'We the People' Should Receive what We Paid for: Public Access Policies for S&T Funding

David A. Wheeler (dwheeler *at* dwheeler *dot* com)

2009-12-17

Here are my responses to the Office of Science and Technology Policy (OSTP) Executive Office of the President request for information (RFI) "Public Access Policies for Science and Technology Funding Agencies Across the Federal Government", which "focuses on approaches that would enhance the public's access to scholarly publications resulting from research conducted by employees of a Federal agency or from research funded by a Federal agency".

The U.S. federal government was established and is funded by "We the People" (sic). **If "We the People" pay for unclassified research, then "We the People" should be receiving the results**. The current system instead gives monopoly rights to other organizations at a fraction of their development cost. Thus, "We the People" are unable to fully receive the results we paid for. This is unconscionable, impedes the progress of science, and impedes job creation from the jobs that *could* have resulted from the research.

NIH's system is better than many previous approaches; its biggest problem is that it does not go far enough:

1. NIH has a one-year embargo on research results, but the rationale for this embargo is inadequate. The people paid for the results; why should they not receive them once they are available? For an organization to receive a special one-year monopoly rights over any government work, they should have paid for a significant proportion of that research effort (say, 10% or more) before it began.
2. NIH primarily focuses on the papers, but modern science is more than papers. Supporting data (field and experimental) and software should also be included in such releases, if unclassified and developed using public funds, and they should be released to the public under very broad licenses. For example, imagine that climate modeling software was created and a paper written based on it; without the climate model software, it is impractical for others to determine if various extensions to the model would invalidate the paper's conclusions. These licenses should at least permit arbitrary use (including commercial use), modification, and re-release of either the original or modified information. One exception: personally-identifying data must be handled specially, to protect privacy. Science must be repeatable, and today that requires the release of data and software.

Given the above, here are my responses to the questions the RFI poses.

1. *How do authors, primary and secondary publishers, libraries, universities, and the federal government contribute to the development and dissemination of peer reviewed papers arising from federal funds now, and how might this change under a public access policy?*

There are many different development/dissemination approaches, but the government should *not* focus on current techniques at all. These existing approaches were based on the fundamental assumptions that final printing of documents is required, expensive, and must be done by a professional publisher using dead trees. Now that the Internet is widely available, these assumptions are false. Instead, most people want to use a search engine (such as Google), find the results in seconds, and download the results immediately. They may then use screens or print them locally, but any case, there is no need for the current obsolete system.

The Internet's impact is as profound as the invention of the printing press, and thus, we should *expect* that processes will radically change. We should instead focus on what the new processes *should* be, and move to them, instead of worrying that these changes will affect someone. Of course they will!

In particular, many traditional publishers and societies will need to undergo major changes or cease to exist. Publishers/societies will need to undergo changes such as *funding* research, instead of gaining monopolies over research at little cost to them. There is a surfeit of publishers and societies, and there is **no public reason to prop up their obsolete business models**. We now have a better way to spread scientific research results; **we do not need to coddle the manufacturers of buggywhips and quill pens**.

2. *What characteristics of a public access policy would best accommodate the needs and interests of authors, primary and secondary publishers, libraries, universities, the federal government, users of scientific literature, and the public?*

A public access policy should focus on the needs and interests of the *public*. It should briefly consider the interests of others, but the public *must* be the *top* priority. "Accommodating" all other parties is completely *inappropriate*. If "We the People" paid for the research, then "We the People" should get it. Period.

Such a policy must be easy to understand and easy to implement. The policy should include a *requirement* that ***all* government-funded unclassified research results must be "*open access*"**, that is, they must be free for all users to acquire, read, and use via the Internet or any successor network. Exceptions may be granted, but only in cases where significant research funding (say, at least 50%) was from non-public funds and where the government has specially agreed to this. If exceptions are too easy, then private interests will work to capture the public's research. Licensing should be simple, in particular, papers be should be public domain or released under either the Creative Commons CC-BY-SA or CC-BY licenses. Research is cumulative—it often depends on previous research—so works should be released in a way that enables remixing and recombination to enable future research, and in a way that eases commercialization.

This policy should cover not just academic papers, but also **all the supporting unclassified data and software whose creation/generation was funded by the public**. Such supporting data and software should be usable for any purpose, modifiable, and re-releasable in either original or modified form.

To help implement this policy, **a *simple* centralized U.S. government web site should be established**, in which the papers and supporting data/software can be deposited and made directly available to the public. This should make results accessible to all, without a "paywall" requiring payment for these results. This repository would be backed up in geographically-separated locations so that U.S. research will no longer be easily lost. It should be maintained as a U.S. government site under the .gov domain. Each addition should be posted using web standards (such as RSS or ATOM). It would be inexpensive yet yield large benefits. I believe it should be centralized, not decentralized, because centralized systems are easier to scale, back up, and ensure that *all* of these results are not lost. It is too easy to lose results from a decentralized system.

For each paper, a simple set of metadata (data about the paper) should be captured. This should include title, authors, release date, abstract, keywords, peer reviewer(s) by individual or group, contact information, stable URLs on the government site, and perhaps a few more fields. This information should be provided to users in a format that is easily read by computer (e.g., a simple XML file with a reference to formatting information; that way both humans and computers can process the same file). There should be at least two stable URLs for each paper: A stable URL for the paper metadata (such as title) in XML format, with human-readable format, and a stable URL for the paper itself. If supporting data is also released, then a stable URL that is a directory for this data should be included. The web repository should be able to automatically determine what other papers in the repository it refers to, and what papers refer to it, and report that information to users. **There is no need for the government to implement a search system**; commercial organizations can perform this service once the data is available to the public.

The policy should require that data be submitted and provided to users using ***open standards*** for formats. The term "open standard" is often misused, so the definition of "free and open standard" as defined by Digistan (http://www.digistan.org/) should be used. Given current technology, research papers should at least be submitted and provided to others in PDF format. Other open standards that should be acceptable include: HTML, JPEG, PNG, and Open Document. Before acceptance, the papers should be tested by being viewed by at least two independent implementations.

The government should withhold part of the research money until researchers submit their materials as required by this policy, and hold them liable for all research funds (with penalties) if they continuously fail to provide what they were paid to provide. Given this, compliance will be high. And it is easy to justify: the government is paying for research, so researchers should only be paid when they provide their results.

I would suggest that the names of peer reviewers (either individually or via some group) be credited as part of the information about a research paper, to give them credit. This would simply have the same meaning as it does now: "We have reviewed it, and did not see any glaring errors". It would not be a guarantee that there were no errors; should someone find errors, that would be grist for another paper.

The NIH policy currently requires a one-year embargo. This is a nod to obsolete processes and impedes research progress. The U.S. is competing in a world that is moving faster, not slower. Such embargos should be by special exception, requiring significant prepayment, and not a normal part of the process.

3. *Who are the users of peer-reviewed publications arising from federal research? How do they access and use these papers now, and how might they if these papers were more accessible? Would others use these papers if they were more accessible, and for what purpose?*

Currently, much peer-reviewed research is only available via large, wealthy companies or large universities. Publishers often charge monopoly rents for works they did not even pay to create.

The *potential* users of peer-reviewed research is anyone interested in or impacted by U.S. research results. Essentially, that is everyone, because U.S. research covers a wide variety of areas. These users should be able to obtain this research information, without fee, from anywhere in the world. All they would need is a web browser; search engines and other sites would help them find the "most relevant" results, and then they could go to the government sites to actually read those results.

4. *How best could Federal agencies enhance public access to the peer-reviewed papers that arise from their research funds? What measures could agencies use to gauge whether there is increased return on federal investment gained by expanded access?*

Public access would be best enhanced by *true* public access. **Make the research results available on a government site, without fee, via a .gov website**. Centralizing the site, with routine back-ups to geographically sites, to other locations, will prevent their loss.

5. *What features does a public access policy need to have to ensure compliance?*

As noted above, to ensure that researchers comply, **do not pay researchers until they provide their materials to the government for public dissemination**. Problem solved.

The website which provides research results must be owned by the government itself (a contractor might *run* it on the government's behalf). The public has paid for this research, so the government must ensure that the results become available to the public for its use.

6. *What version of the paper should be made public under a public access policy (e.g., the author's peer reviewed manuscript or the final published version)? What are the relative advantages and disadvantages to different versions of a scientific paper?*

At the *least* the "final published version" should be made available to the public under a public access policy. It may also be useful to have older versions available as well. But even *asking* this question makes the incorrect assumption that the publishers with obsolete business models are more important than the public who paid for it. That assumption has it backwards: **the *people*, who paid for it, come first**.

7. *At what point in time should peer-reviewed papers be made public via a public access policy relative to the date a publisher releases the final version? Are there empirical data to support an optimal length of time? Should the delay period be the same or vary for levels of access (e.g. final peer reviewed manuscript or final published article, access under fair use versus alternative license), for federal agencies and scientific disciplines?*

Peer-reviewed papers should be made available to the public **immediately** if the work was publicly funded. The typical purpose of a delay is to give some organization temporary monopoly rights over work that they did not significantly fund. This should be rejected as unwarranted. Peer review does *not* now require that

anyone be granted a monopoly over distribution of a paper. Peer reviewers are often unpaid or paid a small fee; they could receive such amounts without granting monopolies to others.

If private companies decide to fund their own research, then they can decide what kind of delay they want to have. But the government is not a private company; it represents the people. Again, "We the People" paid for it, so "We the People" should get it, directly and immediately, through open access.

8. *How should peer-reviewed papers arising from federal investment be made publicly available? In what format should the data be submitted in order to make it easy to search, find, and retrieve and to make it easy for others to link to it? Are there existing digital standards for archiving and interoperability to maximize public benefit? How are these anticipated to change?*

As noted above, policy should require that data be submitted and provided to users using "**open standards**" for formats. The term "open standards" is sometimes misused; I recommend using the definition of "free and open standards" as defined by Digistan (http://www.digistan.org/). Given current technology, research papers should at least be submitted and provided to others in PDF format. Other open standards include HTML, JPEG, PNG, and Open Document.

9. *Access demands not only availability, but also meaningful usability. How can the Federal government make its collections of peer-reviewed papers more useful to the American public? By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections? What are the best examples of usability in the private sector (both domestic and international)? And, what makes them exceptional? Should those who access papers be given the opportunity to comment or provide feedback?*

The most important type of "usability" is legal usability. I recommend that all papers be released as public domain, CC-BY-SA, or CC-BY licenses (the CC-BY-SA license is used by Wikipedia and many other projects). Software should be released under common licenses that allow anyone to use them for any purpose, modify them, and redistribute them unmodified or modified, enabling future research.

"Number of readers" is a misleading metric. If only one person reads a given paper, but in the process cures cancer, who cares that there was only one reader? Instead, I would emphasize these metrics:

- % open access: This would be the percentage of unclassified government-funded research papers that are open access, within 1 week of their release from the researcher(s) and peer review by a small group. This number should be "100%"; anything less means that some paper is being "stolen" from the public who paid for it. Since researchers should not be fully paid until they provide these papers, this value should quickly move to nearly 100%.
- Hours before open access: Once the researchers have released a paper so people (other than reviewers) can read it, how many hours does it take to become available to the general public? This should be measured in seconds to hours, not days or months.

There is no need for the U.S. government to develop a complex search system; commercial organizations like Google do a better job. Simply make the documents available to all, and allow commercial organizations to download them for searches (ensure that there is no robots.txt file, or that its settings permit all to search it). This merely requires that the material be easily processed electronically; open standards and section 508 compliance go a long way toward that goal.

Basic metadata, including contact information, should be provided on each paper in a standard form (I suggest XML plus format information so that human readers would see a "nice" format of it). It would be nice to support some sort of comment system, but the effort to set up a comment system should *not* interfere with releasing research results to the public. I believe that the government should *first* focus on releasing all unclassified government-funded research results as open access; *then* consider comment systems and other methods to improve usability. That way, commenting issues will not impede the more important task: ***We must release research results to the people who paid for it***. ***Starting now.*** Thank you.