EnablingOpenScholarship
Université de Liège - Rectorat, place du 20-Août, 7 (Bât. A1), B-4000 Liège, Belgium

Public Access Policies for Science and Technology Funding Agencies Across the Federal Government

Response from Enabling Open Scholarship (EOS)

A. Opening statement

Open Access to federally-funded research will advance American science and technology, will speed up research, will reduce duplication, will increase the usage and impact of research, will facilitate interdisciplinary research, will improve the quality of research because greater scrutiny will be possible, will enable the deployment of new semantic technologies to create new knowledge from existing research findings, will provide the wherewithal for better, smarter research assessment and management and will provide greater payoff for the US taxpayer from the funds invested in research across all federal agencies. A number of constituencies will benefit: as well as the research community itself, which will have immediate and untrammelled access to the information it needs to do its work, the professional, practitioner and lay public communities will also benefit. These things will lead to wealth creation, improvement in the quality of life and a better informed populace in an increasingly scientific/technological world. Access must be made available in ways that permit full re-use of research results and through services that maximise ease of use and convenience for the relevant user constituencies. We elaborate on these points in our detailed response below. Our response is organised by answering the questions listed in the *Request For Public Comment*.

B. Enabling Open Scholarship (EOS)

EOS is an organisation of universities and research institutes worldwide whose managers have come together to discuss, shape and promote the principles of open scholarship. EOS has members on six continents, from the largest, broad-based universities and research institutes to some of the smallest, most specialized research-based institutions. As well as universities and research institutes, EOS also has government departments and research councils (analogous to the NSF or NIH) in several different countries as members. The EOS Board has collectively developed and signed this submission: Board members are listed at the foot of the document.

C. EOS responses to the questions in the Request For Public Comment

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?

Authors conduct the research using federal funds and disseminate it mainly via the traditional routes, which are writing articles for journals, writing chapters for books, authoring entire research monographs or contributing papers to refereed (peer reviewed) conferences which are subsequently published. In some disciplines, authors are also disseminating their findings through less formal routes, such as by posting them on blogs and other websites, many of which permit and encourage interactive feedback from the target community (so-called 'Web 2.0' technologies). The use of this latter category of dissemination routes by authors is growing but is not systematic. Locating and accessing information disseminated in this way is therefore subject to the possibility of missing items of potential use because there is not yet a structured way of discovering and retrieving them. Authors are also peer-reviewers for other authors' papers. Peer-review is a non-cash cost in the scholarly communication system, but in cash-equivalence it amounts to around 30% of the cost of publishing and distributing an article (CEPA, 2008).

Libraries and universities have until recently done little to *disseminate* research, having instead traditionally been concerned with *acquisition* of research information rather than its publication. This is changing as the Web provides an effective means for dissemination of research by the institutions that create knowledge. Institutional repositories are being established at the rate of approximately one per day in institutions around the world, usually by their libraries, and these enable research-based institutions to collect together and disseminate the fruits of their research efforts. Universities, through provision of peer review services to publishers and through the payment of subscriptions for journals, meet over 80% of the annual global cost (total annual cost is US$ 10.37 bn) of publishing and distribution of journal articles. Peer review services provided for free to publishers by researchers equate in cash terms to around US$ 3bn per annum: Subscriptions amount to around US$ 5.4 bn per annum. Universities and research institutions also provide the infrastructure for research to be carried out – the laboratories, offices, study areas, libraries, and the administrative machinery to make it all work. And largely, these overhead costs are underwritten by public funds.

Primary publishers provide the management of peer review, editorial services and other value-adding services. They publish research in the form of journals and books, and these products are mostly sold to the scholarly community under an access-restriction model (subscription-based in the case of journals and as a once-paid-for good in the case of books, though electronic books bring with them new business models based upon licensing in many cases). The scholarly publishing industry is dominated by a very few, very large publishers, all of which are for-profit organisations. The profit margins vary somewhat, but Wiley Blackwell reported a profit margin of 13.56% and Reed Elsevier one of 16.55% as company averages for the last year's trading (margins from journal publishing alone are higher than these company averages). We state these figures not because profiteering in a capitalist economy is wrong, but because there is a point to be made about the overall costs of the system to the US taxpayer, of which profit forms a part, and how these might be reduced by a move to Open Access (a point to which we will turn later).

Some publishers use different business models for journal publishing and produce Open Access products (free of charge for users): there are a number of these business models, but they distil down into two major categories – publishing with and without 'front-end' charges. Front-end charges, or article-processing charges (APCs), are fees that cover the publisher's costs and are levied when the article is approved for publication after the peer-review process is completed. They are usually paid from the author's research grant or by the author's institution.

Secondary publishers provide discover-and-retrieve services, bought by libraries on subscription, that allow users to navigate the literature and access the items of use. Mostly, these services point to the paid-for (subscription) version of articles, though sometimes if a free (Open Access) version exists, it is indicated. These services do not have a dissemination role, but they do provide a metadata-aggregation (harvesting) function. Primary publishers give their metadata to secondary publishers because there is a marketing advantage in having their journal contents indexed by the secondary publishers' services.

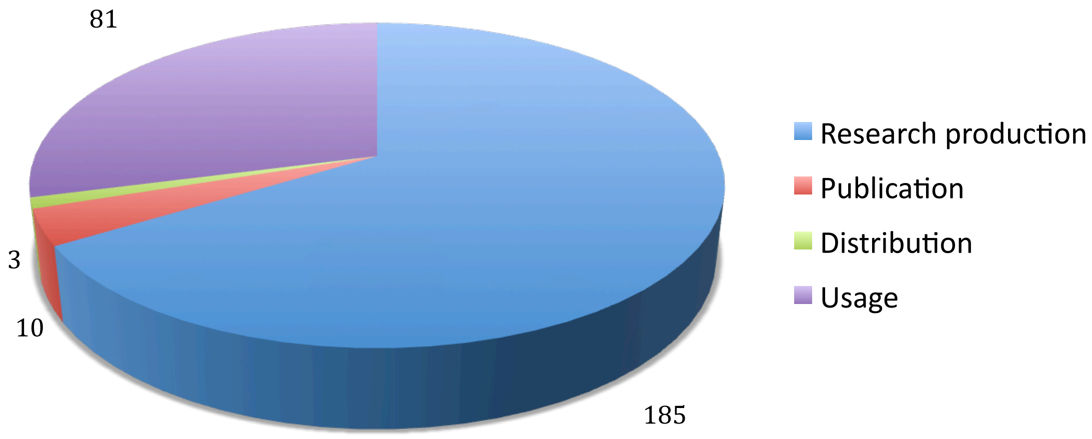
The ***federal government*** funds the research process and, through this funding, in some cases directly finances publishing services by the payment of APCs.

System costs

The costs of the scholarly communication process have been identified for the world (CEPA, 2008) and for the United Kingdom in two studies (CEPA, 2008, Houghton *et al*, 2009a). The UK produces around 6.6% of all research, while the USA produces probably four times this much. Nonetheless, the research system model is similar in both cases and the UK costs therefore probably reflect fairly well, in proportion, those in the US scholarly communication system. Both countries are net contributors to the international peer review system. The global costs breakdown, reproduced from CEPA's study, is as follows and shown in Figure 1:

Research production (activities that create an article):	USD 185 billion
Publication (peer review, editing, etc):	USD 10 billion
Distribution (transport and marketing):	USD 3 billion
Usage (user searching, printing and reading):	USD 80.5 billion

Figure 1: Global scholarly communication system costs, 2008 (USD billions)



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?

The interests of users of research information (researchers, libraries, universities, the federal government, the public) are maximally served by *provision of that research information on an Open Access basis, immediately available after peer review, and with minimal conditions on re-use*. Any public access policy should therefore make this the primary aim.

The interests of primary publishers are predominantly centred upon profit or, in the case of scholarly society publishers, upon creating a 'surplus'. Some exceptions exist, such as certain university presses and societies, but these are a minority. There will thus always be a tension between the main motive of most primary publishers and the public interest in the dissemination of research results. The tension can be resolved by viewing publishers as service providers to the research community, and then structuring a research dissemination system that facilitates open communication while creating a true competitive market for publishing services (peer review management and editorial services). Research information users' need for immediate access to peer-reviewed publications at no charge, and publishers' need for a competitive market for their services, are thus both accommodated. Market forces will determine what added value is 'must have' and what is 'nice to have' and publishers will engineer their value proposition accordingly.

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?

User groups for peer-reviewed publications arising from federally-funded research are as follows:

- Researchers in all disciplines within the academy
- Research managers, for monitoring research progress and direction
- Research funders, for monitoring research progress and direction
- Research communities outside the academy (e.g. R&D-based industries, military research establishments, etc.)
- Professional communities outside of the academy (e.g. medical services, the legal community, accounting professionals, etc.)
- Practitioner communities outside the academy (e.g. civil engineering and energy sectors, architecture and the built-environment sector, extraction/forestry, agriculture etc.)
- Cultural heritage organisations (museums and galleries, in both public and private sectors)
- Freelance and private scholars
- The lay public (interested individuals who can use research information in their daily lives)

The difficulty most frequently expressed by researchers within universities is their inability to access journal articles they have identified as useful because of a subscription barrier (Jubb et al, 2007) and this issue heads the list of top concerns of researchers across all disciplines reported by Swan (2008). The recent report by the UK's Research Information Network concludes that 'many researchers are encountering difficulties in getting access to the content they need and that this is having a significant impact on their research' (RIN, 2009), evidence that even the largest and best-funded universities cannot afford to buy in all the research information that their researchers need, as inter-library lending budgets continue to testify.

So long as journal articles are obtainable only through Toll Access (via subscription or pay-per-view), it will continue to be difficult and costly for would-be users *outside* institutions with well-stocked libraries to find and use all the research material that they need. Into this category of would-be users fall the R&D-based SMBs (small and medium businesses), the bedrock of the knowledge economy and innovation-based industry. Ware (2009) recently reported that 73% of UK SMBs had difficulty in accessing journal articles they need, as do 53% of large companies. The same study found that just 2% of SMEs, 7% of large firms and 17% of higher education-based researchers report having access to all the articles they need for their work. These numbers correlate with those from a much larger study conducted five years previously (CIBER, 2004), indicating that access has not improved in any meaningful way in that half-decade. Perhaps the best way to bring the point home is to quote the managing director of a small biotechnology company, who said:

> *"With a small oncology company … it is imperative that I have access to the literature. But small companies do not have the "deep pockets" necessary... The for-profit journal publishers have effectively barred access to key scientific information except to those who can afford their outrageous fees. Much of the most innovative work is being done at companies like mine that cannot afford to pay $30+ per paper or pay per-search charges in abstracts or journal collections."*[1]

If companies cannot afford access to the journal literature essential for their business, access problems will be all the more hindering for institutions in the less wealthy cultural heritage and practitioner communities, let alone for freelance and private scholars whose numbers are not inconsiderable in the US.

These communities would certainly use journal articles if they were available to them, as would the research community itself. Data from university repositories shows the level of usage that freely-available articles enjoy over an above the usage they get from publisher websites by those who have access through subscriptions. For example, the University of California's eScholarship repository has almost 31,000 items at the time of writing, and since 2002 has seen almost 9 million views from users. Kansas University's ScholarWorks repository has 4,500 items with an average download figure of 99 per item. The physics repository, arXiv[2], receives well over 100,000 hits per day. This is usage that is in addition to usage from the journals themselves and illustrates the latent demand for the fruits of scholarship that is not satisfied through the access-restriction (Toll Access) model that has hitherto dominated the scholarly communication arena.

---

[1] Terence Dolak, SDR Pharmaceuticals, Andover, NJ: Chemistry Central blog, Thursday May 24, 2007: http://blogs.openaccesscentral.com/blogs/ccblog/entry/unemployed_retired_might_lose_touch
[2] www.arxiv.org

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?

To enhance public access to peer-reviewed papers arising from publicly-funded research, federal agencies should develop policies on access along the lines of those that have been developed in other countries. The United Kingdom's seven Research Councils, for example, all have policies on Open Access to research articles and datasets. The Research Foundation Flanders has one, as do the Irish Research Council, the National Health & Medical Research Council in Australia and many others. At continental level, the European Research Council also has a mandatory policy on open access. There are currently 42 mandatory policies on Open Access from research funders and the list of existing policies is given in appendix 1. The policies can be viewed in the Registry of Open Access Policies[3].

Numbers have grown year on year: there were none in existence in 2005, but by the end of 2006 there were 9, by end-2007 there were 23, by end-2008 there were 30 and by end-2009 there were 42. US federal agencies would therefore not be treading new, experimental ground by establishing public access policies: indeed, from our viewpoint, representing the interests of universities and research institutes, we would welcome the swift development and implementation of suitable policies across federal agencies.

Measuring return on investment from public access is a new challenge to society. Benefit to research – and thus to society – can have a payoff far into the future. Better dissemination of research information now from a team working on blood rheology may take 20 years to manifest as a new medical engineering technology for diagnosing disease. Such an example serves to illustrate both the necessity to take a long-term view when developing methods for measuring payoff and benefit, and highlights the fact that open access to research findings facilitates exactly the kind of interdisciplinary approach now becoming so much a necessary characteristic of modern science and technology.

Preliminary work has begun on modelling the cash benefit to society from public access to research findings for the research community alone (i.e. this initial work has not attempted to tackle the modelling of benefit to all of the constituencies, listed in point 3 above, that can benefit from open access). Initial modelling work has been carried out by Houghton (2008, 2009a, 2009b) and Houghton *et al* (2009a, 2009b). These investigators developed an economic model for calculating the benefits of open access and have so far applied it to the national situations in Australia, the UK, Denmark and The Netherlands (and are currently doing so for Germany).

In all cases, the findings are that there is a considerable cash saving from an open system of dissemination of research findings. Of particular interest in regard to the question posed here is the element of benefit that economists call 'return to R&D', which is effectively the savings made by increased efficiency and efficacy in research that arise from increased accessibility to information. This is the saving for the research community as a whole, (including to funders, institutions and researchers) described in monetary terms, arising from better accessibility of research information; savings derive from less duplication,

---

[3] http://www.eprints.org/openaccess/policysignup/

reducing plagiarism, greater overall accessibility, facilitation of interdisciplinary research, etc. These economists modelled recurring annual gains in efficiencies and accessibility from one year's research expenditure and, of course, if these changes are permanent – as they should be with the implementation of a suitable federal policy – then they can be converted to growth rates. Such calculations could be used by US economists to calculate the benefits to the US research economy, of public access. Moreover, the Houghton model has been adopted and adapted by Swan (2010), who has applied it to a number of universities in the UK, showing the sort of cash and efficiency savings that individual institutions could enjoy from open dissemination of research information.

This type of economic modelling is just the start of what will undoubtedly be a growing field of endeavour in economics. As well, benefit of public access to research information will come in non-economic forms, and methods to measure these benefits will need to be developed. Already, some workers are attempting to model the impact of developments in basic research fields in applied areas many years down the line (for example, Lewison, 2005), and we anticipate such approaches growing over time as accessibility matures as a friction factor in research dynamics.

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

A public access policy needs three elements:

(i) It must be mandatory. The policy from the NIH was originally voluntary, with researchers encouraged and exhorted to make their publications openly accessible through PubMed Central (PMC). During the first year or so under these conditions, only around 4% of qualifying articles were deposited by their authors in PMC. Once the policy articulated a requirement (that is, it became a mandatory policy) for authors to ensure their articles were made openly accessible, deposit levels increased rapidly. After one year, the deposit level was 57% and has been growing since. Other studies also demonstrate that only mandatory policies bring the kind of level of open literature that is hoped for by research funders (Sale, 2006).

(ii) The effects of the policy must be monitored to see what the level of compliance with the requirement is, so a system to measure deposit levels against expectations needs to be put in place. The Wellcome Trust began measuring compliance in the first year after its own mandatory policy came into effect. The Trust was able to see that levels of deposit were not as great as had been expected and was able to understand what the causes of this were (partly author recalcitrance and partly publisher non-compliance). As a result of this understanding, steps were taken to make authors more aware of the requirement on them and to remind publishers that immediate deposit of articles that the Trust had paid to be made openly accessible was expected. Compliance by both constituencies rose markedly.

(iii) Sanctions, at least implied ones, must be available to the policymaker to use against non-compliers. The research community is not one where heavy-handedness goes down well, and the general evidence shows that once researchers have begun to see the benefits of making their work openly accessible they readily continue the process, but a new mandatory policy needs to be supported by clear guidelines about expectations and what will happen if those expectations are not met. Many universities, including some of our member institutions, have simply stated that the repository is the place where research managers will look when they wish to see the outputs of any member of research staff (i.e. for promotion boards, appraisals, etc). This brings about the desired effect, for researchers do not wish to be disadvantaged by not compiling a complete record of their outputs in the institutional repository[4]. The University of Liege, for example, implemented its mandatory policy from the end of November 2008 when the number of full-text articles in the University's repository was 974, and by the end of October 2009 the number of full-text articles stood at 14,090.

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?

The version that should be made available under a public access policy is the author's final version after the corrections and changes required as a result of the peer review process have been incorporated. This version is normally called the '***postprint***'. The '***preprint***' is the version before peer review and although this is the version that the high-energy physics community makes available for public access in arXiv[5], this solution is not generally favoured across other disciplines. Indeed, there is a good argument for considering it to be bad or even dangerous practice in some fields (notably biomedicine), where inaccurate or misleading articles placed in the public domain before peer-based quality control may have undesirable consequences.

The ***published version*** of an article is laid out and presented in the publisher's house style. While this may be attractive, and the inclusion of page numbers may be convenient for citation purposes, the published version is almost always in PDF format. This inhibits the re-use of the article by semantic technologies (text-mining and data-mining software), thus reducing the utility and value of the article. Moreover, many publishers still insist on an embargo period before even the postprint can be self-archived and most do not ever permit the published PDF to be made publicly accessible. Delay in accessibility reduces the potential impact of an article for all time, as shown in a recent study by Gentil-Beccot et al (2009). Conversely, early open accessibility gives an article an increased impact over comparable articles that are available only through Toll Access: this impact advantage is persistent and continues to diverge positively from the impact of Toll Access counterparts. Below is Figure 4 from that study which shows the data on which the statements above are based.

[4] http://listserver.sigmaxi.org/sc/wa.exe?A2=ind09&L=american-scientist-open-access-forum&D=1&O=A&F=l&P=866

[5] www.arxiv.org

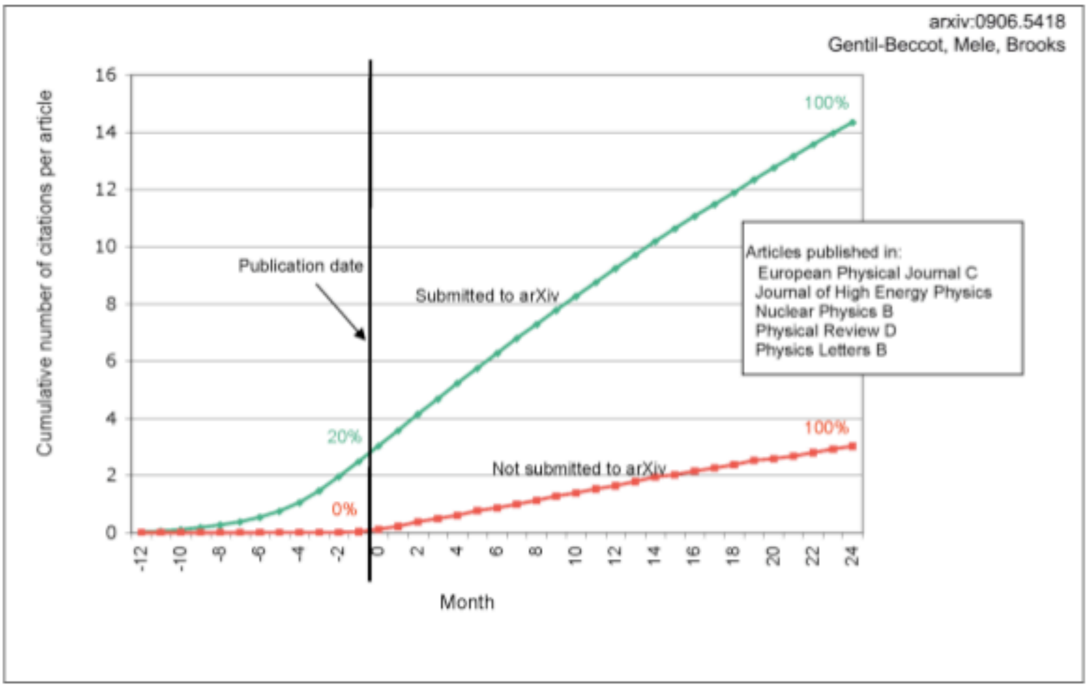


Figure 4. Cumulative citation count as a function of the age of the paper relative to its publication date. 4839 articles from 5 major HEP journals published in 2005 are considered.

(Source: Gentil-Beccot *et al*, 2009: published in *Scientometrics*, December 2009)

Other studies in other disciplines have also shown the same early access advantage (e.g. Henneken *et al*, 2006). It is therefore essential that a public access policy specifies that articles be made accessible as early as possible after peer review. The relative advantages and disadvantages of access to different versions of an article are summarised below:

Version	Advantages	Disadvantages
Preprint	• Very early access, useful in fast-moving fields of research and in small research communities where familiarity and trust between researchers negate the possibility of unreliable information misleading others	• Not appropriate for fields of research where errors and misleading information in uncorrected manuscripts may have serious consequences if used by others
Postprint	• Early access is possible • Quality control has been performed • Article is to all intents and purposes the same as the published version and is an acceptable alternative for those (the majority) without Toll Access	• Has not undergone editorial quality control so may contain grammatical or other errors that should be picked up at proofing and copy-editing stages of the publishing process • Can be subject to delay in access if a publisher's embargo is not counteracted by a properly-worded policy
Published version	• Fully formatted and copy-edited by publisher	• PDF format inhibits machine re-use

- Virtually always delayed-access, even if use is permitted
- Use is usually not permitted by the publisher

7a. 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?

As argued – and supported by empirical evidence – above, the earlier an article is made publicly accessible the greater its impact and the better value it represents for author, funder and American society. The optimal point for articles to be made publicly available if public interests are paramount is immediately after the changes required by peer review have been made.

Publishers will argue for an embargo period (of varying length depending upon discipline), saying that providing immediate access after peer review will harm subscription sales of their journals. We wish to set out quite clearly that this supposition is not supported by any evidence, at least as yet.

The evidence shows the opposite is true. Two society publishers have been doing business in the high-energy physics arena during the whole period when the arXiv has been in operation (since 1991). Effectively, all articles in high-energy physics and allied fields have been made publicly available through the arXiv over that time period, while at the same time being published as usual in subscription journals. If this form of open access were to have an effect on subscriptions, it would be expected to have shown up in high-energy physics during that period.

In 2005, two important society publishers in physics, the American Physical Society (APS) and the Institute of Physics Publishing (IOPP) in the UK, were asked about their experiences of living alongside the arXiv for 14 years[6]. There are two salient points to note from their responses:

1. Neither can identify any loss of subscriptions to the journals that they publish as a result of the arXiv.
2. Subscription attrition, where it is occurring, is the same in the areas that match the coverage of the arXiv as it is across any other areas of physics that these societies publish in.

Both societies, moreover, see actual benefits for their publishing operations arising from the existence of arXiv. The APS spokesman stated that the American Physical Society has *"cooperated closely with arXiv including establishing a mirror (jointly with Brookhaven National Laboratory)... We also revised our copyright statement to be explicitly in favor of author self-archiving. These efforts strengthened (rather than weakened) Physical Review D [an APS journal that covers high-energy physics] ...I would say it is likely we maintained subscriptions to Physical Review D that we may otherwise have lost if we hadn't been so pro-arXiv .."*

[6] http://listserver.sigmaxi.org/sc/wa.exe?A2=ind05&L=american-scientist-open-access-forum&D=1&O=A&F=l&P=6121

In answer to the question "Does arXiv worry or threaten your business?" the APS answered:

> *"We don't consider it a threat. We expect to continue to have a symbiotic relationship with arXiv. As long as peer review is valued by the community (and it seems to be), we will be doing peer review. While the APS aspires to open access and is not threatened by arXiv.org, we do have strong reservations about government requirements for open access."*

The Institute of Physics Publishing's response was:

> *"IOPP's experience as a learned society publisher illustrates the strong synergies and mutual benefits that currently exist between major peer-reviewed journals, such as our Classical and Quantum Gravity, and the arXiv e-print server. Both systems continue to serve the scientific community very effectively. Journals act as the "brand", setting standards for scientific quality. Our authors and editors tell us that they value publishing in a peer-reviewed journal because this continues as an essential requirement for establishing reputation and authority of the research they publish. Whilst posting an pre-print or post-print is becoming more of an essential in some areas of the physics community for immediate and wide dissemination, we do not see the arXiv or repositories threatening our business."*

Gene Sprouse, Editor-in-Chief at the American Physical Society, recently contributed further comments[7] on this issue as follows:

> *"1) One piece of information in the recent work of Gentil-Beccot, Brooks and Mele* http://arxiv.org/abs/0906.5418v1 *is that papers put on the arXiv and never published in a peer reviewed journal are cited 1/6 as much as those that are eventually published. In addition, papers submitted to peer review journals that were never put on the arXiv also suffer the same low citations.(see Figure 2). The American Physical Society, publisher of Physical Review D, has embraced and supported the arXiv, as it provides the "newspaper" for the field, while the journals provide the validation. Both are necessary.*
> *2) Our experience at APS is that we have not seen significant loss of subscriptions of PRD (Physical Review D) as a result of the arXiv. PRD is the largest of the HEP journals, with currently more than 45% of the HEP literature."*

7b 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?

There should not be a delay period past the moment when peer review-required corrections have been made. The world is not yet so perfect, however, and while all rights to dissemination pass routinely to publishers there will always be a difficulty in enabling early accessibility due to publisher desires to restrict access, at least for a period (which may be long) to protect their revenues under an access-restriction business model.

---

[7] http://listserver.sigmaxi.org/sc/wa.exe?A2=ind09&L=american-scientist-open-access-forum&D=1&O=A&F=l&P=70112

There are a number of ways that a policy covering federally-funded research might overcome this problem to enable the early access that is so important:

- The policy could be worded to require that manuscripts are deposited in pubic archives as soon as they are finalised after peer-review (i.e. as postprints). If the publisher of the journal in which a paper is to be published requires an embargo period on access, the full-text of the article may be kept 'dark' for the embargo period. The metadata, however, over which the publisher has no rights, should be made openly accessible from the time of deposit. This at least allows potential users to know that the article exists and they will be able to contact the author about it if they need it during the embargo period. Such a policy has been implemented in a number of our member institutions and it works extremely well as a pragmatic solution to accommodating publisher embargoes whilst enabling the earliest possible access.
- The policy could usefully be modelled on the existing NIH one, where assigning the funder the right to disseminate a publicly available copy is a condition of grant. The NIH terms and conditions of grant require grantees to retain that right and assign the NIH a non-exclusive license to disseminate the author's final manuscript. It is simple, elegant and has proved itself in practice.

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

There are existing standards for archiving and interoperability and any system for exposing the results of federally-funded research must adhere to those standards if it is to achieve the objective of maximizing usage and impact of that research.

(a) Interoperability

i) Search, find and retrieve:

The foundation standard for harvesting content between repositories is ***OAI-PMH*** (Open Archives Initiative-Protocol for Metadata Harvesting). This enables content to be collected from open access research repositories and amassed in others. It thus enables a system whereby original documents are deposited in local, distributed repositories based in research institutions and then these articles – or just their metadata – are harvested for use by other services. These services, such as OAIster or Google Scholar, provide the search-and-discover and retrieve functions that help would-be users to locate and access articles of interest. Both of these examples harvest content across all disciplines from institutional repositories and present a complete corpus to the user. Other services harvest specific types of content, such as all articles in a particular subject area or all electronic theses in Europe. They then present these to the user as a particular 'view' of the research corpus.

A system for making federally-funded research publicly accessible might offer such convenient 'views' as well as constructing a comprehensive national-level collection of US research and presenting that as a service, too. A number of countries[8] already have such

[8] http://repinf.pbworks.com/Harvesters+-+national+and+international

national-level ‘shop window’ services: specific examples are *Intute Repository Search* in the UK[9], *JAIRO* in Japan[10] and *NARCIS* in The Netherlands[11].

ii) A US public access model:
There is an optimal model for how a nation could develop a system for making research publicly accessible in these ways[12]. It was developed in a study for the JISC (Joint Information Systems Committee) in the UK (Swan *et al*, 2005). The study recommends the so-called ‘harvesting model’, much as described above, where institutions collect the content in their local repositories and this content is harvested by national-level services that provide the ‘front-end’ for users. One of the advantages of such a system is that institutions benefit from collecting a complete record of their research outputs, and have a shared interest with their researchers in the enhanced visibility and impact that open accessibility brings to these local collections. It is thus in their interest to ensure that content is collected. Institutions themselves also have policies on open access (there are currently 79 mandatory institutional policies), so a federal policy that covers research carried out by federal agencies could be enhanced by institutional policies that would develop alongside it if the federal policy recommended that universities collect the articles in their own repositories for harvesting. Not only would federally-funded research outputs be collected, then, but also outputs from research funded by other bodies and from unfunded research.

iii) Exchange of digital objects:
A further OAI standard, ***OAI-ORE*** (Open Archives Initiative-Object Re-use and Exchange) has been developed to facilitate the exchange of items between digital collections and enable them to be re-used in their new contexts. Details of both standards are on the OAI website[13]. It is critically important that federally-funded research outputs are stored in digital collections that operate according to this standard, otherwise they will not be maximally accessible.

It is worth noting that the use of OAI-PMH and ORE – while practicable – is not optimal because they rely on a metadata standard named DC (Dublin Core). Since 1999 this has attracted criticism from the euroCRIS community[14] involved in systems concerning research information. The European Union has recommended to member states a standard named CERIF (Common European Research Information Format) and has requested euroCRIS to maintain, develop and promote it. It is being taken up rapidly throughout Europe and has the advantage over traditional repository technologies of a formal syntax and declared semantics including full multi-character set/media type, full multimedia and full multilingual features. CERIF provides much richer metadata than DC not only for publications but also for other outputs (products, patents) and the whole research context (persons, organisations, projects, funding, events, facilities, services, equipment and so forth). CERIF can interoperate comfortably with repositories based on DC and interoperation based on

---

9 http://www.intute.ac.uk/irs
10 http://jairo.nii.ac.jp/en/
11 http://www.narcis.info/background
12 http://www.openscholarship.org/jcms/c_6801/national-oa-model
13 http://www.openarchives.org/
14 www.eurocris.org

OAI-PMH/OAISTER as a value-added overlay. CERIF has – for exchange and interoperation – a XML variant.

(b) Archiving: preservation is an area receiving much attention and there are a number of preservation standards for digital research material. The current picture is described in reference material developed for a project (ongoing) on repository interoperability, International Repositories Infrastructure[15]. There is a wiki for this project that contains extensive background information on the state of play of interoperability developments for digital research repositories. This information, presented under a number of categories, can be found by following links under 'Briefing Materials' from the front page of the wiki[16]. We recommend this resource is consulted if further detail is required on interoperability, since the resource is trusted and has been peer reviewed by the international repository community. We are also appending to this document map versions of these pages.

In our view, it is unlikely that these standards will change materially, though it is highly likely that refinements and modifications will be developed over time as new needs arise and better ways of serving needs are identified.

New developments in the field of interoperability include work on unique identifiers (for authors, institutions, and objects), and technologies to improve ingest of items into repositories, including enabling deposit into more than one repository at a time, thus saving authors from having to make multiple deposits. These sorts of technological developments will undoubtedly lead to new standards in future.

Our overall opinion on these technical issues is, though, that the answers to these questions do not need to be determined for the purpose of writing law, or even mandated across the board. Some experimentation should be encouraged so that the different agencies can find the most appropriate solutions for their needs. Systems may well evolve over time accordingly.

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

i) Use by machines:
If the opportunity for re-use is to be maximised, peer-reviewed papers need to be available in non-PDF format. The ideal format is XML, and mark-up is best done at source. Most authors are not yet accustomed to providing XML documents, however, though the newest edition of Microsoft's Office software produces XML output as default, so the situation will improve. XML conversion can be done after the paper has been deposited by the author. The practice at UK PubMed Central (UKPMC), for example, is for XML conversion (plus

[15] http://repinf.pbworks.com/Preservation
[16] http://repinf.pbworks.com/

additional mark-up to enable particular text-mining technologies that have been developed specifically for application on the biomedical research literature) to be carried out after deposit wherever needed. The process has a cost attached, but the UKPMC sponsors consider that the payoff, in terms of optimising re-use potential, is worth the cost.

An open access system for federally-funded research might implement a similar arrangement since it has the merit of standardizing the technical presentation of documents and optimizing this for re-use. We anticipate that technical standards for maximizing re-use may diverge between different disciplines over time as semantic mining technologies develop to best serve the needs of those disciplines.

ii) Use by people:
Use can be measured in a number of ways, including by metrics such as download numbers, visitor numbers and page views. The Federal government should create a system of monitoring use of its public access collections that provides as much information about users as possible without contravening data protection and privacy rules. For example, most university repositories collect data on page views and download numbers both for the whole collection, and by author and article.

Some of our own repositories additionally record:
- the domain from which users originate
- the referrer (which service refers the user to the article (e.g. Google or a certain blog posting, etc)
- the key words or search terms that the user employed in their search algorithm
- the time of the 'hit' on an article
- the pattern of hits for the whole collection and for each article (thus providing useful information about capacity requirements for the system as well as for more general usage monitoring)

It is possible also to add user feedback in a variety of ways and these will increase and change with time. The high-energy physics arXiv permits 'trackbacks' where blog comments can be linked to from the article in arXiv that sparks the comment[17] though this system has not been without problems or criticism. Inclusion of the facility to add comments by users to items in the Federal collection is worth considering, though due attention should be paid to the issue of quality control for such a system. It may become burdensome or controversial. The arXiv uses a team of editors to verify the validity of any Trackbacks (not without controversy) before allowing them to be added to the database and it would seem that some sort of verification system would also need to be used to ensure quality and propriety of open commentary systems for a corpus of Federal research outputs.

[17] http://arxiv.org/help/trackback/

This submission
Submitted to the Office of Science & Technology Policy by the Board of *Enabling Open Scholarship*, January 2010:

Professor Bernard Rentier (chair), Rector of the University of Liège, Belgium
Professor Tom Cochrane, Deputy Vice Chancellor, Queensland University of Technology, Brisbane, Australia
Dr William Dar, Director General of the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India
Professor Martin Hall, Vice Chancellor, University of Salford, UK
Professor Stevan Harnad, Canada Research Chair, Université du Québec à Montréal (UQAM), Montreal, Canada
Professor Keith Jeffery, Director of IT and International Strategy at the Science & Technology Facilities Council, Swindon, UK
Dr Sijbolt Noorda, President of VSNU, the Association of Dutch Research Universities, The Hague, Netherlands
Professor Stuart Shieber, James O. Welch, Jr. and Virginia B. Welch Professor of Computer Science, School of Engineering and Applied Sciences at Harvard University and Director of Harvard's Office of Scholarly Communication, Cambridge, USA
Professor Ian Simpson, Deputy Principal for Research and Knowledge Transfer, and Professor of Environmental Science, University of Stirling, UK
Professor Peter Suber, Berkman Center for Internet & Society, Harvard University, Cambridge, USA
Professor John Willinsky, Khosla Family Professor of Education at Stanford University and director of the Public Knowledge Project at the University of British Columbia and Simon Fraser University
Dr Alma Swan (Convenor/Coordinateur), Director of Key Perspectives Ltd, Truro, UK

Appendix 1: research funders with mandatory policies on Open Access:

Australia	Australian Research Council
Australia	National Health & Medical Research Council
Austria	Fonds zur Foerderung der wissenschaftlichen Forschung
Belgium	Research Foundation Flanders
Canada	Michael Smith Foundation for Health Research
Canada	Canadian Health Services Research Foundation
Canada	Fonds de la recherche en sante Quebec (FRSQ)
Canada	Canadian Breast Cancer Research Alliance
Canada	National Research Council (NRC) Canada
Canada	Ontario Institute for Cancer Research (OICR)
Canada	Canadian Cancer Society Research Institute
Canada	Canadian Institutes of Health Research (CIHR)
Europe	European Commission
Europe	European Research Council
France	Agence Nationale de la recherche (ANR) (Humanities and Social Sciences Branch)
Hungary	Hungarian Scientific Research Fund (OTKA)
Ireland	Science Foundation Ireland (SFI)
Ireland	Irish Higher Education Authority (HEA)
Ireland	Irish Research Council for Science, Engineering & Technology
Norway	Norwegian Research Council
Spain	Government of the Principality of Asturias
Spain	Madrid Autonomous Community of Spain
Sweden	Swedish Research Council
Switzerland	Swiss National Science Foundation (SNF)
UK	Engineering & Physical Sciences Research Council (EPSRC)
UK	JISC (Joint Information Systems Committee)
UK	Arts and Humanities Research Council (AHRC)
UK	Arthritis Research Foundation
UK	British Heart Foundation
UK	Cancer Research UK
UK	Chief Scientist Office (Scottish Executive Health) Department
UK	Department of Health (UK)
UK	Science & Technology Facilities Council
UK	Natural Environmental Research Council (NERC)
UK	Biotechnology and Biological Sciences Research Council (BBSRC)
UK	Economic and Social Research Council (ESRC)
UK	Medical Research Council (MRC)
UK	Wellcome Trust
USA	Institute of Education Science
USA	Autism Speaks
USA	National Institutes of Health (NIH)
USA	Howard Hughes Medical Institute (HHMI)

References

CEPA (Cambridge Economic Policy Associates) (2008) Activities, costs and funding flows in the scholarly communications system in the UK: http://www.rin.ac.uk/costs-funding-flows

CIBER (2004) *Scholarly communication in the digital environment: What do authors want?* CIBER Group, UCL, London. www.publishing.ucl.ac.uk/papers/2004Rowlands_etal.pdf

Gentil-Beccot A, Mele S and Brooks T (2009) Citing and Reading Behaviours in High-Energy Physics. How a community stopped worrying about journals and learned to love repositories. http://arxiv.org/abs/0906.5418

Henneken E, Kurtz M, Eichhorn G, Accomazzi A, Grant C, Thompson D and Murray S (2006) Effect of E-printing on Citation Rates in Astronomy and Physics http://arxiv.org/abs/cs/0604061

Houghton, J *et al* (2008) Economic Implications of Alternative Scholarly Publishing Models: Exploring the costs and benefits. http://www.jisc.ac.uk/publications/documents/economicpublishingmodelsfinalreport.aspx

Houghton, J., Rasmussen, B. and Sheehan, S.; with Oppenheim, C., Morris, A., Creaser, C., Greenwood, H., Summers, M. and Gourlay, A. (2009a), Economic implications of Alternative Scholarly Publishing Models: Exploring the Costs and Benefits, JISC EI-ASPM Project, Report to the Joint Information Systems Committee (JISC) (UK), CSES and Loughborough University, January 2009.
http://www.jisc.ac.uk/publications/documents/economicpublishingmodelsfinalreport.aspx
http://alpsp.publisher.ingentaconnect.com/content/alpsp/lp/2007/00000020/00000002/art00002
http://www.jisc.ac.uk/whatwedo/topics/opentechnologies/openaccess/reports/learnedsociety.aspx

Houghton, JW, de Jonge, J and van oploo, M (2009b) Costs and benefits of research communication: the Dutch situation. SURF Foundation, Utrecht.
http://www.surffoundation.nl/en/publicaties/Pages/CostsandBenefitsofOpenAccessPublicationlTheDutchSituation.aspx

Houghton, JW (2009a) Costs and benefits of alternative publishing models: Denmark. Danish Electronic Research Library (DEFF), Copenhagen.
http://www.deff.dk/content.aspx?itemguid={EACA73FB-2EFE-44CA-92CD-91C4416C0370}

Houghton J (2009b) Open Access – what are the economic benefits? A comparison of the United Kingdom, Netherlands and Denmark. A report to Knowledge Exchange. Knowledge Exchange, Copenhagen.
http://www.knowledge-exchange.info/Admin/Public/DWSDownload.aspx?File=%2fFiles%2fFiler%2fdownloads%2fOA_What_are_the_economic_benefits_-_a_comparison_of_UK-NL-DK__FINAL_logos.pdf

Jubb M, Look H and Sparks S. (2007) Researchers' use and perceptions of discovery services. *Learned Publishing* **20 (2),** pp147-153.

RIN (2009) Overcoming barriers: access to research information content. Research Information Network, London. http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/overcoming-barriers-access-research-information

Sale, AHJ (2006) Comparison of IR content policies in Australia. *First Monday*, 11 (4). http://eprints.utas.edu.au/264/

Swan, A (2008) The 'Big Picture' and researchers' top concerns about the scholarly communication process. A report for the JISC Scholarly Communications Group. http://eprints.ecs.soton.ac.uk/15450

Swan A (2010) Modelling scholarly communication transition: costs and benefits for universities. In press, to be published early February 2010 by the Joint Information Systems Committee, London.

Ware, M. (2009) Access by UK small and medium-sized enterprises to professional and academic literature. Publishing Research Consortium, London. http://www.publishingresearch.net/documents/SMEAccessResearchReport.pdf