011). Final Report of the 2CUL LOCKSS Assessment Team, Cornell University Library & Columbia University Library Retrieved 7 Dec 2011, from http://2cul.org/sites/default/files/2CULLOCKSSFinalReport.pdf

Peters, M., & Billington, J. H. (2010, January 25, 2010). Mandatory Deposit of Published Electronic Works Available Only Online [revises 37 CFR 202]. *Federal Register, 75 (15),* 3863-3870.

Reich, V. (2008). LOCKSS Retrieved 7 Dec 2011, from http://www.lockss.org/lockss/Home

Van Camp, J. (2011, November 22, 2011). Penguin halts Kindle library lending: Will more publishers disable the feature? *Digital Trends*.

Comment 4

(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?

One class of partnerships is with research universities, particularly university libraries. Such libraries have extensive experience in preservation and archive infrastructure, and tend to have very long time horizons; the oldest private university libraries in the U.S. predate the federal government by more than a century. Some of the best non-Federal examples of successful archives, e.g. ArXiv. represent public-private (specifically Federal-university) partnerships.

Universities are also increasingly becoming active publishers of open access journals, and have a role to play in that regard. For example, the University of Oregon is fairly typical of major university libraries in having established a small program to publish very specialized open access journals, with a current focus mostly on humanities and social sciences. Such journals fill an important niche that is currently largely ignored by traditional commercial publishers. They, along with commercial open access journals such as those from BioMed Central, provide a model that avoids many of the intellectual property issues created by traditional publishers, since the license that the journal requires does not conflict with federal requirements for deposit in an archive such as PubMed Central. However, partnership is needed to develop better tools for automating such deposit along the lines of NIH submission "method B" (National Institutes of Health, 2011) that are specific to the (often open source, e.g. OJS, Drupal, Annotum, etc.) publishing software that university publishers typically use.

Numerous other examples of successful public/private partnerships also provide models. For example, consider the NSF International Children's Digital Library, one of many resources provided by federal agency / private collaborations as part of the FREE website (Federal Resources for Educational Excellence, 2011). In many cases such resources can be greatly enriched by references to the peer reviewed literature and examples of the science that informs more popular presentations of information.

References

Federal Resources for Educational Excellence. (2011). International Children's Digital Library Retrieved 18 Dec 2011, from http://free.ed.gov/resource.cfm?resource_id=2187

National Institutes of Health. (2011). Submission Methods Retrieved 10 Dec 2011, from http://publicaccess.nih.gov/submit_process.htm

Comment 5

(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?

In this comment we primarily address the second and third questions posed. We believe that rich and well documented metadata, combined with implementation of current federated search protocols and with research on such protocols, will contribute to interoperability.

Many of the most important core metadata fields are already captured in standards such as the NLM-XML archiving tag set (NCBI, 2010), though are not always mandatory. Examples include standard citation information such as author names, article title, journal name, and so on. The archiving tag set provides a good tool for describing and ingesting arbitrarily structured journal articles, but fails to distinguish well between descriptive information contained within the article and metadata associated with it, and does not mandate a set of minimum core metadata. We recommend that the PubMed Journal Article DTD (NCBI, 2011) be adopted and extended as the target for metadata associated with new journal articles.

Although the archiving tag set is a good starting point that balances consistency with pragmatics, neither it nor the Journal Article DTD prescribes a sufficient minimal set of metadata elements to guarantee that a user can create a reference, but being able to create a reference to an article seems to be a minimal bar that defines metadata that most users are likely to need. Since citations are not always in NLM format, the required metadata needs to be rich enough to allow generation of minimal NLM, APA, Chicago, and MLA, references, suggesting that any datum that is common to and required in at least 3 of these should be considered mandatory metadata.

In addition, several article-level metadata fields are not widely standardized or encoded but are very important for effective use and should be included in a minimal mandatory set.

One specific metadata element that is needed for individual peer-reviewed publications is a precise statement of the license that a document is released under and who it applies to in a format that can be automatically processed by text mining software. The archiving tag set <permissions> entity is a good start in this direction and more flexible and detailed than the Journal Article <copyright> entity, but is oriented towards documenting restrictions and so for instance would not be used to affirmatively assert that a work was in the public domain because the author was a federal employee.

Another vital field is versioning information that describes the relationship of a particular document to the published version. At a minimum a simple controlled vocabulary that specifies whether a version is an author's preprint, the author's final

manuscript version, a copy of the publisher's copy of record (and if so, as of what date and in what ways it has been modified, since even formatting variations may prove relevant to future researchers), a derivative work, etc. We assume that in most cases publishers will wish to maintain the copy of record for publications that have appeared in their journals, and will associate with that copy a revision history (for example, tracking status if by mischance the article needs to be withdrawn). It is important that other archives provide clear links to the copy of record, for example by providing a DOI if available. Some but not all of this information can be encoded in the Journal Article <pub.status> and <eLocation> entities.

Another article-level metadata element that is needed is information as to what data sets, samples, case studies, and supplementary documents the conclusions rely on. Knowledge of the data sets used is particularly important in minimizing duplicate publication and in facilitating meta-analyses. The APA Publication Manual in Psychology describes numerous reasons to eschew duplicate publication, and states that: "As multiple reports from large-scale or longitudinal studies are created, authors are obligated to cite prior reports on the project to help the reader understand the work accurately... It is also important to make clear the degree of sample overlap in multiple reports from large studies" (American Psychological Association, 2010, pp. 13, 15). Reporting in systematic fashion the particular grant(s) that funded the research is a step in the right direction, but since many grants generate multiple studies and data sets the particular data set used needs to be reported in a fashion that allows machine analysis.

In addition to enhancing the core metadata fields, federal agencies need to improve their archives by clearly distinguishing metadata (what NLM-XML generally considers front matter) from the publication itself, making metadata available through both the repository web interface and through an API interface. One step that would make it more likely that metadata is consistently provided would be to move towards consistent upload of articles in NLM-XML or as an OAI-ORE. We suggest recommending submission of the "best format" (Peters & Billington, 2010, pp. 3869-3870) as now required for Library of Congress deposit of online-only journals. In addition, we recommend that all submissions be required to include a minimal set of metadata, either mechanically derivable from the article itself (as for instance, with NLM-XML), or separately provided by the submitter, and that the ingest software first attempt to fill out all fields based on the article itself, then ask the submitter for additional information as needed.(NCBI, 2010)

References

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, D.C.: American Psychological Association.

NCBI. (2010, June 6, 2010). Archiving and Interchange Tag Set Retrieved 14 Dec 2011, from http://dtd.nlm.nih.gov/archiving/

NCBI. (2011, 2011-10-11). PubMed Journal Article DTD Version 2.6 Retrieved 14 Dec 2011, from http://www.ncbi.nlm.nih.gov/entrez/query/static/PubMed.dtd

Peters, M., & Billington, J. H. (2010, January 25, 2010). Mandatory Deposit of Published Electronic Works Available Only Online [revises 37 CFR 202]. *Federal Register, 75 (15),* 3863-3870.

Comment 6

(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?

Five strategies seem particularly appropriate for maximizing benefit while minimizing costs:

1. Standardize licenses to include the maximum possible reuse rights (such as those embodied in Creative Commons CC-BY licenses) for the public;

2. Ensure consistency (and to the extent possible, temporal stability) in requirements and processes, particularly across multiple agencies;

3. Take advantage of current software technologies to minimize cost, both to agencies and to depositors;

4. Provide simplified interfaces to make mandate compliance as easy as possible.

5. Encourage commercialization, but only in domains where competition exists to avoid monopoly profits; in all cases of commercialization weigh the benefits of standardization and the overall cost efficiencies of providing services centrally or of contracting for services with non-profit entitites such as universities.

Under "standardization," please see this submission's Comment 2, noting that standardization of license terms also contributes to consistency and simplicity. Under the "consistency" umbrella, it is particularly desirable that all federal agencies to which a mandate applies use a standardized set of criteria to determine what materials need to be deposited and how deposit should occur. This is obviously relevant to researchers who may have funding from multiple federal agencies, but is also important because it reduces support costs, e.g. for documentation and for training of support staff (both at federal agencies and within institutions that assist their researchers with compliance). For example, at the University of Oregon our total post-award sponsored project administration staff comprises 8 FTE despite a moderately large number of extramural projects. Minimizing the complexity involved in monitoring regulatory mandates and advising/assisting PIs on compliance is critical to ensuring not only that cost of administration is contained and compliance reviewed quickly, but that mandates are actually followed.

Under "current technology," one observation is that agencies have an interest in encouraging the development of new tools to make compliance simpler. One specific example that relates to consistency is that authors may need to deposit in multiple repositories based on multiple simultaneous mandates. Different components of a University Of Oregon research project might be funded by Wellcome Trust and NIH, plus some departments at the University of Oregon have departmental deposit mandates into our institutional repository. Why should the author have to go through 3 processes to meet three sets of deposit requirements? It should be

possible using protocols such as SWORD for the author to automatically deposit an author's final draft in multiple repositories simultaneously. It should be possible as part of the deposit process to provide automated notification to interested parties such as an institutional grants management system or institutional faculty profile databases. It should be very easy to extract in a standardized output-neutral format such as Citation Style Language (CSL) all citation data for a single researcher or an entire institution. The goal should be an easy process for moving citation data to free and commercial bibliography management and publishing tools such as bibtex, endnote, or mendeley. Standardization on a powerful and open citation format would encourage standardization within the bibliographic software industry.
Under "simplicity" one very useful approach would be for an agency to employ a human interface design consultant to conduct tests and identify issues that make it difficult for real users to comply with the mandates.

Comment 7

(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?

Although it is desirable for all peer reviewed publications resulting from federally funded research to be widely available, the policies under which they are made available may need to differ depending on the type of material. For example, the interests of authors in peer reviewed journal articles where they generally receive no financial remuneration may differ from those of authors of commercial textbooks.

Another plausible next step in expanding types of publications would be to peer-reviewed conference presentations. One issue here is that such presentations show a much wider range of formats than peer reviewed journal articles, and so present more technical challenges. If a conference presentation is a multimedia presentation using Mathematica (rather than a text with an incidental use of a demonstration), must deposit consist of both the Mathematica workbook and a purchased copy of the version of Mathematica required to display it? As conferences increasingly become hands-on the problems of adequately capturing the "presentation" get even greater; what about that presentation that allows audience members to use a Microsoft Kinnect to explore a virtual human?

Given the complexities involved, we at the University of Oregon believe that efforts at this time to extend public access policies to include book chapters or conference proceedings within a single public access policy would increase complexity and make it harder to achieve the more important goal of widespread access to peer reviewed journal articles.

However, one class of materials that are particularly important to make available are research instruments, survey forms, and protocols. Journals are currently quite inconsistent in their requirements for publication of such supplementary materials and tend to be driven by no-longer-relevant concerns about costs associated with hardcopy distribution, but in a number of disciplines the relevant professional organization makes clear that it mandates public access to such materials. For example, in Psychology the APA Ethics Code requires that researcher retain and make available to other researchers not just data but also supplementary materials including "[o]ther information related to the research (e.g., instructions, treatment manuals, details of procedures, code for mathematical models reported in journal articles) …; such information is necessary if others are to attempt replication." (American Psychological Association, 2010, p. 12). Given the central role of replication in the progress of science, mechanisms should be established to routinely and systematically collect such materials and make them available online as supplements to the text of the peer reviewed article.

Although not peer reviewed publications per se nor adequate substitutes for such, research progress reports and final reports are potentially very useful in contextualizing grant-funded scholarly articles. Such reports are not a substitute for the peer reviewed articles that document results, and indeed one notes that it is routine for reports to reference the publications. However, the grant reports do provide useful auxiliary information and in particular may be helpful in determining

relationships between multiple peer reviewed publications flowing from a single grant. Federal agencies should take steps to ensure not only that such reports are freely available to the public in timely fashion, but that tools are available to cross-walk between those reports and corresponding articles.

References

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, D.C.: American Psychological Association [not available open access].

Comment 8

(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?

Embargoes of any length impose a cost in terms of decreased public access and a negative impact on the degree to which an article's availability fosters further research and development. Houghton et al estimate (Houghton, Rasmussen, & Sheehan, 2010, p. 8) that "a six-month embargo reduces the returns [in benefits from increased R&D] by around $120 million (NPV)." Similarly, increasing evidence (Wagner, 2010) indicates that open access in general increases citation rates for peer reviewed publications, suggesting that embargoes may have a negative effect on readership of the embargoed journal. For example, data from a sample of Chinese journals (Cheng & Ren, 2008) indicates that journals in their sample that had embargoes (delayed vs immediate open access) experienced a 20% citation disadvantage (1-1.26/1.57).

The argument in support of embargoes is of course that such embargoes putatively encourage academic libraries to continue to subscribe to the journals. We know of no studies that directly examine this hypothesis or of documented examples of journals whose financial viability has been significantly damaged by public access policies such as the NIH public access mandate. It would be very useful to be able to consider empirical data, preferably in peer-reviewed economics journals, that facilitated measurement of this possible relationship and the more general question of the economic effect of publication embargoes.

Consensus in the library community seems to be that cancellation decisions are budget related, not access related. Anecdotal evidence from our University of Oregon journal cancelation projects does not show that embargoes themselves (including open access moving walls and embargoes on availability of full text within licensed databases) have had any substantial influence discouraging journal cancellation decisions, particularly given the constraints of publisher bundling of multiple journals in a single subscription. One UO library subject specialist and department head, however, goes further and reports that "I do take embargoes into account. I tend to view an embargo such as a 1 year embargo on access to full text in a licensed database negatively" (Frantz, 2011).

It is important to note as well that the benefit of embargoes accrues only to those publishers who use particular economic models. The rapid growth of the open access journal market, where subscription fees are replaced by author fees (or possibly by APCs supplemented by institutional subsidies, advertising, and other revenue streams) have very different needs and no clear benefit from embargoes. It may be in publishers' interests to move to APC funding models, which would

completely vitiate any need for embargoes. On the other hand, if journals move to advertising-driven open access models, then their publishers may continue to have an interest in embargoes.

In examining the length of embargo periods, it is important to note that a maximum embargo period of six months is becoming the norm among biomedical research funders, with NIH an outlier at allowing 12 months. (Carlson, 2011).

The prompt specifically addresses the question of whether delay should be different in different disciplines. Although different disciplines and even narrow but nearby subdisciplines clearly show different patterns of article usage – there is a much stronger premium on early access to preprints and published articles in rapidly changing subdisciplines and in those STEM fields where research results often lead directly to commercialization – it is not clear whether such differences ought to be considered even if they imply differential economic effects. For one thing, the effects on subscription-based publishers should be balanced by the economic benefits of widespread access to the economy as a whole. For another, there are substantial costs in increased administration complexity and user confusion as soon as one allows differential embargoes.

Our overall impression at the University of Oregon is that the PubMed Central model, with variable embargoes from 0 to 1 year, is working adequately, and does not need to be changed at this time. However, we also believe that there is reasonably strong evidence that a standard maximum embargo period of 6 months would be preferable. It is also important that embargo periods be established and approved by the individual author within those parameters, since it is the author who is granting to the federal government the rights allowing public access, and the author who best understands the negative impact of an embargo on rapid readership for his or her work.

References

Carlson, D. (2011). *Open Letter to Dr. Francis Collins, Director, National Institutes of Health*. SPARC. Washington, DC. Retrieved from http://www.arl.org/sparc/bm~doc/nih-sparc-final-11-0406.pdf

Cheng, W. H., & Ren, S. L. (2008). Evolution of open access publishing in Chinese scientific journals. *Learned Publishing, 21*(2), 140-152. doi: 10.1087/095315108x288884 [not available open access]

Frantz, P. (2011, 12 December 2011). [personal communication].

Houghton, J., Rasmussen, B., & Sheehan, P. (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

Wagner, A. B. (2010). Open access citation advantage: An annotated bibliography. *Issues in Science and Technology Librarianship*(60).