TO: Office of Science and Technology Policy
FROM: Oya Rieger, Associate University Librarian, Cornell University Library, Ithaca, NY
EMAIL: rieger@cornell.edu
RE: Response to "Public Access to Peer-Reviewed Scholarly Publications"
DATE: January 12, 2012

Introduction

I am writing in support of governmental mandates that encourage scientists to share their research outputs, especially if the work is funded by the taxpayers. As the associate university librarian for digital scholarship and preservation services at Cornell University Library, I would like to offer examples from my own program area in support of the RFI, specifically related to the potential impact of open access for improving the scientific enterprise. I oversee the Library's digitization, online repository, digital preservation, electronic publishing, and e-scholarship initiatives with a focus on needs assessment, requirements analysis, business modeling, and information policy development.

arXiv: An Open Access Success Story

The current publishing ecology has a diverse range of stakeholders including commercial and university publishers, scholarly societies, and libraries. Open access to social and scientific information is not mutually exclusive with commercial publishers. An excellent example of this dynamics is arXiv, which is the primary daily information source for hundreds of thousands of researchers in physics, and plays an increasingly prominent role in mathematics, computer science, and other related fields (Gingparg, 2011). With 700,000 e-prints, it provides an instant communication mechanism for scientists and complements the formal publishing process, which may take several months. Faster and unmediated access enables scientists, both in academic and entrepreneurial institutions, to incorporate new findings into their research faster. Since its launch in 1991, arXiv has achieved iconic status as an effective online distribution system and is often cited to illustrate digital repositories' potential role in transforming scholarly communication. Such an impact is difficult to measure in financial terms due to its deep scholarly communication infrastructure roots.

One of the premises of arXiv has been making science more democratic by allowing for the rapid worldwide dissemination of scientific findings. The RFI is focusing on the U.S. economy an commerce; nevertheless, open access to scientific information has global implications without national boundaries. Figure 1 illustrates the international reach of arXiv. It is a global initiative, involving dedicated mirror sites in 17 countries and collaboration with U.S. and foreign professional societies and other international organizations. It has also provided a crucial life-line for isolated researchers in developing countries. Most scientists and researchers who post content on arXiv also submit it for

publication in traditional peer-reviewed journals. However, famously reclusive Russian mathematician Grigori Perelman's decision to post his proof of the 100-year-old Poincaré Conjecture solely in arXiv underscores the repository's increasing importance and its role in transforming scholarly communication.[1] We are in the process of parsing the 24% usage shown as "other" in Figure 1 but our early analysis indicates that there is strong use by commercial entities.

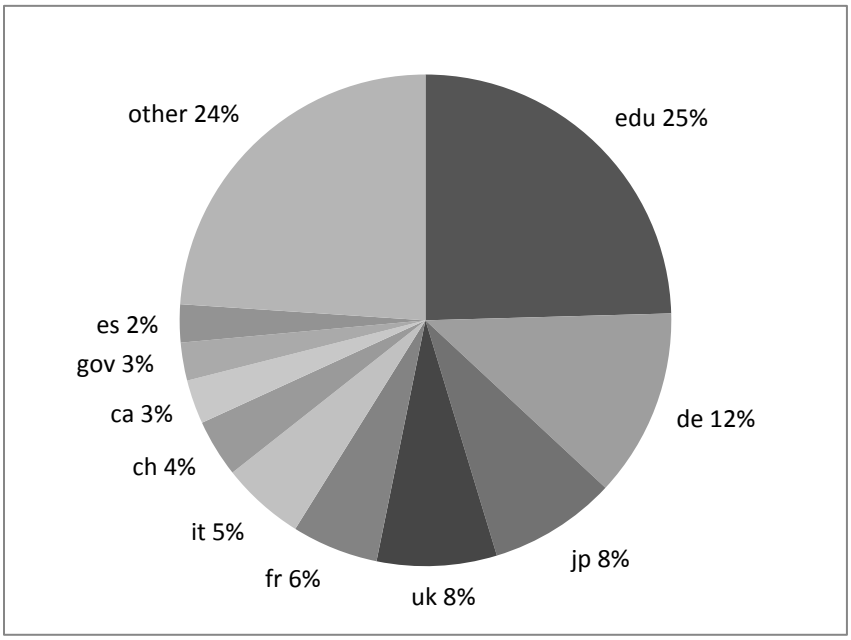

Figure 1: arXiv institutional downloads at main site by Internet domain of institutions (2010)

arXiv does not compete but co-exists with commercial publications. There is a unique role for each mode of dissemination. arXiv facilitates rapid and global dissemination of research results (*place* to *stake* intellectual precedence claims) whereas commercial journals continue to provide a venue for peer review and tenure requirements. In celebration of the arXiv's 20th anniversary, on September 23, 2011 Cornell University Library (CUL) hosted a meeting at Cornell with the representatives from several publishers and societies, including Elsevier, Wiley, and Springer. The goal of the forum was to discuss the feasibility and desirability of establishing a research and innovation collaboration in

[1] The Poincaré Conjecture is the only one of seven famous mathematical problems identified by the Clay Mathematics Institute that has been solved. For this work, Perelman was awarded the prestigious Fields Medal (which he declined) and in 2006 the journal Science named his proof of the Poincaré conjecture as its annual Breakthrough of the Year.

support of arXiv. The commonality among all the participants is their interest in understanding and meeting the needs of the scientific community.

arXiv is an example of a cost-effective open access delivery method. Since it moved to Cornell in 2001, the Cornell University Library has provided the bulk of arXiv's operating costs, which are projected to be approximately $500,000 for 2011.[2] In January 2010, Cornell has established a voluntary institutional contribution model and invited pledges from the top 200 libraries and research laboratories accounting for more than 75 percent of annual institutional downloads (Rieger, 2011). Based on a budget of $330,000 and 40 million paper downloads for 2010, each e-print costs merely 0.08 cents per download and the cost per submission is $4.70.

Digital Preservation Through Open Access

Scholarship has been created and sustained through the interoperation of three key agents that have formed an infrastructure for sustainability:

- Scientists and their academic and professional networks
- Publishing organizations including scholarly societies
- Libraries and archives

This network has been changing due to the new modes of digital scholarship. Academic libraries are increasingly dependent on commercially-produced, born-digital content that is purchased or licensed. E-journals have replaced the majority of titles formerly produced in paper format. Cornell University Library spends more on e-materials than other forms of content. The finding of a recent Cornell and Columbia University Libraries study that assesses the role of LOCKSS and PORTICO in preserving each institution's e-journal collections was alarming (LOCKSS Team, 2010). Although LOCKSS and PORTICO are considered successful digital preservation initiatives, only 15-20% of the e-journal titles in the libraries' collections are currently preserved by these two initiatives.[3] From users' perspective, there is an implicit assumption that today's electronic journal content will be refreshed and digitally manipulated as required to carry it forward indefinitely over time.

As we move to greater dependency on digital content, we must rethink how we go about managing our preservation responsibilities. An important benefit of open access mandates is supporting enduring access through redundancy of e-prints available from different

[2] The arXiv budget for 2011 is available at http://arxiv.org/help/support/2010_budget. It is based on an estimate and will be updated throughout the year to reflect the actual expenses.

[3] Information about LOCKSS and PORTICO is available at http://www.jisc.ac.uk/publications/briefingpapers/2007/pub_ejournalspreservationbp.aspx

sources. For instance, if an article is published only in a born-digital commercial journal, the long-term accessibility of the work is solely based on the preservation provisions put in place by the publisher. Unlike journals with print and digital version, for born-digital articles, there is not a physical copy that will be archived and maintained by libraries and archives. Whereas an open access journal article is much more likely to be deposited in multiple repositories and therefore creating a security network through redundant digital copies maintained through different systems. So, unlike the proprietary publisher scenario, there is not a single point of failure and the risks are distributed.

The long-term viability of digital scholarly content managed solely by publishers' is uncertain. They are entrepreneurial entities with focus on return-on-investment. Due to the volatility in the publishing market, commercial publishers should not have the sole responsibility for providing long-term access to the output of the publicly funded research. When a publisher goes out of business, it is very likely that the content they manage will become inaccessible. To ensure that federally funded scholarly research outputs are permanently available online, there must be federal mandates to enable and authorize depositing articles to open access repositories. NIH Public Access Policy is exemplary as it requires that grantees submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central upon acceptance for publication.

To sum, although it is ideal to separate the issues of access to peer-reviewed research and digital preservation of the published intellectual and cultural record, these two areas are often intertwined with co-dependency. Preservation of the published versions of research outputs requires digital preservation infrastructures and services. However, due to the lack of a scalable and reliable preservation infrastructure for born-digital scholarly articles, the retention of prints and post-prints through institutional repositories or centralized subject-based repositories constitutes a critical archival strategy through redundancy and multiple copies.

Creating Sustainable Infrastructure for Open Access

Open access model allows transparency and accountability about how knowledge is created, verified, analyzed, and interpreted. Although I am supportive of encouraging free and open access wherever feasible, federal directives requiring uninhibited discovery of information is not an end in itself. Open access mandates from federal and state agencies are fundamental enablers; however, there needs to be a sustainable infrastructure based on the following principles:

- Scalable and cost-efficient databases to store, discover, access, and re-purpose information.
- Management policies and procedures in place to ensure the enduring usability, authenticity, discoverability, and accessibility of content over the very long-term (digital preservation).
- Interoperability arrangements that link a given repository to related systems, services, and communities.
- Leveraging existing technical and policy infrastructure to minimize unneeded duplication of efforts and support cost-effectiveness and long-term sustainability.
- Features that support supplementary information objects such as underlying data, auxiliary multimedia content, and research methodologies.
- Flexibility to accommodate different embargo periods in support of professional, academic, and entrepreneurial requirements of scientists and research institutions.
- Incentives, rewards, and recognition for scientists who share and archive the outputs of their research endeavors.
- Support for IPR, privacy, and confidentiality – especially if research entails human subjects or other sensitive and potentially misleading information.
- Functionality and arrangements that lower barriers for scientists to contribute content to multiple complementary repositories.
- Sensitivity to disciplinary cultures, practices, norms, and aspirations.
- Active participation in the development of community standards for deposition, use, and maintenance of scholarly information.
- Provisions for collaboration among communities and information types (e.g., articles, data, images) in order to encourage interdisciplinary scholarship.

In closing, I also urge the White House to adopt a more inclusive definition of research to include the work of humanities scholars. This will require that in addition to scientific organizations such as NSF and NIH, agencies such as NEH is also involved in developing open access policies. Interdisciplinary mandates will be instrumental in facilitating and

promoting collaboration among all kinds of scientists, including the humanities and cultural heritage community.

References

Ginsparg, P. (2011). arXiv at 20. *Nature*, no. 476, p. 145–147, http://www.nature.com/nature/journal/v476/n7359/full/476145a.html

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

Rieger, O.Y. (2011) Assessing the Value of Open Access Information Systems: Making a Case for Community-Based Sustainability Models. *Journal of Library Administration*, 51(5-6), 485-506. http://bing.exp.sis.pitt.edu:8080/webdav/new_documents/rieger_open_access_value.pdf