Policy Forum on Public Access to Federally Funded Research: Features and Technology

By Rick Weiss

This morning OSTP is launching Phase Two of our forum on public access publishing, which will focus on Features and Technology. (Phase One began on Dec. 10 through Dec. 20, and a wrap-up of that Phase is posted here.)
It is one thing to talk about the philosophy of public access and open government generally, and quite another to get serious about how, exactly, to implement some of those ideas. So through the waning hours of 2009—until midnight of Dec. 31, that is—OSTP is inviting you to weigh in on some of the nuts and bolts aspects of public access publishing. Among the questions we hope you will address:

- In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
- Are there existing digital standards for archiving and interoperability to maximize public benefit?
- How are these anticipated to change?
- Are there formats that would be especially useful to researchers wishing to combine datasets or other published results published from various papers in order to conduct comparative studies or meta-analyses?
- 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?
- What are the anticipated costs of maintaining publicly accessible libraries of available papers, and how might various public access business models affect these maintenance costs?
- By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?

On Jan. 1 we will move to Phase Three of this discussion, which will focus on questions of Management. That discussion was originally scheduled to run through Jan. 7. However, we have heard from many of you that the scheduling of this forum has posed difficulties, especially because of the intervening holidays. So we have decided (and will soon announce in the Federal Register) to add two weeks beyond the scheduled end of this forum. We will use that period from Jan. 7 to Jan. 21 to revisit, on a more detailed level, all three focus areas that will have been addressed by then—perhaps asking you to dive deeper into a few areas that, by then, show themselves as deserving additional attention.

Thanks for your continued involvement in this experiment in open government and public engagement. We look forward to learning from you!

Rick Weiss is Director of Strategic Communications and a Senior Policy Analyst at OSTP

This entry was posted on Monday, December 21st, 2009 at 9:00 am and is filed under News, Public Access Policy, Requests for Comment. You can follow any responses to this entry through the RSS 2.0 feed.

Responses to “Policy Forum on Public Access to Federally Funded Research: Features and Technology”

+1 Stevan Harnad said on December 21, 2009 at 11:43 am:

FORMAT: There is no need at all to be draconian about the format of the deposit. The important thing is that the full, peer-reviewed final draft should be deposited in the fundee’s (OAI-compliant) institutional repository immediately upon acceptance for publication. A preference can be expressed for XML format, but any format will do for now, until the practice of immediate Open Access deposit approaches global universality (at which time it will all converge on XML as a natural matter of course anyway).
It would be a needless handicap and deterrent to insist on any particular format today. (Doc or Docx will do, so will HTML or PDF or any of the open formats.) Don’t complicate or discourage compliance by gratuitously insisting on more than necessary at the outset, and trust that as the practice of public access provision and usage grows, researchers will converge quite naturally on the optimal format. And remember that in the meanwhile the official published version will continue to be generated by publishers, purchased and stored by subscribing institutions, and preserved in deposit library archives. The public-access drafts are just supplements for the time being, not substitutes, deposited so that it is not only paying subscribers who can access and use federally funded research.)
STANDARDS: OAI will do for a start. Institutional repositories elicit somewhat richer metadata. http://www.eprints.org/software/ Once mandates become more universal, metadata standards can be raised still higher at the deposit (institutional) level and/or enriched at the harvester level.
http://eprints.ecs.soton.ac.uk/11000/
COMMENT AND FEEDBACK: Once the research content is openly accessible online, many rich new tagging, commenting and feedback mechanisms will grow quite naturally on top of them (and can also be provided by central harvesters and services commissioned by the funders themselves, if they wish, or the metrics can simply be harvested from other services for the funder’s subset of their content).
The institutional repository software allows comments. http://www.eprints.org/software/ These can be implemented at the central harvester level too. There are also ways to elicit peer commentary at the refereed journal level. (See references on open peer commentary below.)
COSTS: Institutional Repository costs are minimal (set-up and a few days per year maintenance), distributed across institutions and the IR software is free. http://www.eprints.org/ Harvester and metadata-enhancement costs can be funded centrally. The most important thing is to mandate (institutional) deposit. Further federal funding is welcome and useful, but not as essential as the mandate.
http://www.eprints.org/software/
METRICS OF SUCCESS: Institutions already have an interest in monitoring the usage and impact of their research output, and their institutional repositories already have means for generating usage metrics and statistics (e.g., IRStats). In addition there are now central means of measuring usage and impact (free services such as Citeseer, Citebase, Publish-or-Perish, Google Scholar and Google Books, as well as fee-based ones such

as SCOPUS and Thompson-Reuters Web of Science). These and other rich new metrics will be available to measure success once the deposit requirements are adopted, growing, and supplying the content from which these rich new online metrics are extracted. Which of the new metrics proves to be the "best" remains to be tested by systematically assessing their predictive power and their correlation with peer evaluations. http://openaccess.eprints.org/index.php?/archives/369-guid.html

Open Access will not only generate but also increase many existing and new metrics of research uptake, usage and impact (downloads, citations, growth curves, hub/authority scores, co-citations, etc.), but the metrics need to be validated. See citebase http://www.citebase.org/ and metrics references below as well as this bibliography: http://opcit.eprints.org/oacitation-biblio.html

PRIVATE SECTOR USABILITY: Metrics will not only make it possible for deposit rates, downloads, citations, and newer metrics and their growth to be measured and monitored, but it will also be possible to sort uptake metrics into those based on public access and usage, researcher access and usage, and industrial R&D and applications access and usage. But the urgent priority is first to provide the publicly accessible research content on which all these uptake measures will be based. The measures will evolve quite naturally once the content is globally available.

REFERENCES ON OPEN PEER COMMENTARY AND OPEN ACCESS METRICS

COMMENTARY:

Harnad, S. (1978) BBS Inaugural Editorial on Open Peer Commentary. Behavioral and Brain Sciences 1(1). http://users.ecs.soton.ac.uk/harnad/Temp/Kata/bbs.editorial.html

Harnad, S. (ed.) (1982) Peer commentary on peer review: A case study in scientific quality control, New York: Cambridge University Press. http://eprints.ecs.soton.ac.uk/3389/

Harnad, Stevan (1985) Rational disagreement in peer review. Science, Technology and Human Values, 10 p.55-62. http://cogprints.org/2128/

Harnad, S. (1990) Scholarly Skywriting and the Prepublication Continuum of Scientific Inquiry Psychological Science 1: 342-3 http://cogprints.org/1581/

Harnad, S. (1991) Post-Gutenberg Galaxy: The Fourth Revolution in the Means of Production of Knowledge. Public-Access Computer Systems Review 2 (1): 39-53 http://cogprints.org/1580/

Harnad, S. (1992) Interactive Publication: Extending American Physical Society's Discipline-Specific Model for Electronic Publishing. Serials Review, Special Issue on Economics Models for Electronic Publishing, 58-61. http://cogprints.org/1688/

Harnad, S. (1995) Interactive Cognition: Exploring the Potential of Electronic Quote/Commenting. In: B. Gorayska & J.L. Mey (Eds.) Cognitive Technology: In Search of a Humane Interface. Elsevier. Pp. 397-414. http://cogprints.org/1599/

Harnad, S. (1998/2000/2004) The invisible hand of peer review. Nature [online] (5 Nov. 1998), Exploit Interactive 5 (2000): and in Shatz, B. (2004) (ed.) Peer Review: A Critical Inquiry. Rowland & Littlefield. Pp. 235-242. http://cogprints.org/1646/

Harnad, S. (1996) Implementing Peer Review on the Net: Scientific Quality Control in Scholarly Electronic Journals. In: Peek, R. & Newby, G. (Eds.) Scholarly Publishing: The Electronic Frontier. Cambridge MA: MIT Press. Pp 103-118. http://cogprints.org/1692/

Harnad, S. (1997) Learned Inquiry and the Net: The Role of Peer Review, Peer Commentary and Copyright. Learned Publishing 11(4) 283-292. Short version appeared in 1997 in Antiquity 71: 1042-1048. Excerpts also appeared in the University of Toronto Bulletin: 51(6) P. 12. http://cogprints.org/1694/

Light, P., Light, V., Nesbitt, E. & Harnad, S. (2000) Up for Debate: CMC as a support for course related discussion in a campus university setting. In R. Joiner (Ed) Rethinking Collaborative Learning. London: Routledge. http://cogprints.org/1621/

Harnad, S. (2002) BBS Veledictory Editorial on Open Peer Commentary. Behavioral and Brain Sciences 1(1). http://users.ecs.soton.ac.uk/harnad/Temp/bbs.valedict.html

Harnad, S. (2003) Back to the Oral Tradition Through Skywriting at the Speed of Thought. http://eprints.ecs.soton.ac.uk/7723/

METRICS:

Hitchcock, S. Carr, L., Jiao, Z., Bergmark, D., Hall, W., Lagoze, C. & Harnad, S. (2000) Developing services for open eprint archives: globalisation, integration and the impact of links. Proceedings of the 5th ACM Conference on Digital Libraries. San Antonio Texas June 2000. http://cogprints.org/1644/

Harnad, Stevan (2003) For Whom the Gate Tolls? Published as: (2003) Open Access to Peer-Reviewed Research Through Author/Institution Self-Archiving: Maximizing Research Impact by Maximizing Online Access. In: Law, Derek & Judith Andrews, Eds. Digital Libraries: Policy Planning and Practice. Ashgate Publishing 2003. http://cogprints.org/1639/

Brody, T., Kampa, S., Harnad, S., Carr, L. and Hitchcock, S. (2003) Digitometric Services for Open Archives Environments. In Proceedings of European Conference on Digital Libraries 2003, pp. 207-220, Trondheim, Norway. http://eprints.ecs.soton.ac.uk/7503/

Harnad, S., Carr, L., Brody, T. & Oppenheim, C. (2003) Mandated online RAE CVs Linked to University Eprint Archives: Improving the UK Research Assessment Exercise whilst making it cheaper and easier. Ariadne 35. http://www.ariadne.ac.uk/issue35/harnad/

Hitchcock, Steve; Woukeu, Arouna; Brody, Tim; Carr, Les; Hall, Wendy and Harnad, Stevan. (2003) Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service http://eprints.ecs.soton.ac.uk/8204/

Harnad, S. & Brody, T. (2004) Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals, D-Lib Magazine 10 (6) June (Japanese translation) http://eprints.ecs.soton.ac.uk/10207/

Hajjem, C., Harnad, S. and Gingras, Y. (2005) Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact. IEEE Data Engineering Bulletin 28(4) pp. 39-47. http://eprints.ecs.soton.ac.uk/11688/

Brody, T., Harnad, S. and Carr, L. (2006) Earlier Web Usage Statistics as Predictors of Later Citation Impact. Journal of the American Association for Information Science and Technology (JASIST) 57(8) pp. 1060-1072. http://eprints.ecs.soton.ac.uk/10713/

Harnad, S. (2006) Online, Continuous, Metrics-Based Research Assessment. Technical Report, ECS, University of Southampton. http://eprints.ecs.soton.ac.uk/12130/

Brody, T., Carr, L., Harnad, S. and Swan, A. (2007) Time to Convert to Metrics. Research Fortnight pp. 17-18. http://eprints.ecs.soton.ac.uk/14329/

Brody, T., Carr, L., Gingras, Y., Hajjem, C., Harnad, S. and Swan, A. (2007) Incentivizing the Open Access Research Web: Publication-Archiving, Data-Archiving and Scientometrics. CTWatch Quarterly 3(3). http://eprints.ecs.soton.ac.uk/14418/

Harnad, S. (2008) Self-Archiving, Metrics and Mandates. Science Editor 31(2) 57-59 http://www.councilscienceeditors.org/members/secureDocument.cfm?docID=1916

Harnad, S. (2008) Validating Research Performance Metrics Against Peer Rankings. Ethics in Science and Environmental Politics 8 (11) doi:10.3354/esep00088 The Use And Misuse Of Bibliometric Indices In Evaluating Scholarly Performance http://eprints.ecs.soton.ac.uk/15619/

Harnad, S., Carr, L. and Gingras, Y. (2008) Maximizing Research Progress Through Open Access Mandates and Metrics. Liinc em Revista 4(2). http://eprints.ecs.soton.ac.uk/16617/
Harnad, S. (2009) Multiple metrics required to measure research performance. Nature (Correspondence) 457 (785) (12 February 2009)
Harnad, S. (2009) Open Access Scientometrics and the UK Research Assessment Exercise. Scientometrics 79 (1) Also in Proceedings of 11th Annual Meeting of the International Society for Scientometrics and Informetrics 11(1), pp. 27-33, Madrid, Spain. Torres-Salinas, D. and Moed, H. F., Eds. (2007)

Stevan Harnad said on December 21, 2009 at 11:53 am:

Online Open Peer Commentary Journal (Psycoloquy):
http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?
Cambridge University Press Open Peer Commentary Journal (Behavioral and Brain Sciences):
http://www.bbsonline.org/

Hilton Gibson said on December 21, 2009 at 4:26 pm:

In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
* http://ir.sun.ac.za/wiki/index.php/Digitisation
Are there existing digital standards for archiving and interoperability to maximize public benefit?
* http://ir.sun.ac.za/wiki/index.php/Digitisation
How are these anticipated to change?
* http://ir.sun.ac.za/wiki/index.php/Digitisation
Are there formats that would be especially useful to researchers wishing to combine datasets or other published results published from various papers in order to conduct comparative studies or meta-analyses?
* http://ir.sun.ac.za/wiki/index.php/Digitisation
What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?
* http://openrepositories.org and http://www.openrepository.com
Should those who access papers be given the opportunity to comment or provide feedback?
* Yes
What are the anticipated costs of maintaining publicly accessible libraries of available papers, and how might various public access business models affect these maintenance costs?
* http://www.ijdc.net/index.php/ijdc
By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?
* http://www.doaj.org

Hilton Gibson said on December 22, 2009 at 4:31 am:

In addition the above, please see the following:
* http://ir.sun.ac.za/wiki/index.php/Open_Access
* http://ir.sun.ac.za/wiki/index.php/Digital_Preservation

Barbara Kirsop said on December 22, 2009 at 10:09 am:

The Electronic Publishing Trust for Development, a UK-registered charitable Trust, works both to improve access to research information by the research communities in the developing world and to ensure the incorporation of research arising from these regions into the global knowledge pool. We greatly welcome this initiative to accelerate the process. Without the worldwide distribution of research knowledge on a 'level playing-field basis', the many problems facing the planet (climate change, infectious diseases, agricultural challenges through drought/flooding . . .) will not be resolved. Open access to publicly funded research publications is an essential first building block to sharing research between all countries. This strategy, the details of which have been described by others, was proposed some years ago. It is now well-established in many universities, institutes and funding organisations around the world and is increasingly understood by the research community as being the best way to increase the use of published findings for the public good. The framework exists and is operating successfully, providing access to all researchers through internationally developed standards.
As evidence of growing global acceptance, the following statistics can be found from online open access databases:
National/institutional/departmental mandates requiring open access; see ROARMAP
Total numbers 176 (from developing countries 13 - 7%)
Institutional repositories; see Registry of OA Repositories
Total numbers 1552 (from developing countries 324 - 21%)
OA Journals; see Directory of Open Access Journal
Total numbers 4507 (from developing countries ~773 -17%)
From these figures (at December 17th 2009) it can be seen that developing countries are increasingly aware of the benefits of open access and beginning to adopt open access policies and establish open access repositories and open access journals. While the open access repositories and journals already provide free access to hundreds of thousands of research publications, there remains much to be done in raising awareness of such benefits both to administrators and the research communities. Furthermore, training in the establishment of institutional repositories and setting up open access journals is essential. Fortunately, there is an increasing volume of information available online, and a good example of this is from the Open Society Institute-supported Open Access Scholarly Information Sourcebook (OASIS).

But the best mechanism for accelerating adoption of what is now clearly accepted as the right way foreword for the global distribution of knowledge, would be the adoption and implementation of policies by major research countries. The EPT greatly welcomes this initiative by the OFST and hopes that you will be encouraged by developments so far and by the increasing need for sharing research publications to meet the urgent needs of the world, particularly of those in resource-poor nations. The poorer nations need an independent science base on which to strengthen their economies. Without free access to existing research information they will remain forever dependent.
Trustees of the Electronic Publishing Trust for Development
Barbara Kirsop (UK)
Subbiah Arunachalam (India)
Leslie Chan (Canada)
Margaret Ling (UK)
Judy Ugonna (UK)
Vanderlei Canhos (Brazil)
Daisy Ouya (Kenya)
Virginia Cano (UK)
Brian Kirsop (Treasurer, UK)
Web site: http://www.epublishingtrust.org
EPT Blog: http://www.epublishingtrust.blogspot.com

Anali Perry said on December 23, 2009 at 2:14 pm:

On behalf of Arizona State University Libraries, I'd like to specifically address the following questions:
In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
Papers should be submitted to a repository in whatever format is most convenient to the author. However, the repository should ensure that the version that is accessible is displayed in a robust, searchable standardized format, such as the XML-based National Library of Medicine DTD standard (see http://dtd.nlm.nih.gov/), which is supported by both the Library of Congress and the British Library (http://www.loc.gov/today/pr/2006/06-097.html).
Are there existing digital standards for archiving and interoperability to maximize public benefit?
The Open Archives Initiative (OAI) at http://www.openarchives.org/ is an excellent standard already in use by most major digital archives and repositories, including PubMed Central (see http://www.ncbi.nlm.nih.gov/pmc/about/oai.html). Using common standards such as Digital Object Identifiers (DOI) (http://www.doi.org/) to facilitate discovery and Creative Commons Licensing (http://creativecommons.org/) to streamline permissions would be other standards to consider.
How are these anticipated to change?
These standards evolve through the support of a community of practice dedicated to the perpetual preservation and access to information. Any changes that arise will continue to support this mission and ensure stable archival and access standards.
I appreciate this opportunity to comment on such an important initiative. It's encouraging to see the OSTP solicit ideas and comments in this public forum. I hope that all of our remarks will be useful as you consider the next steps.
Thank you,
Anali Maughan Perry
Assistant Collections & Scholarly Communications Librarian
Arizona State University Libraries

B Klein said on December 23, 2009 at 4:32 pm:

This recommendation falls under implementation, standards, metrics and management.
A basic first step is to mandate, standardize, and implement policies and procedures to identify and give Notice of US Government Sponsorship. The notices should be included on the document in two formats. The first should be readable by humans and the second should be electronic metadata readable and searchable by machines. Currently there is no requirement or format in use to identify "works of the U.S. Government" (Title 17 USC 101 & 105) authored by Federal Government employees.
There are at least three (3) different funding instrument categories: Grants, Contracts, and Government Works. The first two already require and specify the format of the notice.
CURRENT REQUIREMENTS:
GRANTS:
See "Assistance Terms & Conditions" (2008). National Science Foundation http://www.nsf.gov/pubs/policydocs/rtc/termsidebyside.pdf
Sec 215.51(a) Monitoring & Reporting Requirements:
(a) Publications. The recipient is expected to publish or otherwise make publicly available the results of the work conducted under the award. An acknowledgment of awarding agency support must appear in the publication of any material, whether copyrighted or not, based on or developed under this project, as follows
*****(1) The acknowledgment will be:
This material is based upon work supported by the [name of awarding agency(ies) under Award No. [recipient should enter the awarding agency(ies) award number(s).
CONTRACTS:
See "Federal Acquisitions Regulation." Section 52.227-14 — Rights in Data – General. As prescribed in 27.409(b)(1): (c) Copyright— (1) Data first produced in the performance of this contract.
(i) Unless provided otherwise in paragraph (d) of this clause, the Contractor may establish, without prior approval of the Contracting Officer, claim to copyright in scientific and technical articles based on or containing data first produced in the performance of this contract and published in academic, technical or professional journals, symposia proceedings or similar works. The prior, express written permission of the Contracting Officer is required to assert copyright in all other data first produced in the performance of this contract.

*****(ii) When authorized to assert copyright to the data, the Contractor shall affix the applicable copyright notices of 17 U.S.C. 401 or 402, and acknowledgment of Government sponsorship (including contract number).
(iii) For data other than computer software, the Contractor grants to the Government, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in such copyrighted data to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.
GOVERNMENT AUTHORED WORKS
There is no standard notice of government funding for articles and papers authored by a Government employee. Articles created on-the-job are "works of the U.S. Government" and are not eligible for copyright protection (Title 17 USC Sec 105). Scholarly journals routinely require scientists to sign standard copyright or royalty agreements which purport to transfer copyright in the employee's work to the publisher. There being no copyright to give away for U.S. Government works, such agreements have no force. Nevertheless, they cause confusion for Government authors, publishers and the public. Publishers may only claim copyright in original material they add to a Government work. Section 403 of the Copyright Law is aimed at a publishing practice that, while technically justified under present law, is misleading. In cases where a Government work is published or republished commercially, publishers frequently add some "new matter" in the form of an introduction, editing, illustrations, etc., and then include a general copyright notice in their name. This in no way suggests to the public that the bulk of the work is uncopyrightable and therefore free for use.

+2 Evans Boney said on December 23, 2009 at 6:42 pm:

I'll only respond to the questions I know enough about to have an opinion. Otherwise… Stevan's thoughts above are clearly more well-researched than my own, so I'll defer to his opinion on the more technical questions.
* In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
Format doesn't really matter, but it would be useful (very) to standardize some fields one way or the other. The reason is that this would make searches of scientific literature much faster, as the search engine would have a set number of fields to work with for every article. As image searches come into their own, making searchable figures is a particularly tantalizing prospect. I know several times I've gotten a really odd function that I know I've seen somewhere before, but there's currently no way to search for "similar graphs". There could be, and there should be, it would be a boon to research.
Another important, small, change is to make an AOI (like an author's doi number) to avoid problems like finding my papers as E. Boney, E. T. Boney, E. T. D. Boney, or Evans T. D. Boney, as all will likely appear eventually, depending on journal and format. Any my name is spelled with English characters, imagine the trouble this causes our naturalized scientists! This was addressed in last week's Nature, and I would suggest following these suggestions to a tee.
* What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?
Google Scholar. Frankly, before GS came online, the science community was so scattered that any one search was a virtual waste of time, as it would only index a certain subset of journals. I think they clearly have the best search algorithm, and they have been far better for interdisciplinary research than any other engine, public, private, or otherwise that I've encountered.
In terms of ordering the articles, ArXiv and PubMedCentral are an amazing resource for many communities and, perhaps with a better GUI, they could provide a skeleton architecture for the new, larger database.
In terms of commenting on an article, I like the new site journalfire.com for indexing all the databases, but I admittedly have a conflict of interest there (see below).
* Should those who access papers be given the opportunity to comment or provide feedback?
Not attached to the paper in the government database. That should be, in my mind, a place only for peer-reviewed papers and comments. I will also declare a conflict of interest by way of a suggestion: I'm on the student board of directors at a Caltech startup, journalfire.com , that seeks to be a better place for a more general sort of comment or clarification. My participation in this company is driven by my vision of a better future for scientific collaboration, but I also have some stock options in the case that they do wind up making any money in the next decade.
* By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?
For older articles, average unique visitors per article per citation should be a good metric of the efficacy of your setup. More citations mean that a paper is more important in the science community. Unique visitors mean that the community is accessing them in your database. So unique visitors per article per citation is a way of gauging the success of the database in connecting the important articles with their intended audience, regardless of discipline. Clearly this metric has problems near the time of publication (when both number of citations and number of visitors will start at 0), and so perhaps it should only be measured after 12 months? After both visitors and citations reach a threshold of size 2? Just some food for thought, clearly I'm not 100% sure what the right way to do this is.
Thanks so much for taking our comments on this subject, I really look forward to the implementation of a robust open-access policy!
Evans T. D. Boney
4th year PhD Candidate in Chemistry Theory
Noyes Lab of Chemical Physics, Caltech

Ajay Ohri said on December 23, 2009 at 8:10 pm:

In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
The Best format fo submission is Text. Other forms may have problems of size and searchability.
Post submission A LATEX format, XML document or a PDF format may be used by the Federal Agency.
For easy viewing- the best format for PDF is Slideshare.net as well as Google Doc presentation.
Are there existing digital standards for archiving and interoperability to maximize public benefit?
Yes.
How are these anticipated to change?
Change to make it easier to index, view manually as well as store data more optimally.

Are there formats that would be especially useful to researchers wishing to combine datasets or other published results published from various papers in order to conduct comparative studies or meta-analyses?
Yes- Best is Citrix server with R Datasets or SAS datasets. This enables browser based processing.
An example of using big datasets is at University of Tennessee at http://analysis.utk.edu
What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?
Google- and Microsoft research/ Slideshare
Great way of searching huge amounts of data for easy viewing.
Should those who access papers be given the opportunity to comment or provide feedback?
Yes, but moderated for spam and with auto submission to paper authors.
What are the anticipated costs of maintaining publicly accessible libraries of available papers, and how might various public access business models affect these maintenance costs?
Costs are likely to shrink with time, and using advertising, or even a simple banner ad for branding will help the Agency cut costs.
By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?
Recency How recent are the views /comments
Frequency (Number of Views, Number of Sharing, Number of Forwards by Email or Social Media, Citations)
Satisfaction of Users (Star Ratings, Up or Down Ratings, Time spent reading the article)
Ip Addresss level analysis of people visiting the post/paper

Mike Serfas said on December 23, 2009 at 9:37 pm:

1. Format
In a sense, submission is not strictly necessary at all: what matters most is to establish that individual readers have the right to download or scan the journal article and place it online. This could actually be the way that material in some areas is "submitted" while people are busy getting their central servers up and operational, as there is no reason to sacrifice the right to access weeks or months of publications solely due to budget shortages or server logistics.
I would also suggest that an offprint or good copy of a paper should always be acceptable as a fulfillment of the author's legal requirement. If the author fails to submit anything, then personnel at the repository should obtain this on their own. The public access system should not be enforcement-driven, but should instead work to repair any deficiencies. It is far faster and cheaper to scan in papers than to deliberate whether a researcher should lose funding for failing to comply with public access requirements, and this also avoids damaging controversy. Consider for example that because 340,000 papers are submitted yearly from the U.S., hundreds of hard drive failures will likely occur and several authors will likely die on the day of submission.
Of course the vast majority of researchers are highly motivated to encourage distribution of their works, and will gladly make submission in a convenient format. Typically this might include a text file in the OpenOffice native ODF text format (.odt), or Word document formats (.doc) that are covered by the Microsoft Open Specification Promise, accompanied by figures in open formats such as .xcf, .png, or .svg. To conserve research funds, repository personnel should make sure not to require submission in proprietary formats that would require software purchases by researchers. Any repository should expect to create .pdf-style files from scratch, because the fonts and layout from a journal publication are pretty clearly prone to copyright claims by parties other than the publicly funded researchers.
The most trouble is likely to come from "supplementary data" (or "supplemental data") referenced in published papers. This often appears on a journal Web site, but sometimes at the author's institution or at some other location. It is not always actually accessible at the time when the paper appears online or in print, and it may not remain accessible afterward. It can appear in any format however obscure. Despite all these things, supplemental data can be crucial for understanding a paper - especially for certain of the "best" journals where authors may be told to cut manuscripts to half the original size, then submit all the leftovers as supplemental data. The repository will need to look closely to archive all of these supplements correctly. Authors might be allowed to submit an extra version of their work, which has been peer reviewed but remains unabridged with the supplemental data fully integrated. This should be made available at the repository in addition to the final published form.
6. Feedback.
I much approve of online discussion forums, but they are not a part of this project. This program should only archive those comments which are made by federally funded employees, submitted to peer-reviewed journals, and approved for publication. The repository should never need to hire anyone capable of deciding whether a comment is reasonable or advocates a fringe theory.
As feasible, the central repository should offer links to notable online resources, including forums, if they exist.
7. Costs.
We know that Wikipedia maintains three million articles online, with hundreds of archived revisions and very heavy reader access, for a fraction of its $6 million yearly budget. The total number of scientific publications is around one million papers yearly, so we should not expect the cost of this program to become excessive.
The publications are also an outstanding opportunity for targeted advertising, with specific manufacturers identified in the text by name and product number. Offering direct links to their catalog pages would actually improve the usability of the site, and could generate useful funds.

Stevan Harnad said on December 24, 2009 at 9:48 am:

The free, open-source, open-access repository softwares, EPrints and DSpace — and the dream of data-sharing:
Origin of OA IR Softwares: http://bit.ly/4i2wDG
Developer Rob Tansley: http://bit.ly/5gynrf
CalTech Review: http://bit.ly/5PCw2k
EPrints: http://www.eprints.org/
DSpace: http://www.dspace.org/
Data Sharing: http://bit.ly/8aWAf7

Stevan Harnad said on December 24, 2009 at 10:24 am:

Worldwide and US Repository Stats from ROAR:
Repositories Worldwide (1557): http://roar.eprints.org/
Institutional Repositories Worldwide (867): http://bit.ly/82Sv0d
Central Repositories Worldwide (147): http://bit.ly/58kCWM
Repositories US (301): http://bit.ly/7Atxhk
Institutional Repositories US (172): http://bit.ly/6CWmQm
Central Repositories US (30): http://bit.ly/8A6pJQ
DSpace Repositories Worldwide (502): http://bit.ly/7wqy2X
DSpace Repositories US (88): http://bit.ly/5L9Pn3
EPrints Repositories Worldwide (358): http://bit.ly/7sU2VX
EPrints Repositpries US (64): http://bit.ly/4CokNZ

+1 ann viera said on December 24, 2009 at 10:26 am:

Are there formats that would be especially useful to researchers wishing to combine datasets or other published results published from various papers in order to conduct comparative studies or meta-analyses?
See Philip Bourne, et al. "Open Access: Taking Full Advantage of the Content" PLOS Computational Biology March 2008 V. 4 No. 3
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2275780/

+2 Hope Leman said on December 24, 2009 at 1:02 pm:

• In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
I second Ann Viera's suggestion that the article by Philip Bourne, et al. "Open Access: Taking Full Advantage of the Content" PLOS Computational Biology March 2008 V. 4 No. 3
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2275780/
be referred to in your deliberations on formatting issues.
And, although, PDF is not, as I understand it, Semantic Web and search-engine friendly, all materials should be offered in that format in addition to other formats. PDF is an easy way to send articles of interest to colleagues and other interested parties. A key point of your initiative, it seems to be me, is not merely Open Access, but the widest possible dissemination of the materials you are going to make available and PDF facilitates, via email and other transmission modes, such dissemination.
And I hope it is a given that accessibility for the disabled will be kept in mind as a paramount concern and that all materials will be optimized for accessibility in a cutting-edge fashion from the get-go and that features for the disabled not be merely grafted on as a perfunctory afterthought to meet regulatory requirements such as Section 508 http://www.section508.gov/index.cfm?FuseAction=Content&ID=3
• What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?
Nature Precedings http://precedings.nature.com/
is an exemplar of an Open Access site. The papers are easily downloadable. The site is handsome. The materials are eminently searchable and subscribeable. It is much better looking than PubMed Central http://www.ncbi.nlm.nih.gov/pmc/
• Should those who access papers be given the opportunity to comment or provide feedback?
Absolutely. That is imperative. Indeed, that would be one of the primary reasons for your entire Open Access initiative. The promise of Open Access is that it will create a system of worldwide peer review and ensure that disasters such as the stem cell fraud that occurred in the journal Science 2004-2005, for instance, will not be repeated. The more people that can scrutinize data and articles, the better. That is the promise and power of Science 2.0 and Open Science.
Additionally, as we enter the age of Participatory Medicine, shared decision making in medicine and the rise of the e-Patient, laypeople should be able to access as much data as possible. See this useful article for instance: Mayer M. A seat at the table: a research advocate's journey. J Participat Med. 2009(Oct):Launch Issue.
Published: October 21, 2009.
http://jopm.org/index.php/jpm/article/view/25/29
Moreover, your Open Access initiative promises to address some of the problems of peer review as it is currently practiced. See Smith RW. In search of an optimal peer review system. J Participat Med. 2009(Oct): Launch Issue.
http://jopm.org/index.php/jpm/article/view/12/25
and Frishauf P. Reputation systems: a new vision for publishing and peer review. J Participat Med. 2009(Oct):Launch Issue.
Published: October 21, 2009.
http://jopm.org/index.php/jpm/article/view/11/21
• By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?
I suggest that you hire as consultants on these matters such authorities as Peter Binfield Managing Editor, PLoS ONE
http://www.plos.org/about/people/one.html and Cameron Neylon. They are both leading experts in this area and are respected throughout the Open Science and Open Access communities. See the article Neylon C, Wu S (2009) Article-Level Metrics and the Evolution of Scientific Impact. PLoS Biol 7(11): e1000242. doi:10.1371/journal.pbio.1000242
Published: November 17, 2009
http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1000242

+1 Tito Jankowski said on December 27, 2009 at 2:49 am:

I graduated from Brown University's BioMedical Engineering program 2 years ago. Now, while pursuing non-institutional biotech for a startup company, I regularly get my articles by emailing colleagues an html link to the paper, requesting a PDF copy. I have a handful of associates who do the same.
An important question is who is the target audience of these improvements.
Tito

mike marchywka said on December 27, 2009 at 7:51 pm:

To: publicaccess@ostp.gov
Subject: comments re docket E9-29322 "Public Access Policies …"
From: Mike Marchywka, Marietta GA 30064 marchywka@hotmail.com
[First let me apologize for sloppy editing, I was going to submit via email but finally decided to try blog first]
Hi, I'm responding to the RFC published at
http://edocket.access.gpo.gov/2009/E9-29322.htm
[Federal Register: December 9, 2009 (Volume 74, Number 235)]
[DOCID:fr09de09-111]
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
Public Access Policies for Science and Technology Funding Agencies Across the Federal Government
AGENCY: Office of Science and Technology Policy (OSTP), Executive Office of the President.
ACTION: Notice; request for public comment.
_____
I became aware of this solicitation via wikipedia, http://en.wikipedia.org/wiki/Wikipedia:Village_pump_(miscellaneous)#Public_access_to_U.S._federally_funded_science as I often contribute to and read this resource which benefits greatly from freely available government sponsored works. Previously, I have responded to requests from other agencies [3-6] about seemingly unrelated regulations but my conclusion and suggestions have always been the same: we need better computer readable information to reach the widest possible audience with the most complete results. In this case, the details or what and when to release full text seem to be largely an issue of publisher revenue models. I will focus on question "8• from the RFC, more on "how" or in what formats to release the data and text, and generally advocate that publications and suitable raw data be made available via an "API"[8,10], and not just a human readable web interface. In the detailed responses below, I have tried to reword essentially the same "API" notion to show how it applies in each case, this leads to some redundancy however. An API provides flexible and vendor neutral access to the information for automated analysis or repackaging by anyone. I would also comment that "simpler is better" and make sure that current standard formats not obscure or confuse information with various features. When works are available in formats such as pdf, be sure to include a requirement that the computer readable information, usually the text, be easily available- I couldn't even get text from some PDF IRS instructions in a comprehensible format and presumably there is no reason to restrict these. Some agencies accept "scanned" or other types of pdf files (see for example submissions on Drugs@FDA [2] which contrast with the structured and versatile documents that the SEC[7] is adopting, I recall problems extracting information from FCC submissions too) which do not allow for best use of computers to automate data processing. Often, these alternative formats encumber the information with unhelpful formatting and security "features." The NCBI eutils facility[1] provides an excellent example of an "API" from which most information is available in a simple, appropriate computer readable format using automated access tools. Overall, we need to think about what computers can do and not just try to make computers act like paper nor just throw every high-tech "standard" and high margin feature into the "solution". The governmment need not anticipate the needs of every possible user, just make sure that their chosen interface doesn't limit those with simpler equipment or those who wish to repackage the information (Wikipedia being one example of such a group). Where needed, revenue models can be changed too(I guess if banks can get TARP what about a bailout for publishers?) but this is a political and business issue and I will confine my remarks largely to less controversial issues I have found while trying to use available publications.
Responses to Specific Questions:
==========
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?

The federal government is an originator of both data and analyses in many fields. I'm not sure if "peer review" is generally associated with academics or hard sciences, but just to be clear on the obvious, the government's data and analysis authoring role extends into many fields including weather, securities, demographics, etc. Often, government generated raw data is invaluable to authors elsewhere.

Except in some cases, everyone wants unrestricted access to information but the limiting issue is often the revnue problem. Probably the private sector publishers are the biggest moderating influence against unrestricted free access to everything. They contribute valuable editorial and peer review facilities and often their financial objectives restrict availability of content. Any regulations which restrict their ability to profit could impact review quality, even if many peer reviewers are paid by other institutions.

For me, the libraries and universities serve as "access domains" within which more journals are available. Without publisher imposed restrictions, they would not be relevant to most electronic journal access.

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?

Regardless of issues on full text access, we need a good database of articles searchable via an automated "API." Even if some papers are not available for free, we need to know that they exist and generally what they have concluded. Teaser abstracts should be avoided when it is known that full text will be restricted for any length of time(" did our experiment work? Buy the paper and find out"). The biggest problem I've seen so far from many government information sources isn't even so much the access policy as much as the usage of confining computer tools that are often skewed towards human readability at the expense of automated data processing or favor one company's products, which tend to be similarly limiting. In the case of journals, this is probably driven largely by publishers' desire for digital rights management that comes from some software formats, as well as sales efforts from some software vendors. I have not yet found an author who wanted to restrict usage of his own work and most users of course want flexible unencumbered access to either human or machine readable results. The general public presumably benefits if more people can repackage the scientific results for different audiences, a task facilitated with computer readable publications.

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?

It probably is worth stating that peer reviewed govt funded work is often the best source of information on most topics and everyone would end up using it directly or indirectly if it was easy to repackage at places like Wikipedia. I use peer reviewed publications for for everything from stock research to general interest. Essentially every aspect of life today is better served by examination of primary sources and original scientific works. I tend to use the NCBI eutils facility for almost everything related to medicine and would use similar facilities if they existed for agriculture, physics, or weather. With the NCBI eutils facilities for example, members of the public can and have written their own computer programs to reformat citations into a form that can be used by wikipedia, allowing even more people to benefit[9], " This uses a cygwin bash script to invoke java code and some other pubmed eutils scripts [...] You should be able to integrate this into your own scripts if desired for testing."

The existence of a computer readable API, and not just web interfaces or display formats which confine the use to special purpose applications like a pdf reader, allows it to be used more flexibly to reach a wider audience.

I often use articles for general background or exploratory work and may post excerpts on various message boards or generate my own notes. This usually involves skimming lots of articles and abstracts. I could not do this if I had to click and download a pdf file for each article of interest. Obtaining a short text file of all abstracts in one place, which I can do with the NCBI eutils facility without the need to even use a SOAP API, is a big benefit as I can use it with a large number of software products including scripts I have written myself.

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?

Generally, I think the best strategy here is to let the private sector worry

about addressing specific audiences by allowing companies to repackage government information for best use by their own audiences. This means providing computer readable publications with a simple and flexible API. Also note that private sector doesn't mean for profit, I think I mentioned wikipedia earlier.

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

Most researchers want publicity and indeed those I have talked to are often confused that some file formats are restrictive and many authors don't even like publisher restrictions. I guess if you could find a way around publisher revenue models that would eliminate most problems with compliance.

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?

I have gotten everything from early manuscripts to actual reprints in response to specific queries to authors. Personally I find reviewer and editor comments and questions helpful but these are not supposed to serve any archival purpose or even become public, just fix the paper. The final product is presumably the best intellectual quality.

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?

AFAIK, the only concern here is publisher revenue. I think some people on the blog mentioned patent or security issues but as I understand the question it is only over publisher revenue, public disclosure has already been made in a journal. Optimal of course depends on your figure or merit- for a publisher, never would be good (up to the point of creating disinterst maybe) but for those who are constantly trying to sanity check various ideas, immediately would be optimal. I'd be wary of anyone who has a quantitative figure of merit. Researchy results lead to undpredictable usages and their relevance to particular usages can be short or long lived.

8. How should peer-reviewed papers arising from federal investment be made publiclyavailable? 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?

See related comments under the other sections. Generally, I have seen that information gets obscured as people try to inflict paper models, even with the latest hitech features, onto information in an attempt to enhance human readability while destroying automated usages of the data. The format needs to be simple, computer readable and computer searchable, and not something supported only with tools predicated on human interaction even if these claim to be "Standards" based. Generally for articles, this means a simple text format that can be separated from the article artwork and page layout/columns, often this is difficult to obtain from pdf submissions. Tabular data should be available in something like a line-oriented text format (csv maybe). The SEC is moving towards highly structured XML submssions, the NCBI eutils facility cited previously offers an excellent and simple API for searching and repackaging of research results. While I have made several negative comments about some formats that are well supported with some authoring tools , the existence of various commercial authoring tools should be considered a positive but with a federal standard commercial products should become available and more versatile. Commercial tools of course are often designed to lock an author into formats most beneficial for the vendor in the absence of strong sentiment otherwise.

I'm not sure I can emphasize this issue enough as it isn't just in this area where commercial interest exist but in many publications that have no reason to be obfuscated. Many govt and private authors may not be aware of these issues. Indeed, I couldn't even extract text instructions from PDF IRS publications in a useful text format. Presumably there is no reason to "protect" these and even filled out forms like 1040 should allow a private user to extract information without all the archaic formatting or use of proprietary commerical tools. Federal courts seem to encourage documents be submitted in an unreadable format even when these submissions will be public but the SEC requires structured documents be submitted for ease of computer readability. Generally people probably need to think about computers as something that auomates data processing and not just try to make them act like paper or papyrus or clay

tablets that paradoxically contain lots of high-tech resource hungry features .
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

The largest issue here is computer readability. Many govt and private sites focus on human readability- and this makes sense if you are supported by advertisers but it prevents the best use of information and computers. Certainly govt access does need to accomodate the casual or naive user. Interactive web pages that present human readable information is important but no website designer is omniscient and computer readable information allows others to address the needs of many potential users. Many human readable formats obscure, clutter, and even remove information needed by others. This needs to be avoided as the only means of making data and information available. Computer readability insures that the private sector can repackage and use this information to make it available to even more people who may currently not even know about it. The revenue model in the private sector has actually created some barriers to usage that the federal government has and should continue to avoid.

Metrics are a huge problem if you need to measure results such as utility to a reader and not just click-thru rates. Tying a publication to economic value would be difficult but just publishing raw usage rates would at least tell us if anyone is using a facility. I guess there is a concern that “artistic merit” will be used as an evaluation criterion. While a friendly human readable website is a big plus, it is largely a dead end for most users and its usage shouldn’t be considered as the only factor. Private sector figures like click thru rate or interaction measures may not make much sense here. You may only need to use a API once to download the paper and data that let’s you license the govt patent you need to start your business.

Since many issues are technical and “Computer related”, I’d like to be able to cite someone like IEEE as a good example (and skimming other comments at blog.ostp.gov they are mentioned and represented here) but they don’t have much for the public and in any case I’m not sure if they offer any API to members for automated access to their own journals. Personally for the human readable website, something as simple as Google Scholar would be fine,
http://scholar.google.com/scholar?q=marchywka
which really isn’t much different from Citerseer,
http://citeseerx.ist.psu.edu/search?q=marchywka&submit=Search&sort=rel&ic=1
but Citeseer generally offers full text in a format which is not computer readable. As far as I know there is no API for automated access (I’ve gotten kicked off of one commercial search engine for using automated access) .
Neither of these are much different from the pubmed website,
http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed
but again I wouldn’t worry too much about a super fancy and complicated website- often these mean you need the latest browser and lots of memory, CPU, and bandwidth. Something simple that runs on older PC’s or cell phones is great, especially if there is an API to let others add value to your article database. Last time I checked, pubmed had a hard time with cell phones but they do have a text version. A special text version would not be needed if their main site html was simpler. For my needs, however, I have to option of writing my own interface using eutils results.

Thanks.

References

==========

[1] http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
[2] http://www.accessdata.fda.gov/Scripts/cder/DrugsatFDA/
[3] http://www.sec.gov/comments/s7-04-09/s70409-2.pdf
[4] http://www.sec.gov/comments/s7-27-08/s72708-19.pdf
[5] http://www.federalreserve.gov/SECRS/2008/December/20081210/OP-1338/OP-1338_7_1.pdf
[6] http://files.ots.treas.gov/comments/fddad554-1e0b-8562-eb37-34e416089fee.pdf
[7] http://xbrl.sec.gov/
[8] http://www.mail-archive.com/bbb%40bioinformatics.org/msg00145.html
[9] http://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Medicine/Archive_15#pubmed_cits_to_wiki_conversion_tool.2C_temporary_test
[10] http://en.wikipedia.org/wiki/Application_programming_interface

note new address

Mike Marchywka
1975 Village Round
Marietta GA 30064
415-264-8477 (w)<- use this
404-788-1216 (C)<- leave message

+1 Johann van Reenen said on December 27, 2009 at 8:59 pm:

In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
I agree that papers should be submitted to a repository in whatever format is most convenient to the author. However, I agree with Harnad that "there is no need at all to be draconian about the format of the deposit. The important thing is that the full, peer-reviewed final draft should be deposited in the fundee's (OAI-compliant) institutional repository immediately upon acceptance for publication. A preference can be expressed for XML format, but any format will do for now, until the practice of immediate Open Access deposit approaches global universality… " The onus should be on publishers to provide the author with a final copy for deposit ("self-archiving") as the author and most editors provide free products, services, and expertise to publishers.
Are there existing digital standards for archiving and interoperability to maximize public benefit?
Yes, there are globally approved standards that are continually being upgraded based on developments in interoperability and – in future – semantic web functionality. The Open Archives Initiative (OAI) at http://www.openarchives.org/ was the first and still in use by most major digital repositories. There are many others, see for instance "Training and Capacity building. Recommended Reading for Repository Managers. From an Investigative Study of Standards for Digital Repositories and Related Services" online at: http://plip.eifl.net/eifl-oa/training/reading/from-investigative-study
How are these anticipated to change?
Changes are occurring regularly through user groups and other organizing bodies, such as the DuraSpace group, e-Prints group, et cetera. Ongoing research at Cornell, Los Alamos National Laboratory Library Protocol group, e-Prints group at U. Southampton, and others are exploring new ways to provide seamless services but much depend on the cooperation of publishers.
Should those who access papers be given the opportunity to comment or provide feedback?
Social software systems are now mainstream and should be included in any repository and archival service as a value added for future data mining.
Thank you, Johann van Reenen, University of New Mexico

Stevan Harnad said on December 28, 2009 at 12:05 pm:

Please bear in mind, in considering what should be mandated, that the one who is bound by the federal mandate is the federal fundee, not the publisher.
The advantage of the fundee's final, revised, refereed, accepted draft over the publisher's proprietary version is that the fundee's draft is far less bound by publisher constraints, and hence leaves fundees far less constrained in their choice of journal. http://romeo.eprints.org/stats.php
In contrast, the gains to the user from having access to the publisher's proprietary version rather than the fundee's final refereed draft are minimal — and especially when compared to having no non-subscription access at all (which is what open access is primarily intended to remedy).

+1 Hope Leman said on December 28, 2009 at 12:25 pm:

It is clear how important your initiative is given
Tito Jankowski's comments above, "I graduated from Brown University's BioMedical Engineering program 2 years ago. Now, while pursuing non-institutional biotech for a startup company, I regularly get my articles by emailing colleagues an html link to the paper, requesting a PDF copy." Apparently, even the author of a paper has to spend precious time and that of colleagues simply to acquire copies of his own work and share it with others. Good for you for addressing such needless impediments to the advance of science.
I have just visited the Life Scientists room of FriendFeed
http://friendfeed.com/the-life-scientists
(and suggest that everyone involved in your initiative join that community, given the wealth of knowledge and expertise on Open Science, Open Data and text mining that its members possess and evidence literally every few minutes) and therein came across a reference, "The Semantic Biochemical Journal experiment"
http://duncan.hull.name/2009/12/11/utopia/
to this outstanding, edifying article, "Calling International Rescue: knowledge lost in literature and data landslide!"
http://www.biochemj.org/bj/424/bj4240317.htm
which, among other things, discusses the idea of "liquid publications"
http://project.liquidpub.org/
which is somewhat akin to Elsevier's "Article of the Future"
which, however, seems to have fallen somewhat flat:
The "Article of the Future" — Just Lipstick Again?
http://scholarlykitchen.sspnet.org/2009/07/21/the-article-of-the-future-lipstick-on-a-pig/
Johann van Reenen points above are important, "Social software systems are now mainstream and should be included in any repository and archival service as a value added for future data mining." And his suggestion here, "The onus should be on publishers to provide the author with a final copy for deposit ("self-archiving") as the author and most editors provide free products, services, and expertise to publishers…" would address the lamentable rigmarole that Tito Jankowski has to engage in in order to disseminate scientific information.

+2 Eric Patridge said on December 28, 2009 at 1:55 pm:

FORMAT. Peer-reviewed articles generally move from author to publisher and from publisher to reader. In this process, there are generally two opportunities for published papers to be "submitted," depending on your perspective. In accordance with this two-sided workflow, it is important to consider both the convenience of submitting articles to a publisher and the accessibility and format of papers made available by publishers. Here I offer input regarding the best formats for peer-reviewed papers that are made available by publishers. (I believe this includes both general readership audiences and data-mining search engines.)
In order to maximize federal research funds, the accessibility of previous research should be fast, thorough, reliable, and measurable. This includes both the accessibility of papers published for general readership audiences and the accessibility of data published for search engines, meta studies, or reviews.
Perhaps the easiest and most common format for current general readership is PDF format. There are numerous tools that are freely available which are able to open, edit, save, print, and copy from PDF files. In terms of new infrastructure, PDF format aligns well with current infrastructure and would require little restructuring. PDFs are widely searchable and there are several methods available for creating searchable PDF databases.
Importantly, general readership is not the only audience; data-mining search engines should be remembered when considering the format of published papers. It is increasingly important to keep information in a searchable and catalogued format. In order to be most accessible for data-mining search engines, it is reasonable to consider scripting techniques such as LaTeX and XML as possible formats. Such scripts can be highly descriptive, specific, and ordered. The publication resulting from such scripting is significantly more ordered and searchable than a PDF. In addition, the resulting publications can more easily be digitally re-structured if needed. More than this, it is also reasonable to remember that PDFs can easily be created from scripted publications – thus also satisfying the needs for the general readership.
Therefore it would seem that a scripted publication would ultimately be the best format for published articles – with PDF outputs available for the general readership. As a side note, it may be counterproductive to expect that authors be capable of scripting in LaTeX or XML. Rather, to incorporate LaTeX or XML into current infrastructure, a more productive model may include a few personnel at each publisher or at central locations who work together to solidify a model scripting language that is useful and specific enough for the STEM fields.
CHANGE. It is evident that digital infrastructure is currently ever-changing. Email has been popular for almost 30 years. The internet has been popular for about the last 20 years. Online journals have only become widely used in the last 10 years. It is difficult to anticipate the future of digital infrastructure and it will be increasingly difficult to predict changes more than 10 to 20 years ahead. In anticipation for unexpected needs, a more ordered digital format would maximize current efforts. An ordered format such as LaTeX or XML would enable online data and information to be computationally re-structured with ease.
FEEDBACK. An online database of published papers would significantly augment the audience size and population, and therefore the readers' needs are less predictable. It is always reasonable and necessary to collect feedback regarding the service that furnishes access to anything. Otherwise it will be more difficult for the general readership to alert authorities of infrastructure malfunctions – especially by underrepresented communities who may have limited access to resources.
METRICS.
- Enhanced accessibility could be one basic measurement. (i.e. the "lag time" between accessing a search engine and successfully obtaining a published article)
- The number of graduate students, postdoctoral fellows/associates, associate or senior research scientists, and professors who will have enhanced accessibility.
- The number of non-profit organizations that will have enhanced accessibility.
- The number of underrepresented organizations that will have enhanced accessibility.
- The estimated amount of time saved by total researchers in searching for articles.
- The estimated savings passed on to taxpayers by enhancing literature accessibility.
- The number of articles made public by X amount of time.
- The amount of government funding that went into the research being sought.

+5 Victoria Stodden said on December 28, 2009 at 10:24 pm:

We address each of the questions for phase two of OSTP's forum on public access in turn. The answers generally depend on the community involved and (particularly question 7, asking for a cost estimate) on the scale of implementation. Inter-agency coordination is crucial however in (i) providing a centralized repository to access agency-funded research output and (ii) encouraging and/or providing a standardized tagging vocabulary and structure (as discussed further below).
Agency-funded research output will contain at least a peer-reviewed final paper, and if computational, should also contain data and code ensuring that the work is reproducible (the paper, code, and data together are described as the research "compendium"). It is imperative to provide public access to taxpayer-funded scientific output — not only to the final published paper but also the supporting data and code — for the reproducibility and skepticism fundamental to scientific communication and progress.
We address these eight questions in turn:
1. In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
As a general rule publication formats and standards evolve over time as technologies develop and should not be mandated. Any development of research sharing platforms should take into account the evolving nature of standards and formats, and permit this innovation in an open community-driven way. Likely the easiest format for searching, at present, is that of XML; however, as this is not a publishing standard, a more reasonable intermediate goal is that of annotated PDFs and LaTeX comments which can easily be converted into XML given their rich use of structured environments (e.g., tables, figures, and citations). PDF is largely standard for scientific publications today, but is a proprietary format and should not be regulated as a standard. Proprietary formats, particularly those requiring purchase of specific commercial software, should strongly and unambiguously be discouraged by OSTP.
2. Are there existing digital standards for archiving and interoperability to maximize public benefit?
For manuscripts, there are at least two examples of widely-used standards for archiving. The first is the NIH's use of PubMed and PubMedCentral. PubMed is a list of pointers with unique stable IDs (a.k.a. PMIDs) pointing to the peer-reviewed manuscript's citation or, if available, online presence. The second serves as an archive of published, peer-reviewed manuscripts. PubMed couples both to the dynamics of

publishing as well as funding, in that the final requirement the NIH makes of grant recipients is to use the PubMed Central identifier at the end of citations. The use of unique identifiers of papers, as well as of data and code, can encourage the release and hence citation of all forms of research. PubMed also assists in citation by exporting citations in several formats (though, unfortunately, not in BibTeX, the most widely-used format among quantitative and computational scientists). Such a unique identifier would also indicate compliance with agency open access policies.

The second example is http://arXiv.org, which originates from a different set of communities and is used purely for archiving; uploaded manuscripts need not ever be submitted for peer-review. ArXiv entries are given a unique "tag" pointing to the uploaded manuscript. After April 2007, the format was changed to a simple YYMM.NNNN, serving as a date-specific quantitative ID.

Not yet developed is a similar set of IDs for research compendia (defined above as the manuscript, code, and data required for reproducing the work). Tagging of research compendia is an important issue for communicating work, facilitating topical web searches, and aggregating a researcher's contributions, including their data and code. Development of a standard RDFa vocabulary for HTML tags for agency funded research would enable search for data, code, and research as well as facilitating the transmission of licensing information, authorship, and sources. Enabling search by author would allow a more granular understanding of a researcher's contributions, beyond citations. This would provide an incentive to release data and code, and give others — such as funders, award committees, and university hiring and promotion committees — access to a more representative assessment of the researcher's contributions to the community than mere publication-counting. Such a tagging vocabulary could include unique identifiers for data and code, ideally the same as those required for repository deposit as discussed in the previous section, and thus facilitate and encourage their citation.

The leading efforts on these topics include <a href=http://www.datacite.org/<http://www.datacite.org/ and http://www.openarchives.org/ore/. The issue is not restricted to data however; for computational work the entire research compendium must be incorporated into the semantic structure. A recent talk by one of the authors on this issue, proposing HTML+RDFa tagging for research compendia, is available via http://www.stanford.edu/~vcs/talks/CCTechSummitVCS06262009.pdf.

3. How are these anticipated to change?

Technical challenges ahead will be set, as they have for the past decades, by growing sizes of the data files and code bases to be shared. The flexibility of XML (allowing future defined environment tags, for example) has so far kept up with the unpredictable changing demands of users. We anticipate such a mark-up language standard, which includes the possibility of defining new environments, the likely best option for moving forward.

The recent increase in research collaboration and virtual organizations suggests another possible pressure on standards. As scientific research becomes more highly tied to massive computation, for example the NSF's TeraGrid computing infrastructure, research will tend to proceed through virtual environments allowing intensive collaboration by researchers separated geographically. The sharing of code and data in concurrent use is already happening, in addition to the downstream reuse of code and data by subsequent researchers. These virtual environments are developing standards for sharing that could exert pressure on the evolution of formats and protocols for code, data, and manuscript communication.

4. Are there formats that would be especially useful to researchers wishing to combine datasets or other published results published from various papers in order to conduct comparative studies or meta-analyses?

Formats should emerge from the researching communities (as was the case with the Protein Data Bank (PDB), at http://pdb.org), with encouragement toward HTML+RDFa standards for inclusion of meta-data. Careful consideration should be given to the locus of the digital archiving however. The creation of multiple, community-specific or agency-specific repositories does not facilitate interdisciplinary communication and thwarts scripted search and API usage; a national research repository should be established to house released agency funded manuscripts including supporting digital materials such and data and code, and provide links to research housed elsewhere. Many institutions do not have repositories, nor do they have the resources to maintain them. For computational work, supporting data and code must accompany article release creating additional demands on a repository. For papers whose results can be replicated from short scripts and small datasets, many computational scientists who do engage in reproducible research are able to host their research compendia (paper, data, and code) on their institutional web-pages or using hosting resources their institution is willing to provide. These individual contributions, however, may not conform to standardized formats that facilitate scripted search, and nor display transparent versioning and crucial time-stamping of edits and revisions, and may not be labeled with unique object identifiers as required by the NIH Open Access policy. These desiderata could be implemented in a straightforward manner by a neutral third-party site such as one coordinated among multiple funding agencies (as is the case with PDB). Not all computational research involves small amounts of supplemental data and code and an inter-agency repository could host very large datasets or complex bodies of code in cases where institutional support is not available to the researcher. Such a repository could extend the capabilities of http://arXiv.org or PubMed Central for all federally funded research (data, code, and peer-reviewed final manuscripts; perhaps renaming PubMed Central the more representative "PubSci" or "PubCentral"). A centralized repository is especially useful in encouraging researchers to combine datasets and/or code, as opposed to siloing the research by topic area.

5. What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?

There are few in the private sector, in which there are often disincentives to transparency and interoperability. Successes at standardizing the maintaining and submission of code, for example, can be found in the private sector efforts at http://code.google.com, http://sourceforge.net, and http://github.com which are actively used by some academic researchers.

In the academic sector, notable examples to be emulated include http://arXiv.org (for manuscripts) and the Protein Data Base (http://pdb.org ; for protein structure data, one specific data type), which has worked since 1971 to solve the complexities of data sharing as well as the loosely-aligned interests of publishers, scientists, and funding agencies. There are many successful examples of data sharing in academic communities, such as Gary King's Social Science research repository at Harvard, http://TheData.org, or Pat Brown's Stanford MicroArray Database at http://smd.stanford.edu. Note that the MicroArray community publishes their data with every publication as a routinely accepted requirement; similar standards have been enforced in protein structure since the 1990s (cf. http://www.nature.com/nsmb/wilma/v5n3.892130820.html).

Since the data and code are being shared and reused, licensing agreements in these repositories come to the fore. This is an open and active problem across academia largely with the goal of securing attribution rights for owners while permitting use and reuse by others, while minimizing or eliminating licensing incompatibilities between different datasets. Licenses must be compatible for different datasets or different programs to be combined.

6. Should those who access papers be given the opportunity to comment or provide feedback?

Online submission is clearly advantageous for the open and democratic sharing of opinion. However, given the very real consequences (including to future funding, careers, and, in the case of such fields as climate and medicine, policy and political decisions), feedback should be moderated, restricted to verified email addresses, and provided via unique IPs.

7. What are the anticipated costs of maintaining publicly accessible libraries of available papers, and how might various public access business models affect these maintenance costs?
Memory and disk space get cheaper with each year, but such a site requires staffing. The answer to this question, however, depends entirely on the scale of the implementation. What is important to note is the principle of Open Access, and such libraries should be considered valuable stewards of our culture just as the Library of Congress and the National Archives.
8. By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?
As mentioned above, the principle of Open Access recognizes that such collections should be considered valuable stewards of our culture just as the Library of Congress and the National Archives. Rewards to the availability of scientific compendia — papers, data, and code — come not only through views and downloads, but through the acceleration of scientific research, technological development, and an increase in scientific integrity.
Victoria Stodden
Yale Law School, New Haven, CT
Science Commons, Cambridge, MA
http://www.stanford.edu/~vcs
Chris Wiggins
Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY
http://www.columbia.edu/~chw2
References These issues were discussed at a roundtable convened by one of the authors on research sharing issues held at Yale Law School on November 21, 2009. The webpage, along with thought pieces and research materials, is located at
http://www.stanford.edu/~vcs/Conferences/RoundtableNov212209/.

Peter Guttorp said on January 13, 2010 at 5:00 pm:

As far as I know, pdf is no longer a proprietary format, but was released as an open standard in 2008 and agreed upon as an international standard (ISO/IEC 32000-1:2008).

+2 Hope Leman said on December 29, 2009 at 2:36 pm:

I found the comments of Victoria Stodden and Chris Wiggins above very useful and well stated. I do have some concern, though, about this statement, “feedback should be moderated.” A major problem with that is that it often takes several days for comments to appear and that delay has a chilling effect on the free and open exchange of ideas and wastes valuable time, thereby destroying the momentum and excitement that is the very essence of Web 2.0 and Open Access at its best.
For instance, I made my comments in this very forum and instead of being able to tweet immediately that I had made some comments here (and thereby, by tweeting, possibly generating interest in this discussion and comment from others) it took several days for my comments to be approved and to appear here.
There is also the problem that often moderators are anonymous people with agendas of their own or who are representatives of the self-perpetuating elites that the Open Access movement is trying to dislodge from their positions of omnipotence from which they reign with no worries about accountability or transparency, which, again, tends to dampen the freewheeling intellectual interchange that leads to advances in science. Open Access should mean immediacy and 24/7 and not be a dance to the moderator’s when-I-get-around-to-it tune.
Also, I hope that this wonderful Open Access initiative will leverage the promise of Open Science and Open Notebook Science as so compellingly elucidated by Jean-Claude Bradley is his slideshow, “Transparency and Crowdsourcing in Chemistry Using Open Notebook Science”
http://www.slideshare.net/jcbradley/leveraging-transparency-and-crowdsourcing-in-chemistry-using-open-notebook-science
After all, the initiative is all about capitalizing on the release of massive amounts of data by enabling legions of scientists worldwide to employ it fruitfully, innovatively and in a continuous round-the-clock fashion that will see science advance at quantum rates.
And on the question of, “By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?” here is some more edifying discussion of such matters, “ResearchBlogging.org and PLoS work together to measure the impact of journal articles”
http://researchblogging.org/news/?p=724
And in addition to my earlier suggestion of Nature Precedings to answer this question, “What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?” I would add the sterling examples of the PLoS journals, notably PloS Biology:
http://www.plosbiology.org/home.action
and PloS Computational Biology:
http://www.ploscompbiol.org/home.action
as models of user-friendliness vis-à-vis Web design and of scientific gravitas and creativity.

+2 Victoria Stodden said on December 30, 2009 at 10:08 pm:

Dear Hope,
We understand your reservations and appreciate the opportunity to deepen our response. The point we would like to emphasize is that a subtext of the larger Open Access discussion is public access to published research, and this is new. It has traditionally been the case that methods and research results were published for an audience of scientific peers. We both support public availability of research but I think this change to the target audience of scientific communication bears some careful consideration. We feel that comments could be left unmoderated in general (with appropriate login and encouragement of lack of anonymity) but then watched to see if some kind of moderation becomes necessary. One point raised during the Climategate controversy was an unwillingness to share data and code due to researchers’ sense that they would be exposed to too many comments that were not made in the spirit of scientific questioning, possibly for the purpose of distracting and stalling their research.

We believe this is not a compelling reason not to release research compendia (published paper with supporting data and code) but may be a compelling reason not to rule out the possibility of some moderating structure on comments. Unlike a comment on a typical blog post or a youtube video, a comment on research can have deep consequences for people's careers, for the flow of funding, and for changing policy, political, or approval of medicines or medical procedures.
Also relevant is that Nature experimented with opening manuscripts for public comment and peer-review in 2006, and received very little engagement either from scientists or the wider public, resulting in the termination of the trial. See http://www.nature.com/nature/peerreview/debate/nature05535.html. It may be that leaving the work open for public comment is no problem at all, and only garners the potential benefit of more minds engaged in science.

+1 Hope Leman said on December 31, 2009 at 5:55 pm:

Dear Victoria:
Thank you so much for your comments above. I appreciate your trouble and the further elucidation of your important points.
I still, though, think that your comment here merely make the case for how imperative it is that there be as little moderation of scientific discussion as possible, "…Unlike a comment on a typical blog post or a youtube video, a comment on research can have deep consequences for people's careers, for the flow of funding, and for changing policy, political, or approval of medicines or medical procedures." Precisely so. That is why I am against too much gate-keeping. We can trust the public, whom we really do need to engage in scientific research, and make as many people as possible feel that they have a genuine stake in what is happening. The rise of the Taxpayer Access movement is evidence of the peril of ignoring their wishes: http://www.taxpayeraccess.org/
and of the desire of the public for a greater say in what is done with their tax monies. The NIH Director's Council of Public Representatives is an example of the response to this need: http://copr.nih.gov/
I think of people, too, like Augie Nieto:
http://www.augiesquest.org/AugieAndLynne.html
and the Heywood brothers of PatientsLikeMe
http://www.patientslikeme.com/
They are not scientists but have done a huge amount in a very short amount of time to advance research on ALS and other illnesses. I think of people like E-Patient Dave http://patientdave.blogspot.com/
and the work he has done in alliance with physicians like Ted Eytan:
http://www.tedeytan.com/
I bridle at the idea of the possibility of such people possible being excluded from decision-making. As E-Patient Dave wrote, "One of the most fundamental rights is the right to be fully engaged in one's well-being, especially in moments of crisis when a life is at stake." Moderators militate against that right simply by wasting time.
And the Nature experiment took place in 2006. We are now almost in 2010 and people are much more comfortable now with social networking, tweeting, mobile technologies and 24/7 scrutiny of government and public institutions than they were even as recently as 2006.
Climategate merely proves that it is going to be increasingly impossible to exclude large numbers of people from important matters of public policy science and that attempts to do so, when they almost invariably come to light, only reinforce feelings of anti-elitism among the masses. A sort of, "See, all those eggheads were manipulating information and attempting to wreck the careers of anyone who questioned their shibboleths. Told you so!"
Think of past cases of fraud (Cyril Burt) or evil (Tuskegee Syphilis Study, lobotomies,) that might have been detected earlier if there had been scrutiny by those outside the scientific establishment or in other disciplines who would have sounded alarm bells. Sometimes research should be stopped, after all.
The process of determining who the moderators are would only perpetuate structures that have not been sufficiently open in the past to voices from people who don't happen to be Ivy League-affiliated or who live, in say, Oregon, etc. I mean, here you and I are engaging in cordial discussion and I am a mere nobody from the provinces. That is the power of Web 2.0 and Open Access. Let us embrace it.

+1 Alexander Howard said on December 29, 2009 at 3:00 pm:

Like Hope Leman, I found the comments of Stodden and Wiggins to be on point. XML and similar flexible machine-readable formats are desirable, and I have little to add to that statement than an affirmation and vote, other than to similarly endorse http://code.google.com, http://sourceforge.net, and http://github.com as useful models. The open source software movement's success and failures should offer a useful model.
I believe those who access papers should be provided with the opportunity and means to comment, provided such feedback is transparent and moderated.
Metrics like number of articles listed or visitors recorded have utility but may not capture network effects. Inbound links to articles and other citations might be relevant there.

+2 David Skurnik said on December 29, 2009 at 5:54 pm:

The populating of STM materials in the NIH PubMed Central database has been extremely successful. Logically, their model should be used as a starting point and afterwards, if necessary, can be tweaked.
The entire HW/SW technological platform they have implemented to store, retrieve and publish materials can be made available to other agencies. There is no need to re-invent the wheel and expend taxpayer money building something different. They have already transferred their technology to other countries for the establishment of the country specific PubMed Central - so it is surely transferable to other agencies.
PubMed receives materials from the Authors/Publishers/3rd Party Vendors etc… in either XML specific to the NLM DTD, or they receive materials in PDF Normal and Word. Those materials that are not in NLM based XML are converted to NLM XML. The conversion is funded by NIH. The cost of conversion should not be overlooked when budgeting for this project.

Similar to the PubMed implementation, although an agency should accept other formats, XML should be the only format stored in the repository since it will enable the complex searching needed for researching the materials.
I also recommend that since NIH has been so successful in their Open Access Initiative, they should be given the responsibility to coordinate and possibly implement the expanded multi-agency initiative. This will dramatically cut the implementation cost and time and ensure a greater chance of success.
David Skurnik
Vice President
Data Conversion Laboratory, Inc.

+2 Arthur Smith said on December 29, 2009 at 11:30 pm:

One of the banes of scholarly publishing is the plethora of document format standards for text, images, and other content, as well as for metadata, that all essentially provide the same functionality. That makes interoperability of the various respositories much more complicated, with computationally intensive, lossy, and failure-prone format conversions requiring significant investments of human effort to do much useful with all that textual and other data. By setting at least a preferred standard, a federal repository requirement could help bring about a convergence and simplification that would in itself be of great benefit.
Various other comments here have mentioned XML as a natural standard - and that makes sense, but just requiring an XML format is not sufficient; more or less standard document structure components should also be specified. Open Document Format (ODF) is one option, already widely supported in word processing software. SVG is a natural format for most images (an XML flavor that handles vector graphics). But the best solution may be to use what seems to be becoming the standard for electronic books: “EPUB”:
http://www.idpf.org/
One important consideration should be accessibility to the visually impaired - DAISY has a close association with EPUB, though I’m not entirely clear on the relationship (EPUB seems to include some DAISY specifications) - more info on that here:
http://www.daisy.org/
As far as metadata goes, important standards are DOI (http://doi.org/) for citations to published literature, and I’d suggest standardizing on the relatively new ISNI for author-identification:
http://www.isni.org/
This may encompass geographic identification, or you may want to specify national and sub-national components of origin separately. Subject classification is another important area - NLM (PubMed) has standard keywords for their subject areas, and perhaps that should be combined with keyword classification schemes from the physical sciences, economics, etc. to produce a sufficiently broad and flexible system. Or that may be something for independent entities like the existing secondary publishers to work on as added value they can provide.
In general, federal standard-setting in scholarly publishing would be a very good thing, and I hope you will consider some of these as recommendations for the planned repository(ies).

+1 David Skurnik said on December 30, 2009 at 11:58 am:

Considering that the materials are reasearch related, in most cases, they will have a common structure. Therefore, the most logical place to start for an XML standard is the STM based NLM XML DTD currently being used by PubMed Central. Since not all materials will fit the DTD, there will have to be additional work to modify the DTD to ensure that the “non-standard” materials can be tagged.
Once the data is in NLM XML, other flavors of XML like NIMAS (DAISY) and e-Pub can be derived via an XSLT script. The most important thing though is that all the agencies utilize the same process and technology so that we reduce cost, implementation time, increase data interoperability and enhance the users experience - since they would only have to learn how to use one interface to access the data.

James Pepper said on January 13, 2010 at 1:50 am:

Content has to be laid out properly so the blind can access the content. The remediation process to make content accessible today is a very time consuming effort and it only makes content partially accessible to specific software that is very expensive and only the rich blind can afford this content. The public standard for accessibility requires the blind to buy software that costs $1200 and that doesn’t include the computer or other devices. Since the blind were 71% unemployed in 2008, their best year ever, it is extremely callous of everyone to think that this is in any way equality.
I developed a means of making interactive PDF files accessible to the blind for which the blind can use free screen readers to access content in many languages (including English and Spanish) and this process was tested by Darren Burton at AFB TECH, the technology division of the American Foundation for the Blind and he called it a “Raising the Floor” Technology because I can make free text to speech engines read and interact with all the content on the page. My process is currently being tested at the Jernigan Institute of the National Federation of the Blind. My process makes content accessible as it is made so there is no need to fix things after the fact. And because the software to read this is free, the blind do not have a blindness tax, they do not have to pay any more than a sighted person for content.
My experiences with the government and private industry to develop this technology are below and you need to understand these problems to build policy properly.
Last year I made the National Voter Registration form to be accessible to the blind using a conventional technology which is different than my current process. I had the form tested by the American Foundation for the Blind, the National Federation of the Blind and Jim Dickson the Vice President of the American Association of People with Disabilities personally presented the forms to the Elections Assistance Commission to try to get it adopted for the 2008 National Election. The EAC decided to ignore the experts and they had their webmaster come up with a form that violated Section 508 regulations.
It should be noted that the EAC was commissioned by Congress to make voting accessible to the blind and they ignored this with their voter registration form until I complained about it in August of 2008. That is when I contacted the Voting Rights Division of the ACLU and they monitored my correspondence with the EAC. Many Civil Rights groups in Washington were aware of my efforts!
So millions of american were denied the right to vote because they couldn’t fill out the voter registration form.

The blind were required to draw a map of where they lived on the back of the forms by locating an X on the back of the form to label it with the nearest cross streets and then they had to draw in their homes relative to those cross streets and draw in local schools and police stations. Think about this for a moment, you are blind, how are you going to find that X let alone draw a map. This is a literacy test under the Voting Rights Act of 1965!
The forms were not accessible, the blind could not read the form and so that is another literacy test.
The forms required the blind to get assistance to fill them out and that is a Poll Tax.
So I rebuilt the forms so the blind could fill them out like anyone else, and mail them back to their state. I demonstrated it in English but I can do this in most languages and I have contacted SIL International to work on translation efforts worldwide.

And before you say, they could vote because the government spent billions on the elections, yes they could vote if they were registered, because the EAC made sure that the polling places were made accessible to the disabled. And the EAC cited that they have increased the number of disabled to vote since the year 2000 by 4 million people in the 2008 election. But the blind and disabled number around 80 million in the US!
This brings up the problem of how many people are blind in the US. According to the Census there are around 1 million people who are sensory impaired, including the deaf and that figure is used to set policy in the US. It is a survey and not an actual census taking. But the CDC is currently tracking 33.5 million americans over the age of 40 who have been diagnosed with the 6 major eye diseases that lead to blindness. 18 million of them are cataracts, the rest are macular degeneration and the diseases caused by Diabetes, etc. So these people are so impaired that they went to the doctor to find out what was going wrong with their sight!
So if you want to know who is disabled I highly recommend that the census ask people their disabilities. That the President orders the Census to measure disability in the US as a matter of preventing the discrimination of the blind and the disabled under the Voting Rights ACt of 1965, which relates to disability under the Rehabilitation Act of 1973.
The CDC figure shows that accessibility is larger than you think and that we are talking about a demographic that is twice the size of African Americans in the United States. And yet what we get from the government is literally … silence.
I developed this system for accessibility because I had lost my vision and then got it back and so I know what it is to be blind and I found the solutions because of that experience. A blind person cannot find the problems and a sighted person does not have the experience or drive to fix the problem!
When I approached software companies they have a policy of non-confidentiality where if you are an ordinary person and you want to sell them your ideas you have to give away the software so they can decide if they want to use it. So this is a barrier to entry, and this is the common practice so that ordinary people cannot compete against this type of intimidation. I am just one person, why should I give away my work when they will use it to make a fortune? Isn't that against the Clayton Antitrust Act?
I can make every government form accessible to the blind using free software to access this content. Of course I cannot do it myself it is too big but I can teach people to do this work. This process is fast and the content only has to be made once and is accessible to everyone. This process can be applied across platforms but since I am only one person everyone I cannot compete.
When you decide policy for access to ideas and innovations just remember that there are a lot of people out there with good ideas but if they don't belong to a government agency or an institution or are in academia, those ideas will be lost.

+2 Heather Morrison said on December 30, 2009 at 1:53 am:

Q: In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
A: XML is the best format. It is important to also take into account how the researchers work; the process of submission should ideally fit into their workflow. Microsoft has been working on an automated upload feature for repositories. Ideally, researchers should be able to cross-deposit to as many open access archives as are desirable for their work (I already have 3 archives myself, and there are good reason to deposit in all of them).
Q: Are there existing digital standards for archiving and interoperability to maximize public benefit?
A:
- The Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) is key to harvesting and cross-searching metadata from all open access archives.
- Stable URLs, preferably ones that meet the standards for OpenURL (and possibly DOI), are essential.
- The SWORD protocol allows for cross-deposit into multiple archives.
- Creative Commons licensing, to facilitate both human and machine reading of licensing terms.
- For archiving (preservation): LOCKS, CLOCKSS, and Portico. For preservation purposes as well as ensured ongoing access, multiple mirror sites is recommended.
- Open standards are recommended. For example, video materials should use a format like MPEG-4. Open standards will allow the most possible people to access the materials, and will facilitate the task of preservation.
Q: How are these anticipated to change?
A: OAI-PMH is quite stable. SWORD is new; the ability to cross-deposit is very important to researchers, so watch for growth.
Q: What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?
- E-LIS, the Open Archive for Library and Information Studies, has exceptional tools for searching, including a custom-designed subject classification scheme – not surprising for a tool developed by and for librarians: http://eprints.rclis.org/
- Google provides a very effective search engine to materials in repositories, particularly for known items. Google strikes me as more effective in this instance than Google Scholar.
- BASE, the Bielefeld Academic Search Engine, aims to be the world's most comprehensive search service for open archives, using OAI-PMH: http://www.base-search.net/
- It is worthwhile looking at initiatives that are using the same standards for journals, conferences, and archives, providing a foundation for cross-searching materials in all these venues. For example, the Directory of Open Access Journals (DOAJ) http://www.doaj.org features an article-level search, based on OAI-PMH. Open Journal Systems (OJS), a free open source software, also supports OAI-PMH and there is a PKP harvester. http://pkp.sfu.ca/?q=ojs OJS is part of the Public Knowledge Project, which also includes Open Conference Systems and Open Monograph Systems (in development, to be released this February).
Q: Should those who access papers be given the opportunity to comment or provide feedback?

A: Of course; the only questions are the best venues for providing comments or feedback. My perspective is that opening up access to these papers has tremendous potential to inform public debates and commenting on a wide variety of issues; this potential will come to fruition over a period of time, as there will need to be time for learning and exploration. The most fruitful discussions, in my opinion, will be when people take ideas from the papers and bring them to their communities for discussion.
For example, it makes sense to me that a patient advocacy group might lead a discussion on research in their advocacy area, perhaps on their own website, including references to articles of interest. Researchers in this area might well wish to participate in special events with such a group from time to time; this would provide them with feedback in a focused way, and could also be a way for researchers to connect with people who might be good candidates for clinical trials.
Another example: a variety of businesspeople, scientists, and the environmentally minded public might well be interested in research that has the potential to uncover new green technologies.
What would be most helpful to facilitate this kind of discussion would be to ensure that papers have stable URLs so that these communities can reference them, ideally an easy way to export a proper citation, and creative commons licensing to ensure that rights issues are clear (and also to encourage broadest re-use rights; for example, allowing a portion of an article to be posted, with appropriate attribution, to the website of a not-for-profit discussion group).
There can be roles for journalists and media here to act as intermediaries in setting up such discussions, and also for government staff to conduct groups on public policy issues, much like this one.
Q: By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?
A: The first important metric is the number of articles that are freely available. This can involve a simple count of articles, percentage of articles covered under policies that are actually freely accessible, percentage of all scholarly articles published anywhere are freely accessible (an indirect measure of extended policy influence; as an example, hundreds of scholarly journals voluntarily participate fully in PubMedCentral in a way that goes far beyond what is required by the NIH Public Access Policy), and (a little harder) levels of inability to access materials; this may require developing a reporting system.
As for use, number of visitors, abstract views, or article downloads would be useful. It is important to focus on this kind of usage in the aggregate, and not at the individual paper level. There are potentially serious issues with using metrics to evaluate scholarly work, as I have touched on in my book chapter, The Implications of Usage Statistics as an Economic Factor in Scholarly Communication:
http://eprints.rclis.org/4889/

+1 Hope Leman said on January 1, 2010 at 3:57 am:

Heather Morrison makes valuable points here, "…opening up access to these papers has tremendous potential to inform public debates and commenting on a wide variety of issues; this potential will come to fruition over a period of time, as there will need to be time for learning and exploration. The most fruitful discussions, in my opinion, will be when people take ideas from the papers and bring them to their communities for discussion.
For example, it makes sense to me that a patient advocacy group might lead a discussion on research in their advocacy area, perhaps on their own website, including references to articles of interest. Researchers in this area might well wish to participate in special events with such a group from time to time; this would provide them with feedback in a focused way, and could also be a way for researchers to connect with people who might be good candidates for clinical trials."
Examples of how researchers are connecting with patients in innovative ways via the Web include the ALSUntangled project of Dr. Richard Bedlack:
ALSUntangled (ALSU): A New Global Scientific Effort to Help PALS Evaluate Alternative and Off-Label Therapies
http://www.alsnc.org/news/alsuntangled-alsu-new-global-scientific-effort-help-pals-evaluate-alternative-and-label-therapi
and the online registry to track children with infantile spasms:
http://record.wustl.edu/news/page/normal/14702.html
developed by Alexander Paciorkowski, M.D and his colleagues. Dr, Paciorkowski is reaching out to online communities of parents of such children. Both Dr. Paciorkowski and Dr. Bedlack are harnessing the Web in just the sorts of way that Heather Morrison outlines.
Kudos to leaders of the Open Access movement like Heather Morrison and Heather Joseph of SPARC http://www.arl.org/sparc/ for their tireless, selfless efforts in these matters. I spent a great deal of time in the ALS room of PatientsLikeMe and can attest to the value to ALS patients and their caregivers of the kind of discussions Morrison envisions. Such forums already exist and will be greatly enhanced by the OSTP initiative.

+1 Dan Lee said on December 30, 2009 at 3:28 pm:

FORMAT: Acceptable formats for submissions should be those that are in common use by the relevant researchers. Submission should be as easy as possible so researchers can spend more time at what they do best and less time on what is largely seen as bureaucracy.
What is more important is what format papers and data are presented in. Papers in particular should be converted to a standard mark-up language that supports full-text searching, linking, and text mining. Assuming multiple archives will be developed, I would also encourage the adoption of a common DTD across all collections. Anything that will improve ease of use and information sharing will be welcomed by the user community. The NLM DTD is a useful standard and a reasonable place to start.
STANDARDS: Others have addressed this well. OA-PMH and OAI-ORE are working standards used in a range of large related databases. As suggested above, XML and the NLM DTD are useful presentation standards.
CHANGE: Standards will evolve as technologies are developed and community demands accordingly change. This should be expected and understood as archives and policies are developed.
FORMATS FOR COMBINING DATASETS: As above, whatever formats are developed or accepted should be open to allow for re-use.
USABILITY: Each commercial article database requires users to adapt to their idiosyncrasies. The seeming simplicity of Google and Google Scholar, with accompanying advances, field based searching is, perhaps, a target, though it may be unrealistic.
FEEDBACK: Although not necessary, this would be a welcome feature. It would give the public an opportunity to interact with the research they invested in. PLoS One is a good example of research articles openly available where conversations have developed around some papers.
On the other hand, interactivity may require monitoring to weed out spam and truly distasteful comments, which adds to the cost.

COSTS: As noted earlier, NIH spends $4.5 million per year (a minuscule portion of its research budget) and receives 80,000 articles per year. Given the anticipated scale for archiving papers from other agencies, this is likely the best cost estimate for a related operation.
METRICS: There are at least two kinds of metrics worth addressing: compliance metrics and impact metrics. Compliance rates should be tracked to ensure the public receives the full benefit of our investment.
There are a number of impact measures for papers already in use and cited in earlier comments. But the impact of the archive broadly should also be tracked. How are the papers being used? What is the ROI for the public? What new research or commercial developments or commercial can be attributed to the wide availability of research findings?
As downloads, views, and visitors are counted, it is also important to note both the desire for various levels of granularity as well as the need for privacy. It would be of interest, for example, to know the ratio of visitors from .com and .edu sites. But tracking individual IP addresses would be unnecessarily obtrusive.

+1 Hope Leman said on January 1, 2010 at 4:44 am:

Dan Lee is quite right when he says, “Compliance rates should be tracked to ensure the public receives the full benefit of our investment.” The public has the right to know which institutions are flouting the rules and grant funding should be allotted (or withdrawn) accordingly.

+2 Peter Jerram said on December 31, 2009 at 12:43 pm:

In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
In order to maximize the potential of scholarly literature in terms of discoverability and reuse, the full text should be made publicly available in XML format which is the most widely adopted standard within STM publishing, and allows value and interlinking to be added post-publication. In addition to the XML files (which are machine readable files), publishers should also be required to deposit a ‘human readable’ version (such as HTML or PDF versions).
All of the content that is published by PLoS (and other open access publishers such as BioMed Central and Hindawi) is freely and publicly accessible on the day that it is published, and the full text is permanently archived in XML and PDF formats in National Library of Medicine’s public archive, PubMed Central. For publishers that do not deposit full text at PMC, the NIH public access policy requires that at least the accepted author version be deposited by authors themselves for example in Word format. PMC has developed workflows to produce XML, HTML and PDF versions from the author-supplied versions. These processes could be emulated by other funder repositories.
Are there existing digital standards for archiving and interoperability to maximize public benefit?
Within biomedical publishing, PubMed Central (and its mirror sites) is established as the standard public archive for full text content. Similar archives could be created for other fields, such as physical sciences.
The National Library of Medicine DTD is also emerging as a standard format for structuring the XML of scholarly articles. For more information on the NLM DTD see http://dtd.nlm.nih.gov/publishing/ and http://www.inera.com/nlmresources.shtml. The articles within PubMed Central are all made publicly available using this standard. If other repositories are developed for federal funders, it would beneficial if the NLM DTD was also used (or expanded for this purpose).
The Creative Commons licenses represent another standard that is helping to ensure that scholarly content can be reused without restriction and to maximum effect. Open access publishers typically use the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.5/), which allows all reuse, commercial and non-commercial, subject only to the requirement that the work and its authors must be properly credited and cited in any derivative work. Public access policies should require that scholarly articles are made available under the terms of the Creative Commons Attribution License, after an embargo period if necessary.
Are there formats that would be especially useful to researchers wishing to combine datasets or other published results published from various papers in order to conduct comparative studies or meta-analyses?
As mentioned above, literature should be made available under terms (such as Creative Commons Attribution License) that allow for maximum reuse. This will allow and encourage new resources and tools to be developed for navigation, organization and mining of text, for example as described by Bourne et al. (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2275780/).
An illustration of the value that can be added to literature post-publication, and the way that open access articles can be integrated with relevant database entries was recently provided by Shotton et al, who semantically enriched an article published in PLoS Neglected Tropical Diseases - http://imageweb.zoo.ox.ac.uk/pub/2008/plospaper/latest/. See also http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000361 for a description of this work. The work by Shotton et al demonstrates the kind of value that can be added to make literature more powerful, but tools now need to be developed to allow such value to be provided in an automated fashion. The public availability of all literature resulting from federal funding under terms that allow liberal reuse will provide the critical mass of content necessary to drive the development of such tools and services.
What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?
Although it is a public sector initiative, the best example of how value can be added to literature post-publication is provided by PubMed Central. Open access content within PMC is linked and integrated with relevant entries in the wealth of biological databases also hosted by the National Center for Biotechnology Information. These databases include DNA sequence data, human genetic disease and gene expression data, and the links that are added to the open access articles in PMC allow users to navigate seamlessly from literature to data and back again. Open access content from PMC has also been integrated with data outside of the NCBI databases – for example the Protein Databank now includes automatically generated links from protein structures to relevant articles in PMC (http://www.rcsb.org/pdb/general_information/news_publications/newsletters/2009q2/query.html).
Within the private sector, Mendeley is a relatively new service that allows users to organize, share and explore research bibliographies. Although only around a year old, 8 million articles have already been uploaded to the database (http://www.mendeley.com/blog/progress-update/100000-users-and-8-million-articles/), which is adding significant value to research literature. Related services include Zotero, CiteULike and Connotea. Google Scholar is also becoming a more widely used and valuable resource. This service shows the potential for providing integration of searching and navigation (via citation linking) of the scholarly literature. However, all of these initiatives are limited because of restrictions on access and reuse. If the underlying research literature were all open access, this would drive broader participation and the development of more powerful services on top of the literature.

SciVee (http://www.scivee.tv/) is another private sector initiative that demonstrates how open access literature can be reused to create an alternative 'view' of a particular article – in this case an author can create a video presentation to describe the work, and anchor the presentation to the relevant sections of the article (eg http://www.scivee.tv/node/53). Such approaches can help to bring literature to life for new audiences, but can only be achieved when the content is open access, such that the content can be reused and repurposed without restriction.
Should those who access papers be given the opportunity to comment or provide feedback?
Discussion and commentary about articles happens in many different locations (around water coolers; in journal clubs; on twitter; in conferences etc) and the vast majority of that valuable insight is invisible to the reader of the article. As a result, much time is wasted by readers who are unable to learn from people who have considered the article before them. If these discussions were archived and linked to the article, there could be substantial time savings for researchers and gains in research efficiency.
PLoS is one of the organizations that is pioneering the use of commenting, annotation and rating of scholarly articles. Introduced initially on PLoS ONE, all of the PLoS journals now allow users to comment and rate articles. Although the uptake of commenting activity has been modest so far, we feel that post-publication commentary is a valuable and important development for scholarly communication that is likely to gain in significance in coming years. One of the challenges will be to develop appropriate standards that allow relevant community discussion to be integrated, so that users can benefit from all discussions wherever they are taking place. PLoS articles for example now incorporate automatic links to useful blog discussions - see http://everyone.plos.org/2009/12/17/new-addition-to-article-level-metrics-blog-posts-from-researchblogging-org-2/.
By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?
The level of compliance to any public access policy can be estimated by the percentage of articles that have been deposited according to that policy (as compared to all published articles).
Then, usage metrics would be the most obvious metric that could be used to evaluate the impact of the content that is being made available via federal agency public access policies. Usage metrics can be considered at several different levels (ranging from simple page views, to number of unique visitors, number of returning visitors, time spent on article and so forth) and several publishers and academics are actively experimenting with this range (see, for example, the Article Impact Analytics provided by the 'Frontiers' series of journals - http://frontiersin.org/)
It is also worth noting that the federal government is already working to develop metrics that could be useful here. Specifically, Julia Lane, who is working on the STAR program (Science and Technology in America's Recovery http://nrc59.nas.edu/star_info_background.cfm) should be consulted in this respect.
From a user perspective, it is also very helpful to know about the impact that individual articles are having in their respective communities. In 2009, PLoS developed tools, which we refer to as 'article-level metrics' that summarize online usage, citation statistics (from Scopus, PubMed Central and CrossRef), blogosphere coverage and social bookmarking activity. For further information see http://article-level-metrics.plos.org/.
Recently, an initiative has been funded in the UK to develop standards around usage metrics for scholarly articles(http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-index.php.
Peter Jerram
Chief Executive Officer, Public Library of Science.

+1 J. Alex Speer said on December 31, 2009 at 12:47 pm:

Policy Forum on Public Access to Federally Funded Research: Features and Technology
The short comment period does not allow for an official statement by Mineralogical Society of America (MSA). What I write are my own, not those of my employer.
The questions about Features and Technology contain a number of presuppositions, but comments on the questions for phase two of OSTP's forum on public access:
[A] In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
[B] Are there existing digital standards for archiving and interoperability to maximize public benefit?
[C] How are these anticipated to change?
[E] What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?
The initial premise of a depository was for a researcher (or his or her institution) to post their government-supported research results for public access. The unstated assumption appeared to be to accept such reports in any format. However, there has been expectation creep to a desire for the peer-reviewed publication to be posted on a feature-laden, state of the art electronic publishing site to match the usability of the most exceptional private sector sites. Perhaps this comment period is to help to decide if the nature of the depository is to be redefined in this way. If that is the case, the decision will need to consider aspects and consequences beyond features and technology. Features and technology are the easiest parts.
In order to facilitate compliance, you ought to accept whatever format is currently acceptable for author-submitted manuscripts in the discipline field. Perhaps a minimal requirement would be that what is submitted is electronically searchable. This would rule out page image-only files, and make logical an electronic site.
After submission, the challenge of dealing with what has been submitted will be over time as technology and new requirements evolve or arise. Formats, software, and hardware will change. The government agency(ies) or institutions supporting or hosting the envisioned repository will have to assume the responsibility to migrate previously posted files to new formats and systems as well as dealing with the appearance of new or evolving regulatory requirements such as Section 508.
[D] Are there formats that would be especially useful to researchers wishing to combine datasets or other published results published from various papers in order to conduct comparative studies or meta-analyses?
MSA has offered its authors the ability to deposit supplemental data since the 1960's. These are our "deposit items" and they contain data referenced in the published papers. Currently these "deposit items" are posted on the MSA website with the paper and are freely available. For pre-2000 deposit items, we will scan and post those as they are newly requested. We will charge if someone wants a paper copied mailed to them.
Two observations from our experience:
The use of "deposit items" by authors is much less than one would imagine or hope. I surmise two reasons. First, it is much easier to expect that everyone else will make their data available to you then to prepare your own data for others. Secondly, and perhaps more importantly, authors find it difficult to imagine what of their data others might find useful in the future.

The datasets deposited over the last 50 years were collected in all sorts of formats illustrating a veritable history of platforms and storage media. No doubt each format and set of cards, tape cartridge, magnetic tape, floppy disks, etc. was considered state of the art and perfectly suitable for comparative studies or meta-analyses - at the time. For the first 40 years all deposited items were submitted to MSA on paper. In looking at these I would say that if they had been submitted electronically, all would still be useable today if they had been submitted as text files. The task for a data depository is archiving. Future researchers can assume responsibility for formatting the data as input for their own studies.
Authors should have guidance or best practices as to what data of theirs should be made available, the preparation task ought not be onerous, and the format should be simple with an eye to use over the long term.
[F] Should those who access papers be given the opportunity to comment or provide feedback?
The capability is seductive. MSA has a list serve where, among other things, people can ask questions. I am often surprised, but pleased at the substantive feedback. At meetings, often the most important part of a talk or presentation is the question and answer period afterwards. That said, if the proposed depository is to be one of peer-reviewed science and technology literature, it should be kept to that vision. If someone wishes to respond to a published paper, they ought to submit a comment for peer-reviewed publication.
This approach would avoid the necessity of federal government becoming editors, seeking peer-reviewers for comments and replies, deciding whether or not a comment is reasonable, relevant, commercial, or advocates unscientific methods or theories, and assuming legal liability for comments they allow to be posted. On a more human scale, the approach will avoid posting of impetuous comments characteristic of today's electronic world that their authors come to regret.
The governmental purpose of the depository is to make government-supported, peer-reviewed research available to tax payers. It will quickly loose its way and impact if it becomes an interactive social networking site or a portal.
[G] What are the anticipated costs of maintaining publicly accessible libraries of available papers, and how might various public access business models affect these maintenance costs?
The cost will depend on the type and scale of any implementation. At the basic end, the costs might, with much luck, be those mentioned thus far by others. With a feature-laden site, the expectation should be that the costs would be higher. However, the costs are likely to be much higher. Much of the research in the US is government-supported in some form and the depository will likely contain the majority of US research. It is hard to imagine that a free, feature-laden depository site would not eventually displace most other publishing sites. At that point the depository will have evolved into a single-payer, single-publisher system, where both roles are filled by the government. The expense will be more than just the narrow costs of hosting and maintaining a publicly accessible electronic library. The expenses of the entire publishing enterprise would need to be covered.
[H] By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?
The importance of the depository would be in its use, not its holdings. Use meaning either the number of views, downloads, or citations. Views or downloads might be distorted by spiders and bots indexing the site. Someone might be curious to survey users to learn more about them, but given privacy concerns, you will be limited in obtaining such data.
J. Alex Speer
Mineralogical Society of America
[[The Mineralogical Society of America (MSA) is a relatively small, non-profit professional scientific society with a world-wide membership and, among other things, publisher of a journal, books, and a magazine. The publications from 2000 onward are available in electronic form by subscription both on our own website and GeoScienceWordl.org. The journal from 1916 through 1999 is freely available as open access in pdf format (page image with OCR text behind) and is is indexed by Google and searchable with site search. MSA books before 2000 have not yet been digitized. MSA attempts to make what it publishes available in print and electronic forms at cost and is well known for its affordable and widely distributed publications.]]

+1 Hope Leman said on January 1, 2010 at 4:33 am:

I have read J. Alex Speer's articulate, fascinating and well-argued comments. I wish to respectfully disagree with Mr. Speer here, "It will quickly loose [sic] its way and impact if it becomes an interactive social networking site or a portal." Pace Mr. Speer, the impact will in fact come from the social networking aspect. We have seen that repositories do not in and of themselves necessarily lead to advances in science. The goal should be the sort of active engagement with scientific matters such as we see in the Life Scientists room or FriendFeed http://friendfeed.com/the-life-scientists
but this time a community with vast amounts of data readily at hand to bolster strong arguments or immediately undercut weak ones and bad science or sloppy thinking.
OSTP has the opportunity to create a vast community of scientists and members of the public with literally life and death stakes in the matters under discussion. What OSTP can create is this: a portal like that of http://worldwidescience.org/ combined with the user-friendly, vibrant social networking power of FriendFeed. That is not trendy–it is just the way science should go–please, let us not any longer stagnate in the way it has always been. Mr. Speer suggests, "If someone wishes to respond to a published paper, they ought to submit a comment for peer-reviewed publication." That takes months and months. ALS patients, for instance, literally don't have the time for that. Let's ramp up science and adopt a faster pace. We will lose nothing in the process–peer review will be just as rigorous and far more comprehensive vis-à-vis geographic and numerical reach but far more quickly accomplished.

+3 David Karger said on December 31, 2009 at 4:14 pm:

I am professor of Computer Science at MIT working actively on questions around information-sharing on the web. I am the most recent program chair for the International Semantic Web Conference (http://iswc2009.semanticweb.org/) which addresses many issues surrounding the effective sharing of data. Nonetheless, I concur with J. Alex Speer's previous contrarian post regarding the scope of the proposed open access archive feature set. The core objective—to make scientific publications openly accessible—can be met much more easily if we exclude the many wonderful but optional requirements and features discussed in previous comments. The comments reflect feature creep from "providing open access" to "creating a national archive". These are not the same thing! It would be a tragedy if rapid progress toward the core objective were derailed by standards negotiations and research around issues whose answers are not yet clear.
I strongly support many of the previous comments' proposals: to upload machine-readable text versions of articles, to require posting of data as well as papers, to develop a standard (ideally based on RDF) for representing metadata about such papers, to standardize document identifiers

around the doi framework, to host public forums for discussion of the content, and so on. All make great sense. But none is necessary to meet the goal of open access.
Instead, I recommend that we race to quickly achieve the following “minimalist” open access framework. First, that the Library of Congress operate a server that will accept, permanently store, mint a new url for and serve on demand any submitted scientific publication. Second, that our government legislate that any government-funded publication be submitted to the archive in an open format (and thus made available to the public) within one week of publication elsewhere.
This proposal intentionally leaves out many “obviously necessary” specifications and features. It would be nice for LoC to offer metadata and machine readable text for articles, but it isn’t necessary: services such as CiteSeer (http://citeseer.ist.psu.edu/) and Google Scholar (http://scholar.google.com/) have reasonably effective solutions for extracting bibliographic metadata and text from the raw articles. What they don’t have is a centralized repository of the contents they are indexing. Similarly, features such as public commentary on submitted articles, moderated or not, could be offered by other entities once they had access to the underlying articles. Indeed, there may be multiple competing indexers and forum systems, and this would be a good thing, fostering innovation. On the other hand, for the government to choose and enforce a particular standard for documents, indexing, forums, or any other service would hinder innovation.
Deciding on a particular format for articles—HTML, PDF, or some other—also seems like a distraction. Different communities have made different choices, and hosts of tools are available for processing documents in all of these formats. As long as the formats are open, there is no significant barrier to making use of them. (Contrary to the assertion of Victoria Stodden, PDF is no-longer a proprietary format http://en.wikipedia.org/wiki/Portable_Document_Format and http://www.digitalpreservation.gov/formats/fdd/fdd000030.shtml).
After listing all these things the government need not do, I’ll emphasize to the two that really demand government intervention: the open access requirement, and the central repository. The former, because no other entity can really override the commercial interests that wish to restrict access. The latter, because the government’s longevity and heft will give other entities confidence that they can build services around the central repository without fear of its vanishing 5 years later.
In summary, I would argue for OSTP to solve the core problem—that many scientific publications simply are not accessible online—without getting tangled up in solving many other interesting and related but non-core problems. If the government simply provides an accessible archive of the content, then other entities will index it, organize it, inter-operate with it, and make it usable. Open access to the raw content is the one thing that the government can and must provide.
I elaborate further in response to the posted questions:
In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them? In whatever format they are submitted for publication to their scientific conference. This is presumably the standard format for that scientific community, making it the most immediately useful for other members of the community accessing that content.
Are there existing digital standards for archiving and interoperability to maximize public benefit It is worth distinguishing between the broad notion of archiving and the narrower goal of open access. Effective archiving requires capture of standardized metadata but open access does not. It is a mistake to assume that the open-access repository must also be an archive.
How are these anticipated to change? They certainly are anticipated to change, and in ways we cannot currently predict. This is another reason for the government to avoid fixing on a current mechanism that could stifle future innovation.
Are there formats that would be especially useful to researchers wishing to combine datasets or other published results published from various papers in order to conduct comparative studies or meta-analyses? Yes, but they differ for each scientific community and should not be enforced from above.
What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional? Arxiv.org is an example of choices made by a single community (such as requirements to submit all papers in LaTeX format) that are very valuable for that community but would be a disaster to many other communities. The open access repository should remain neutral on such choices so that all communities can benefit from it. CiteSeer (http://citeseer.ist.psu.edu/) and Google Scholar (http://scholar.google.com/) are fine examples of how separate entities can offer competing services (metadata extraction and indexing) over a corpus (currently whatever they can spider from the web) that offers no such services itself.
Should those who access papers be given the opportunity to comment or provide feedback? This would be a lovely feature, but is one that again can be offered by separate entities such as SideWiki (http://googleblog.blogspot.com/2009/09/help-and-learn-from-others-as-you.html), ReframeIt (http://reframeit.com/), or Diigo (http://www.diigo.com/).
What are the anticipated costs of maintaining publicly accessible libraries of available papers, and how might various public access business models affect these maintenance costs? These costs are significant but will grow even higher if more and more features and requirements are added to the system. If service is limited to storage and serving of raw content, pricing should be of the same magnitude as services such as Amazon S3, currently listed at roughly $100 per terrabyte stored and $200 per terrabyte served.
By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections? Like any other web site the repository can measure the amount of content served. Longer term, the government could measure the relative impact of open-access versus non-open access content by measuring the amount citation of both kinds of work.

+1 omdwsjzhbursz omdwsjzhbursz said on December 31, 2009 at 10:43 pm:

a clarification: you wrote
requirements to submit all papers in LaTeX format…would be a disaster to many other communities.
according to the arxiv (see http://bit.ly/6oDSOG), arxiv accepts:
“(La)TeX, AMS(La)TeX, PDFLaTeX
DOCX (Word 2007)
PDF
PostScript
HTML with JPEG/PNG/GIF images
Our goal is to store papers in formats which are highly portable and stable over time. Currently, the best choice is TeX/LaTeX, because this open format does not hide information. Note that for this and other reasons we will not accept dvi, PS, or PDF created from TeX/LaTeX source. Users of word processors such as Microsoft Word (versions prior to Word 2007) should save their documents as PDF and submit that.”

+1 Michael Guthrie said on December 31, 2009 at 5:12 pm:

* In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
It would be advantageous if a variety of formats were offered in a repository. The master, so whatever format it was orignally authored in, and preferably this would be in an Open format, such as ODF, and then derivatives that can be viewed by mostly anyone, such as PDF, and OCR'ed at that. An Open dwerivative if the orginal was not, such as ODF, and there should be no reliace on proprietary software or formats that are only readable by proprietary software.
* Are there existing digital standards for archiving and interoperability to maximize public benefit?
I woutld currently archive images in Tiff uncompressed fomrat if an image, or perhaps some would advocate something like RAW for images, or JPEG2000, and if documents then as above, in ODF for openness, and for interoperability a repository should have the various versions of OAI enabled, such as PMH, ORE, METS, and perhaps the latest SWORD initaitve for easy deposit from varios sources. A RESTful API sould also be enabled for easy interoperability
* How are these anticipated to change?
They will get refined, but the basic mechanisms are there for these to be good for interoperability and federation.
* Are there formats that would be especially useful to researchers wishing to combine datasets or other published results published from various papers in order to conduct comparative studies or meta-analyses
* What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?
There are many DSpace repositories that are providing excellent usability. http://www.dspace.org They are interoperable with other repositories. At BioMed Central, we are providing SWORD deposit of items into repositories, providing interoperability on a very large scale.
* Should those who access papers be given the opportunity to comment or provide feedback?
This has become a hot topic in the repsoitory world, and it would seem ineveitable that this type of interaction will occur, so a moderated forum of commenting rather than a complete free for all would be advantageous.
* What are the anticipated costs of maintaining publicly accessible libraries of available papers, and how might various public access business models affect these maintenance costs?
The costs are coming down in regards to storage, and the cost is typical of any enterprise applicaiton, but not as expensive as some, and is very reasonable for the serivcde it provides.
* By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?
Metrics should be evaluated on the basis of the number of times an item's record or page is accessed, and then also how many times the full text is downloaded, whether viewed via PDF or actually downloaded. A new metric may come about as to how much commentary has been generated by an article. And then how much the article is cited. This metric should not be reliant on a third party and is potentially the most important.

Stevan Harnad said on December 31, 2009 at 5:21 pm:

How To Integrate University and Funder Open Access Mandates: Deposit Institutionally, Harvest Centrally. Otherwise funder mandates compete needlessly with institutional mandates, which cover all of research, funded and unfunded, across all disciplines, instead of reinforcing them. No functionality whatsoever is lost by mandating institutional deposit and central harvesting, but an enormous amount of additional OA content is gained.

Stevan Harnad said on December 31, 2009 at 5:24 pm:

Links for above:
http://openaccess.eprints.org/index.php?/archives/369-guid.html
http://bit.ly/8EBE2i

+1 Michael Carroll said on December 31, 2009 at 11:17 pm:

I will be submitting more extensive comments in response to Phase III and the bonus round for commentary. Below, I respond to the questions as posed, including to the implicit assumptions buried within each question.
* In what format should published papers be submitted in order to make them easy to find, retrieve, and search and to make it easy for others to link to them?
This question conflates issues. The general principle should be that the public record of publicly funded research should be freely available over the Internet within a reasonable time after publication in the format considered by the community to be the publication of record, e.g. in PDF format.
However, the format of record may actually impede achievement of fundamental goals of open science, as many other commenters have mentioned, so there should also be a version of papers reporting the results of federally funded research available in flexible, interoperable formats, such as XML, and/or the relevant agencies should have the legal rights necessary to convert any submissions into such formats.
* How are these anticipated to change?
All aspects of scholarly and scientific communication are subject to change. What is to say that the standard format for an "article" is likely to stay stable in light of the potential public availability of the underlying data set(s) reported and interpreted by an article in the future? Consequently, it is important to have legal and technical flexibility to reuse or republish articles in a flexible manner to take account of future developments. Requiring that research papers reporting the results of federally-funded research be available under a Creative Commons license, such as the Creative Commons Attribution-Only License, would ensure that future archivists could maximize the value of such articles in light of available technologies for linking articles, data, and related digital objects without unnecessary intellectual property-related concerns.
* Are there formats that would be especially useful to researchers wishing to combine datasets or other published results published from various papers in order to conduct comparative studies or meta-analyses?

The question presumes technical formats. But the legal terms-of-use associated with any particular article are as important to the utility of an article. Creative Commons licenses offer standardized, liberal terms of use. The CC0 protocol offers owners of data sets a legal means of disclaiming copyright adn related legal interests in databases subject to community normative claims for attribution to foster the goals the question seeks to promote.
* What are the best examples of usability in the private sector (both domestic and international) and what makes them exceptional?
A number of publishers that use the supply-side funding model offer their published articles under a Creative Commons license. For example, the Public Library of Science offers press-friendly summaries of articles published by some of its journals under a CC license, and these summaries often positively influence the accuracy of the press coverage of the results reported in the underlying article.
* Should those who access papers be given the opportunity to comment or provide feedback?
Yes, absolutely. The fundamental lesson of an open Internet is that useful information - wisdom - comes from a dispersed and unpredictable crowd. While some feedback is likely to be useless from a scientific perspective, it is useful for members of the public to have access to publicly funded researchers, and some commentary may be scientifically useful as well.
* By what metrics (e.g. number of articles or visitors) should the Federal government measure success of its public access collections?
I will have more to say in Round 3 on this issue, but it is important to recognize that provision of public goods has ripple effects that are difficult to quantify or measure. So, a metric based solely on unique visitors to a web site is almost sure to undervalue the public benefits of provision of public access to the results of publicly funded research.

Mike Serfas said on January 1, 2010 at 1:05 am:

Several excellent comments have recommended the LaTeX format already, but one more advantage is the BibTeX system for reference management - an alternative to the $100-$200 Endnote software. The public access program may find it useful to fund the development of user-friendly installer tools and manuals to make LaTeX submissions easier, but it should not be in a position to mandate its use.

omdwsjzhbursz omdwsjzhbursz said on January 1, 2010 at 11:28 pm:

does anyone know why nih’s Pubmed supports (exports citations to) proprietary software like endnote but not bibtex?

+1 Howard Goodell said on January 1, 2010 at 2:04 am:

As We May Publish
(by analogy with http://www.theatlantic.com/doc/194507/bush)
I agree with previous posters that open access to scientific and scholarly papers, data and source code resulting from Federally-funded research is a cost-effective way to increase the value of research investments, especially by facilitating research using published data in ways the original authors didn’t expect.
I also agree with poster David Karger that open access should be initiated in a minimalist fashion; so that the future ideal does not become the enemy of substantial present good. It has been said that changing Internet technology is like rebuilding an airplane in flight. Changing the research literature — the journals whose editorial judgment is our criterion of what is valid in science; the yardstick against which our careers are measured — must be done carefully.
With these caveats, I believe we must not stop with open access. Initiating open access to the significant portion of the global research literature funded by American government is a priceless opportunity to shape its future. DARPA’s 1960’s network research program ultimately produced TCP/IP and the global Internet. Support for the development of interchangeable parts in manufacturing begun by President Thomas Jefferson led to American dominance of manufacturing in the 19th century.
Open access to the research literature should be accompanied by a program of R&D whose approach must be evolutionary, but whose ultimate goals are radical: to transforming the research literature from its paper origins to forms that fully use computer technology to increase its effectiveness. I propose four goals, to be achieved in stages:
1. All data, analysis and results will remain permanently available.
2. A “paper” will not be a disconnected artifact, but will maintain the connections between data, analytical process and results in an executable forms that readers can inspect and manipulate. Every reviewer or reader or scientist wishing to build on the research in an online paper will have capabilities comparable to a high-end data visualization and analysis system today. They may change any part: exclude some data or merge in their own; change assumptions or algorithms; zoom and pan graphics through the data. With each operation, the paper’s charts, tables and numbers in text will show the results.
3. They may post these changes; other readers may comment and rate (weighted by their own qualifications); other readers may choose to see comments or not. Comments become a form of publication in their own right that tenure committees, etc. see as part of an author’s corpus.
4. Publications may be processed fully automatically when future knowledge requires it.
Literature needs to become an “active information resource” designed from the start for active exploration by readers and fully-automatic processing. Standard and tool development can make scientists’ time investment to achieve this small, and their payoff will be enormous.
The processes by which we produce and consume research literature have not changed fundamentally for centuries; they can fairly be called pre-industrial.
Like skilled Victorian mechanics using judgement and expertise to hand-fit parts to each individual machine, researchers writing a paper rarely produce a machine-readable algorithm for exactly how they converted experimental data into its tables, graphics and text; so human judgment and expertise and considerable time are required to reproduce them.
Our consumption of scholarly papers is similarly pre-industrial.
Understanding a paper’s concepts cannot be automated, but lots of intellectual busy-work in today’s process should be.
First, consider data analysis. Today, reviewers and readers can only manually spot-check results and analysis. They cannot re-analyze the data with different assumptions or methods or merge it with other results.
Worse, paper-paradigm literature is difficult to process automatically. “Text mining” literature designed for human reading is imprecise. Researchers who want to replicate, meta-analyze or incorporate published results with their own waste person-millenia each year manually

retrieving and making sense of published results other researchers manually generated. Cutting and pasting data from PDF files, corresponding with authors to get datasets and reverse-engineering analysis operations from text descriptions are massive wastes of time. Any industrial planner would call the expensive detour through text formats “unnecessary operations”, like loading and unloading boxes in the same factory. Worse, much large-scale analysis of previous research that could be quite powerful isn’t done because it would be prohibitively expensive.

The best way to evaluate the potential is to imagine that you are reading such an “active information resource”. You will have abilities like those found in sophisticated visualization and analysis tools, far beyond the minimal affordances of an online paper today.

A table or graphic in such a publication is not just manually-entered text or pixels, but an active object permanently connected to the data and analysis process that produced it. Click on any data point to drill down to the raw data and/or algorithmic transformations that produced it. Link forward from raw data to see how it was represented in one or more papers based on it. Many elements of the paper have active controls that let each reader manipulate the view: zooming or panning or even re-analyzing the data with different assumptions. If you select points to remove or pull a slider to change a P-value threshold, or modify an equation term, every other linked element in the paper changes to match.

Elements in the text may also be links to data that change automatically as it is manipulated.

Some logical relationships now expressed only in text would become more formalized. Instead of only choosing keywords to describe an article, authors might classify its concepts in Semantic Web ontologies and encode logical relationships using high-level graphical tools (see below) instead of writing. Readers can choose to see statements as text or a logic diagram; computers can process them unambiguously. Readers can manipulate these relationships, too, and see text highlighted where conclusions were affected by the new assumptions.

Readers should be allowed to comment on an online paper, showing their different analysis or additional data. They can classify their comments by ontology terms, and specify exactly which logical relationships they are disputing or extending. Authors might respond to all comments about one term or logical relationship; all these comments will get a link to the response, and might be color-coded as responded to, and whether the author disputed or agreed with them on that point. Other readers should be allowed to rate comments as they do in this comment process, but weighted by relevant criteria such as their own publications in the field.

Non-frivolous comments are preserved permanently; they could be linked to, analyzed and even cited, becoming a very quick form of publication and part of the author’s corpus. Authors could link corrections and additions and later work to previous papers; readers decide whether to see the updated or original version, and comments with different weights or categories.

Because “papers” will now be full-fledged data objects with a host of detailed properties rather than just files on a server, paper references can be much more than just hyperlinks. For example, you might filter all the references of a paper or group of papers by many properties: impact factor; subject keywords; datasets they are based on on, experimental techniques used, or even the presence of a specific items in the document such as data about carbon-rich Near-Earth Asteroids or lists of genes associated with a certain disease.

If you select some references or other papers of a certain type, you can cluster other references that are similar. You might select one reference cluster and search their references, or find similar papers the author didn’t reference, visualizing results at each stage. Editors might cluster submitted articles with the literature to find appropriate reviewers. Paper referees might cluster it with recent work to help them assess its originality.

Textbooks, theory papers and review articles can not just cite papers, but link them specifically to paragraphs, tables and raw data that support each assertion. So even beginning students can answer questions by probing as deeply into a subject as their interest takes them. This is a productive paradigm to stimulate creativity and independent thinking from the beginning.

Designing literature for automatic processing will make it far more valuable. Reliable automatic integration of data between experiments to update, review, meta-analyze or extend previous work will let us answer many new questions from existing data and target new experiments more precisely. The success of 19th-century industry’s long and difficult quest for interchangeable parts and mass-scale production made goods far cheaper. Effort spent standardizing data formats and carefully categorizing and annotating each published dataset will have a very high return.

Imagine setting hundreds of CPUs loose on the literature of your field, to re-analyze thousands of papers and their supporting data based on new findings! Imagine them generating in hours the material for comprehensive meta-analysis papers that would have tied up your post-docs and students for months, complete with their own active visualizations of which papers’ conclusions were supported or refuted. Such massive automatic processing could speed our work as much as giant power shovels replacing legions of miners with picks and shovels, or blast furnaces and continuous rolling mills out-producing dozens of small, inefficient mills did in the 19th century, or developing the Internet and WWW in the 20th have done for our recent work.

The implementation of open access to publicly-financed research literature is an opportunity to redefine the functions of scientific and scholarly literature. This has more potential to speed progress than twenty breakthroughs in individual fields: to help us solve our hardest challenges; to understand more; to heal our bodies and our planet sooner and more effectively.

Will needing to post papers in these new formats impose major costs and changes in how scientists work? To provide such sophisticated functions today would take a Web programming team, database and Semantic Web experts and inordinate effort from scientists producing the data. However, just as developing the Internet and WWW were public goods that changed the game for millions of people and institutions with content to share, most of this investment only needs to be made once. Funding R&D to develop standards and seed implementations is a public good worthy of public financing.

So are education and other help for commercial and open-source tool writers to support them. So are creating more publicly-financed online repositories like NCBI http://www.ncbi.nlm.nih.gov in biology, and collaborations with existing sites in each scholarly field to offer compatible capabilities.

With appropriate tools, there’s no fundamental reason why it should be harder for a scientist to, for example, link a permanent repository of his data to a pipeline of analysis and charting operations and create an active chart in an online document than it is to build a static chart using a spreadsheet or R script today and paste it into a word-processor document or PDF submitted to a publisher. In fact, they could continue to use many existing tools.

The NCBI’s PubMed Central accepts articles in its Journal Publishing XML DTD, but also accepts several other widely-used XML DTDs http://www.ncbi.nlm.nih.gov/pmc/about/pubinfo.html#techreqs .

Similarly, standards for defining data processing will probably include common programming tools such as R and SQL. Other tools such as proprietary spreadsheets could include tools for exporting data processing and graph generation code in a standard format.

Another point is that the goal of automatic processing and active information resources can be met in many ways. The Internet’s distributed hierarchical structure with local management solved many problems and greatly aided its acceptance. Similarly, standards for producing and accessing scholarly publications may be implemented in many different ways; yet must interoperate. The same data might be stored as XML or relational database or various file formats in different repositories; yet can be retrieved uniformly. Any document or data element may be retrieved by desktop programs, online repositories, or compute clusters using the same interfaces.

Different institutions with different policies must support the same interfaces. For example, if publishers have a 6-month or 12-month embargo, they should still be required to establish constant URIs for the document components and support all the standard interfaces for their subscribers during the embargo period that the whole public will have later. Once the public has access, they need to see the comments and corrections that were added when it was subscription-only.

A few technology “existence proofs”:

A biological example of the value of posting reliably-annotated data is the MIAME (Minimum Information About a Microarray Experiment) http://www.mged.org/miame standard. Once journals demanded that raw data in MIAME format be deposited in open repositories before papers would be accepted, it became common for researchers to compare their microarray data with other experiments. Many new kinds of research became possible, such as meta-analyses of genes present in the datasets but not considered in the original papers.

Readers of Stephen Wolfram’s Mathworld http://mathworld.wolfram.com/ can parameterize graphs of mathematical functions by editing numbers in boxes.

Galaxy http://galaxy.psu.edu/ is an online biological data analysis website that lets biologists analyze data while saving a history of their actions. These histories are executable programs that could be incorporated directly into active online papers. Readers could execute them to reproduce the authors’ analysis, or copy and change them if they had different ideas.

Many browsers now support XML Scalable Vector Graphics (SVG) http://www.w3.org/Graphics/SVG/ diagrams (including IE via Google’s ExplorerCanvas http://excanvas.sourceforge.net/). SVG can easily be resized by client-side script controls and include links on individual graphics elements for drill-down. A Java WebStart(TM) downloaded program or a browser plugin might provide more features or better response speed for serious users.

The W3C’s Semantic Web activity http://www.w3.org/2001/sw/ is an obvious framework, not only to build many specific languages for specific fields but as a general framework for encoding relationships between diverse objects with Uniform Resource Identifiers (URI)s — anything from a data point to a chemical name or a concept of physics to a comment on a paper. The Semantic Web Resource Description Framework (RDF) and the Web Ontology Language (OWL) can encode many logical concepts.

ClearMethods’ http://clearmethods.com/ experimental tool Justify(TM) provides graphical tools to build a tree of logical assertions and denials and supporting facts that can be rendered graphically or as text.