James Ryley, PhD, RPA

December 1, 2011

Mr. Ted Wackler, Deputy Chief of Staff
Office of Science and Technology Policy
Executive Office of the President
725 17th Street Room 5228
Washington, DC 20502

Re: RFI: Public Access to Peer-Reviewed Scholarly Publications Resulting From Federally Funded Research

Dear Mr. Wackler,

The following comments are made in my capacity as a scientist, the President and Founder of SumoBrain Solutions, and a Registered Patent Agent. It will help put my comments in context to explain my background, what our company does, and why.

I am a biologist by training. As such, I am quite familiar with biomedical research and NIH's PubMed data, which will be discussed later. I am also a Registered Patent Agent with the US Patent & Trademark Office. And, in addition to being a Registered Patent Agent, I have been immersed in field of Intellectual Property for the past 7 years at SumoBrain Solutions. SumoBrain Solutions provides software, data and data analysis, and consulting in the Intellectual Property and Business Intelligence arenas.

SumoBrain Solutions also creates and maintains free web sites aimed at serving various research communities. For example, www.FreePatentsOnline.com (FPO) provides a free patent search engine, and www.BioMedSearch.com (BMS) provides a free biomedical search engine.

Both web sites mentioned above rely upon public access to large government document collections. In the case of FPO, it is patents and patent applications from many patent offices around the world. In the case of BMS, it is NIH's PubMed database.
Our mission is to add value to important sets of documents like these, in many ways, including the creation of federated databases, advanced search tools, data analytics, and the creation of user communities.

FPO was our first, and continues to be our most popular, web site. FPO has existed for over 7 years, and is arguably the most popular patent site in the world. We serve over 10 million page views per month. Our audience consists of attorneys, patent searchers, academics, librarians, independent inventors and others.

By improving upon the access to these documents offered at government sites, we have been able to provide a great deal of value to the patent community. We have over 1 million

registered members, and I think the fact that so many people choose to use our web site instead of the free sites offered by the USPTO, EPO, WIPO, and others is a testament to the amount of value.

And, though many large companies use our site, I like to think that it is particularly helpful to small businesses that may otherwise be unable to afford expensive subscription-based access to patent databases (which frequently cost between $150 and over $1000/month per person).

BMS also serves a substantial number of users each month. However, it has not attained nearly the community acceptance in the biomedical field that FPO has attained in the intellectual property field. There is one main reason for the different levels of community acceptance of these two sites: Patent documents are truly public access, while biomedical documents are not.

Patents have no embargo period, no restrictions on reuse and every document (going back to the founding of the US Patent Office in 1790) is available. The result of this unfettered access is that we have been able to add a tremendous amount of value to the patent documents without any of the drawbacks that would occur were the documents subject to any kind of restrictions (e.g., missing documents, lack of full text, or lack of ability to creative derivative works).

In contrast, NIH's PubMed database, while also operating under a form of public access, is not freely available in a practical sense. With substantial embargo periods, reuse restrictions, and a great deal of historic data that has not been grandfathered into public access, comprehensive data is not available to the public. And comprehensive data is what is needed; no one wants to search a database where they might be missing crucial documents, no matter how good its other features may be.

Consequently, BMS has not been able to, and never will, add the amount of value to its community that FPO adds to the intellectual property community.

The contrast between FPO and BMS provides an important case study in the effect of different levels of public access: restrictions on public access, even "minor" restrictions that may seem like reasonable compromises, can completely negate innovation and competition in providing access to the documents that are the product of federal research grants.
The reason is simple: Users will always prefer a database with 100% of the documents over a database with less than 100%, even when the incomplete database might be better in every other way. Said another way, if a government or publisher web site, due to restrictions on public access, has documents that are not available to third parties, you will effectively have public access at only that government or publisher web site. This is not conjecture; as explained above, we have done it both ways.

My answers to specific RFI questions are below. Please note that my ideas on this topic are hardly revolutionary. So, I will frequently not go into great detail and instead provide references to those who have already eloquently expressed these concepts. I do not believe that the value of my input is in originating new solutions to the problems involved in implementing public

access. I think the solutions are already known. The problem is the number of possible solutions and the competing interests of various stakeholders.
Therefore, I see the task at hand as choosing among the many possible solutions already proposed in order to establish a public access policy that provides the most benefit to the stakeholders, and the nation as a whole. I hope that my input as someone familiar with public access from a variety of different viewpoints will help with that goal.

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?**

   Provide true, unfettered public access. With the appropriate data available, access and analysis will occur through competition. Our own company and the many companies which develop search, data mining, and data visualization tools are evidence of this.

   The appropriate data and access to the data means full text and meta-data, with the ability to reuse in its native form, create analytics based on the data, and create derivative works.

   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?

   Public access can only help science and the economy if it truly adds value beyond the methods of accessing these documents that currently exist. That means that the private sector must be willing to provide search and analysis tools which improve access to information and productivity across many fields of science. Encouraging the participation of the private sector means lowering barriers to entry, both technical and financial, and allowing for viable commercial products.

   While I will go into more technical detail elsewhere, lowering technical and financial barriers means that public access policy should provide for free bulk access to normalized documents and meta-data, preferably in a single repository [1], without an embargo period. The viability of commercial offerings based on this data, on the other hand, hinges largely upon ensuring that there are no reuse restrictions or embargo periods and 100% participation. A commercial search and analysis solution must rest upon complete data to be viable.

   What are the relative costs and benefits of such policies?

   I see very little cost and huge benefits. The costs to set up the actual document access are vanishing small compared to the cost of the research represented by the documents and data in question. For example, Houghton estimated that the cost of a national system of repositories in Australia would average $10 million per year over 20 years.[2] While certainly the cost for US institutions would exceed that number due to differences in research spending and literature output, the point is that the cost of an open access system would be measured in millions per year, versus billions per year in research spending.

   And, while a complete summary of the benefits of public access to research and the economy goes beyond the scope of this reply, I believe it is fair to say that the benefits are substantial and already well-documented.

For example, the ability to discover and access academic research is paramount to many research-related businesses, the cost savings from Open Access can be substantial, and the burden of locating the full text of papers that are not Open Access is estimated at 60 minutes per problem-publication, creating substantial lost productivity.[3]

Further, a benefit to cost ratio of 37:1 has been estimated for enhanced access to higher education research, a ratio far in excess of the ROI on research itself.[2] Given this information, it would seem that increasing the efficiency of access to research documents may be the single most cost-effective measure that can be taken to bolster research ROI.

The sole drawback I see is that publishers will be negatively impacted. However, I believe that this cost is small when compared to the potential benefits to science and the economy, and that alternative business models, such as author publication fees, can be used by publishers to offset the negative impact of robust public access.

Note that, while my understanding is that NIH's public access policy has not has a negative effect on publishers, for reasons already stated the NIH model of public access is insufficient to realize that vast majority of the benefits that could be reaped from true public access. Therefore, I assume that under future policies allowing for true open access, publishers will need to adopt new revenue models.

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

From the point of view of the private sector interested in adding value to the data, the requirements are free, instant, bulk access to structured data and meta-data, preferably from a normalized, federated database, with no restrictions on reuse. This creates the lowest barrier to the private sector, resulting in more competition to promote access and analysis.[1]

Our experience with patents makes it very clear that public access lacking any one of these attributes greatly increases the barriers to entry. Some of these drawbacks can be overcome. For example, multi-point access instead of single-point access drives up technical costs, but could be dealt with. However, an embargo period or restrictions on reuse rights cannot be "programmed away." If such restrictions exist, third-party entities will not be able to provide competitive solutions and so will shy away from entering the field.

With respect to the access that end-users of the data require, it would be the availability of fast, federated databases with state-of-the-art search and analytics tools (some of which do not currently exist, and which may be area-specific) which allow researchers to save time by quickly sifting through millions of documents to find those truly relevant to their needs. It is widely acknowledged that time savings (as opposed to monetary savings via the elimination of subscription fees) is one of the most important benefits of public access.[1, 2, 4]

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?**

These groups do not have aligned interests, so this question requires two answers.

With respect to intellectual property, let me address patenting first. I have heard concerns voiced that public access somehow compromises the patenting rights of researchers. This is absolutely untrue. The patent law is clear on the fact that once information has been made public, anywhere, it serves as prior art. There is no difference between a journal article appearing on a publisher's web site, in print, or in a public access database. Once the author decides to publish, public access is irrelevant.

Open Access is, however, very relevant to copyright. Full public access requires authors, institutions and publishers to give up some copyrights. However, most of the stakeholders in this equation are not adversely impacted by this.

For example, it would be rare that a researcher would want to limit the distribution of his manuscript, or would attempt to profit from it by selling copies (if we are talking about peer-reviewed articles, as opposed to, e.g., book chapters). On the other hand, the researcher certainly benefits from wide distribution of his work in terms of reputation, citations (and public access articles are cited 45-100% more than non-public articles [5]), collaborations, and possible licensing opportunities. The same comments could be applied to Universities. And, of course the public benefits from increased efficiency which impacts them more in terms of advances in science that improve the economy and quality of living, rather than directly through the availability of documents. On balance, the intellectual property picture for most stakeholders is extremely positive.

The unaligned interest is, of course, the publishers. The publishers wish to retain copyright to documents to which they have added value via the peer-review and editing process. And in the past, limitations on public access were accepted to protect the publishers' interests. For example, embargoes and a lack of reuse rights are key drawbacks in NIH's current public access policy. While effective in protecting the publishers' interests, these limitations (as I described in detail above) severely reduce the benefits of public access to all other stakeholders.

I believe that there is a true conflict here which cannot be avoided with compromises and clever legislation. If the full benefits of public access are to be realized, it must be accepted that such policies will indeed force change upon the publishing industry by requiring new revenue models.

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?

I believe the response to the above question addresses this. Immediate, complete public access does undermine the interests of the publishers (though perhaps not technically if one were to assume than an Author's Copy of each manuscript was made available via public access instead of the Publisher's Copy). However, immediate, complete public access is by far more advantageous than any compromise position to all other stakeholders.

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?

We aggregate data from the US Patent & Trademark Office (USPTO), the European Patent Office (EPO), the World Intellectual Property Organization (WIPO), the German Patent Office (DPMA) and others. After years of dealing with data coming from multiple agencies, and in multiple formats even when from the same agency, I can say with certainty that a centralized approach to public access would be preferable.

If the data is decentralized, particularly if each agency is allowed to create its own data standards, the task of aggregating and normalizing the data (and simply dealing with each individual agency when problems arise) can be substantial. This increases the barriers to entry for private entities wishing to provide access to, and analysis of, the data.

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?

I see no reason to allocate stewardship to private sources. The creation of value-added tools and services surrounding the content is an appropriate task for private entities, as it is in their interest to do that to the best of their abilities. Long-term stewardship however, is not. Nor is stewardship a burdensome task for the agencies. Storage space is inexpensive, as is distribution. The costs associated with creating and maintaining such a system would be trivial in Federal research budget terms, and any savings realized by allocating this role to the private sector would not be worth even a slight possibility of data loss or inaccessibility.

Note that several agencies, including USPTO, NIH and SEC already provide such stewardship of their own public data (although the USPTO shares such stewardship with Google, as discussed in Question 4). Our experiences with access to these repositories have been more than satisfactory. So, existing models would seem to support stewardship by the government as well.

I should clarify though that data access and data quality are two very different things. Our experience with access has been good. Our experience with data quality from various government entities has not been nearly as good, which is why I strongly suggest a centralized repository with stringent requirements for data, meta-data, normalization, OCR, a strong preference for “born digital” data, and other technical issues.

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?**

With respect to ensuring stewardship, this seems rare, though one could look to the arrangement between Google and USPTO for one example. Google now provides bulk data access to some USPTO data. However, this situation is unusual in that Google seemingly did this “because they could,” rather than because there was a viable stand-alone business model.

Without a viable business model for private sector stewardship, this would generally seem problematic. For example, while SumoBrain Solutions could provide such stewardship (after all, we have to maintain the data internally anyway), we could not afford to do so for free. Perhaps there are efficiencies to be realized by assigning stewardship to companies like ours. However, I do not believe the cost savings would justify the possible drawbacks.

With respect to public-private models for encouraging accessibility and interoperability, this is not as rare and I would cite our company, SumoBrain Solutions, as a perfect example of this type of public-private partnership. We have over 50 million patent records from patent offices across the world (including the USPTO), 20 million biomedical records from NIH (PubMed), millions of trademark records from USPTO, millions of SEC records (EDGAR), and millions of documents from some of the largest scientific, technical and medical publishers.

We add value to these data sets in many ways, including the creation of federated searches, advanced search tools (including analytics for business intelligence), standardization of documents into downloadable PDF form, alerts and account features to help organize, annotate and share documents. In this way, we are doing exactly what public access advocates tout as the advantage of getting the private sector involved in public access. And, since there are viable business models surrounding the creation of new databases, search tools, and analytics tools, providing such services does not need to be legislated or funded. Assuming proper access to the data, market forces will ensure that such partnerships occur. They already occur where possible.

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?

I suggest a single archive with standardized data and meta-data. Multiple databases increase the technical barriers to entry, and the notion that there are clearly separable disciplines is problematic in itself.

For example, is a new type of titanium alloy with the promise to improve hip implants to be categorized as materials science or biology? Is a new compound with potential applications to cancer a chemical or medical topic? Worse yet, did the research I seek come out of NIH, NSF, or DARPA? Researchers should not have to worry about such questions.

There is no reason to insist on a one-to-one relationship between documents, areas of research and agencies. Yet, if multiple repositories exist such distinctions are implicitly being made.

In the end, the private sector will federate the databases if need be, since this adds value and therefore is a competitive advantage. But, the requirement that the private sector do this is another barrier to entry.

Regardless of the number of databases, standard formats and fields, both for data and meta-data must be agreed upon to facilitate cross-collection search and analysis.

To provide a simplistic, yet actual, example, consider that most documents have authors, while patents have inventors. Since that example is so simple, the solution is obvious: Consider inventors and authors to be synonymous in a federated search engine.

However, not all examples are so simple, and sometimes the meta-data necessary to make various distinctions does not exist. Consider a Word document which has headings that read “Title,” and “Abstract.” The meaning is very obvious to a human. It is not nearly so obvious to a machine. If the meta-data isn’t present to specify what text constitutes the title and what text constitutes the abstract, then we are reduced to language parsing, which cannot be done with 100% accuracy. And that is a simple case; in more complex cases, language parsing may not be feasible and certainly will not have a high degree of accuracy.

What are the minimum core metadata for scholarly publications that must be made available to the public to allow such capabilities?

This cannot be answered without taking the type of document into consideration. But, assuming a schema that caters predominantly to a typical peer-reviewed journal article, I would suggest the following fields.

Please note that I do not suggest these be the sole fields, or that they be in a flat structure. Rather these fields would probably be best incorporated into an XML structure that allows normalization and one-to-many and many-to-many relationships where appropriate. The table below cannot convey such structure, but given the fields and their relationships, creating an efficient XML version is fairly straight-forward.

I would also note that, due to the technical nature of this question and the extensive work that has already been done in the field (e.g., see Dublin Core and OAI-PMH), the actual implementation details go far beyond what can be provided in this response.

Field	Comments
Abstract	
Author Affiliation	
Author Contact Information	
Author IDs	Unique author IDs to disambiguate authors of the same name
Authors	Separable, with a consistent order of representation (e.g., last name, first name)
Categorization	The actual data for, e.g., MeSH categorization
Categorization scheme	E.g., MeSH headings for biomedical documents
Chemical Formula	Standardized representations of chemicals are very valuable for machine

	search. Patent agencies attempts this now, but current policies have created intractable problems in unambiguous interpretation of chemical structures. The technology exists; this is purely a legislative issue.
Date of Publication	
Date of Submission	
Document ID	
Federal Agency	
Format	
Keywords	
Language	In some cases, when not English, this can be hard to determine by machine
Publisher	
Rights	Specifies, in machine readable terms, reuse rights
Source	Generally a journal, with title, volume, and pages, but other sources, such as a conference proceeding, would have slightly different fields
Table and Figure designations	Allowing machine identification of tables and figures can be important to more sophisticated analytics
Title	
Type of Publication	E.g., conference proceeding, journal article, etc.

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?

No response.

Questions 6-7: No response

8) What is the appropriate embargo period after publication before the public is granted free access to the full content of peer-reviewed scholarly publications resulting from federally funded research? Please describe the empirical basis for the recommended embargo period. Analyses that weigh public and private benefits and account for external market factors, such as competition, price changes, library budgets, and other factors, will be particularly useful. Are there evidence-based arguments that can be made that the delay period should be different for specific disciplines or types of publications?

Immediate access is the only viable model if substantial private sector participation is desired. As I described in detail above, embargo periods are one of the most important aspects of public access policy with respect to private sector participation. This is because embargos are not technical issues that pose problematic, but surmountable, barriers to entry. Only legislation can eliminate embargos and so ensure that many companies in the private sector can create commercially-viable solutions.

I understand that this is not an empirical answer. This question is difficult to answer empirically because ideally what would be cited would be studies on the value added to the economy and the effects on all stakeholders under embargos of different lengths (including no embargo). However, since we do not have true public access at the moment, comparative studies cannot be done. And, while we do have open access journals, they co-exist with traditional journals. There are no major research ecosystems where open access has replaced the traditional model, allowing before and after comparisons to be made.

But, despite the lack of ideal data to help answer this question, I think it would be a mistake to think that the outcomes are not fairly obvious, even if not accurately quantifiable. I say this for several reasons.

First, I have spent the past seven year working in one of the few truly public access fields: Patents. At the same time, my company has attempted to add value to other data sets, including NIH's PubMed (operating under what is also ostensibly public access). The difference in dealing with truly restriction-free data sets versus data sets under some flavor of pseudo-public access is immense. Public access restrictions cannot logically do anything but hamper private sector participation.

Second, consider the research budgets and potential ROI involved. The research spending by the agencies in question has been estimated at $60.5 billion annually, while total national spending on R&D is more in the neighborhood of $378 billion. How much do you need to decrease the return on $60.5 billion (or $378 billion), through barriers such as embargos, before compromises that impair the efficiency of public access cease to make sense? Certainly it seems very plausible that the difference between truly effective public access and a lesser version of public access will cost the nation billions of dollars annually.

Third and finally, ignoring the involvement of the private sector and the important role it would play in the development of time-saving tools and analytics, consider the most superficial and obvious savings potentially afforded by public access: Reduced journal subscription costs for our nation's academic and research libraries. Those potential savings will almost certainly not be realized with any embargo period because libraries supporting research must allow access to

the latest publications. Even one month after initial publication is not acceptable in most fields of science; under any scenario with an embargo libraries will be forced to maintain essentially the same subscriptions they maintain now. Over the past two decades journal subscription costs have risen dramatically faster than inflation or library budgets, posing a burden on libraries and hamstringing researcher access to all the material they need to effectively do their jobs.

End Question Responses

In conclusion, I believe that this is an opportunity to bring research document access into line with what the internet and modern information retrieval technology have allowed for quite some time. The current system is an accident of history. To restrict access, in any substantial way, to the documents which are the products of billions of dollars in taxpayer money is not in the best interest of the people who paid for that research in the first place, much less most of the other stakeholders or the nation as a whole.

When thinking about these issues, I ask myself "How would this be done if we were setting up the system anew today?" I think keeping that question in mind helps avoid getting mired in how it has been done (which I believe resulted in a "splitting the baby" approach with NIH, producing legislation that was better than what existed before, but certainly not optimal), and rather focus on how it could be done.

Public access could be done in a manner that is much more effective than the current system, more effective than the current NIH public access policy, and in a manner that aids science and the economy to the highest possible degree by better leveraging the tremendous amount of research that is already being paid for, but not efficiently shared.

Sincerely,

James Ryley, PhD, RPA

References

1. Thakur, N. *Open access as a path to increased scientific productivity*. in *Berlin 9*. 2011. Washington, DC.
2. Houghton, J., *Exploring the Impacts of Enhanced Access to Publicly Funded Research*, in *The Socioeconomic Effects of Public Sector Information on Digital Networks: Toward a Better Understanding of Different Access and Reuse Policies: Workshop Summary*. 2009, Centre for Strategic Economic Studies, Victoria University, Melbourne: Melbourne.
3. Parsons, D., D. Willis, and J. Holland, *Benefits to the Private Sector of Open Access to Higher Education and Scholarly Research*, 2011, HOST Policy Research.
4. Houghton, J., B. Rasmussen, and P. Sheehan, *Economic Implications of Alternative Scholarly Publishing Models: Exploring the costs and benefits*, 2009, Centre for Strategic Economic Studies, Victoria University.
5. Houghton, J. and P. Sheehan, *The Economic Impact of Enhanced Access to Research Findings*, 2006, Centre for Strategic Economic Studies, Victoria University.