Commentary

This is one of a series of commentaries on the future of scientific publishing. For a listing of the other commentaries, see http://www.jneurosci.org/cgi/content/full/26/36/9077.

As We May Read

Paul Ginsparg
Departments of Physics and Information Science, Cornell University, Ithaca, New York 14853

The e-print arXiv (http://arXiv.org/), initiated in August 1991, has effectively transformed the research communication infrastructure of multiple fields of physics and could play a prominent role in a unified set of global resources for physics, mathematics, and computer science. It has grown to contain >375,000 articles (as of July 2006), with >50,000 new submissions expected in calendar year 2006 and >40,000,000 full-text downloads per year. It is an international project, with dedicated mirror sites in 17 countries and collaborations with United States and foreign professional societies and other international organizations, and it has also provided a crucial lifeline for isolated researchers in developing countries (for some general background, see Ginsparg, 1996).

The arXiv is entirely scientist driven: articles are deposited by researchers when they choose (either before, simultaneous with, or after peer review), and the articles are immediately available to researchers throughout the world. As a pure dissemination system, it operates at a factor of 100–1000 times lower in cost than a conventionally peer-reviewed system (Ginsparg, 2001). This is the real lesson of the move to electronic formats and distribution: not that everything should somehow be free, but that with many of the production tasks automatable or offloadable to the authors, the editorial costs will then dominate the costs of an unreviewed distribution system by many orders of magnitude. Even with the majority of science research journals now on-line, researchers continue to enjoy both the benefits of the rapid availability of the materials, even if not yet reviewed, and open archival access to the same materials, even if held in parallel by conventional publishers. The methodology works within copyright law, as long as the depositor has the authority to deposit the materials and assign a nonexclusive license to distribute at the time of deposition, because such a license takes precedence over any subsequent copyright assignment.

The site has never been a random Usenet newsgroup- or blogspace-like free-for-all. From the outset, arXiv.org relied on a variety of heuristic screening mechanisms, including a filter on institutional affiliation of submitter, to ensure insofar as possible that submissions are at least "of refereeable quality." This means that they satisfy the minimal criterion, that they would not be peremptorily rejected by any competent journal editor as nutty, offensive, or otherwise manifestly inappropriate, and they would instead at least in principle be suitable for review. These mechanisms are an important, if not essential, component of why readers find the arXiv site so useful. Although the most recently submitted articles have not yet necessarily undergone formal review, the vast majority of the articles can, would, or do eventually satisfy editorial requirements somewhere. To adapt to the expansion of Internet usage from the academic community to society at large, an endorsement system (http://arXiv.org/help/endorsement) was implemented in 2004 so that new submitters can first be certified by existing contributors. This helps ensure that the arXiv remains a forum for communication among research professionals, not a mechanism for outsiders to communicate to that community. Additionally, a small group of volunteer "moderators," consisting of interested experts from around the world, cursorily prescans new submissions, typically only at the level of title and abstract, for appropriateness to the proposed primary subject area.

The arXiv repository functions are flexible enough either to coexist with the preexisting publication system or to help it evolve into something better optimized for researcher needs. Although there are no comprehensive editorial operations administered by the site, the vast majority of the 50,000 new articles per year are nonetheless subject to some form of review, whether by journals, conference organizers, or thesis committees. Physics and astronomy journals have learned to take active advantage of the availability of the materials before journal publication, and the resulting symbiotic relation would not have been anticipated 15 years ago. The idea of using such electronic distribution before publication to augment the referee process goes back at least to Rogers and Hurt (1989). Simple proposed modifications of the peer review include a two-tier system (for more details, see Ginsparg, 2002), in which, on a first pass, only some cursory examination or other pro forma certification is given for acceptance into a standard tier. At some later point, a much smaller set of articles would be selected for more extensive evaluation.

There is currently much discussion of free access to the on-line scholarly literature. It has long been argued that this ma-

Received July 25, 2006; revised July 25, 2006; accepted July 25, 2006.

Correspondence should be addressed to Paul Ginsparg at the above address. E-mail: ginsparg@cornell.edu.

DOI:10.1523/JNEUROSCI.3161-06.2006

 0270-6474/06/269606-03$15.00/0

terial becomes that much more valuable when freely accessible (Berry, 2001), and moreover, that it is in public policy interests to make the results of publicly funded research freely available as a public good (Bachrach et al., 1998). Studies have shown a clear correlation between openly accessible materials and citation impact (Hajjem et al., 2005; Kurtz et al., 2005; Metcalfe, 2005, 2006; Henneken et al., 2006), although a direct causal link is more difficult to establish. It is also suggested that the move to open access could ultimately lead to a more cost-efficient scholarly publication system. There are recent indications that the United States and other governments may become directly involved by mandating some form of open access for research funded by government agencies.

The response of the publishing community has been that their editorial processes provide an essential service to the research community, that these are labor-intensive and hence costly, and that even if delayed, free access could impair their ability to support these operations. In short, it costs real money to do quality control via the time-honored methodology, but that cost varies significantly from publisher to publisher, as does the profit. If we choose to persist in the current methodology, the requisite funds must continue to flow to the same, or perhaps more cost-efficient, intermediaries between authors and readers. If we choose to reduce the flow of funds, we need a different methodology for quality control and authentication of the materials. *The Journal of Neuroscience* is already in compliance with a typical form of the proposed government mandate, making articles freely available 6 months after publication. This means that only ~600 *Journal of Neuroscience* articles currently require subscription: i.e., >96% of the >17,000 articles since 1981 are freely available to the public.

A form of open access appears to be happening by a backdoor route: using standard search engines, more than one-third of the high-impact journal articles in a sample of biological/medical journals published in 2003 were found at nonjournal Web sites (Wren, 2005). To assess the extent of this phenomenon less systematically in the neuroscience community, I looked up the publications posted at http://brainmap.wustl.edu/resources/papers.html, the laboratory Web site of the incoming president of the Society for Neuroscience, David Van Essen. The society president can be viewed as a role model, as a representative sample, or as a lower bound on the likely behavior of younger researchers. Of 36 publications listed from 2000 or later, 27 full texts were available without subscription either as preprints, at open-access journal sites, or as copies at nonjournal Web sites. Five of the remainder seemed to be available on-line only as abstracts from journal sites requiring subscription for the full text, and the other four (all very recent) did not appear to be available on-line at all, perhaps still undergoing review. The result is striking: at least 75% of the publications listed were freely available either via direct links from the above Web page or via a straightforward Web search for the article title. If indeed this is representative, then the neuroscience community may already be farther along in the direction of open access than most realize.

Because the current generation of undergraduates, and the next generation of researchers, already takes for granted that such materials should be readily accessible from anywhere, it is more than likely that this percentage will only increase over time and that the publishing community will need to adapt to the reality of some form of open access, regardless of the outcome of the government mandate debate.

There is more to open access, however, than just the free access assessed above. True open access permits any third party to aggregate and data mine the articles, themselves treated as computable objects, linkable and interoperable with associated databases. We are still just scratching the surface of what can be done with large and comprehensive full-text aggregations. A forward-looking example is provided by the PubMed Central database (http://www.pubmedcentral.nih.gov/), operated in conjunction with GenBank and other biological databases at the United States National Library of Medicine. It is growing rapidly and already contains >250,000 recent articles in fully functional Extensible Markup Language (XML) from >250 journals (and additionally >400,000 scanned articles from back issues). The full-text XML documents are parsed to permit multiple different "views." GenBank accession numbers are recognized in articles referring to sequence data and linked directly to the relevant records in the genomic databases. Protein names are recognized, and their appearances in articles are linked automatically to the protein and protein interaction databases. Names of organisms are recognized and linked directly to the taxonomic databases, which are then used to compute a minimal spanning tree of all of the organisms contained in a given document. In yet another view, technical terms are recognized and linked directly to the glossary items in the relevant standard biology or biochemistry textbook in the books database. The enormously powerful sorts of data mining and number crunching that are already taken for granted as applied to the open-access genomics databases can be applied to the full text of the entirety of the biology and life sciences literature and will have just as great a transformative effect on the research done with it.

On the one-decade time scale, it is likely that more research communities will join some form of global unified archive system without the current partitioning and access restrictions familiar from the paper medium, for the simple reason that it is the best way to communicate knowledge and hence to create new knowledge. Ironically, it is also possible that the technology of the 21st century will allow the traditional players from a century ago, namely the professional societies and institutional libraries, to return to their dominant role in support of the research enterprise.

References

Bachrach S, Berry RS, Blume M, von Foerster T, Fowler A, Ginsparg P, Heller S, Kestner N, Odlyzko A, Okerson A, Wigington R, Moffat A (1998) Who should own scientific papers? Science 281:1459–1460.

Berry RS (2001) Is electronic publishing being used in the best interests of science? The scientist's view. In: Proceedings of the second ICSU Press-UNESCO Conference on Electronic Publishing in Science (Elliot R, Shaw D, eds). Retrieved August 16, 2006, from http://www.icsu.org/5_abouticsu/CDSI_web/EPS2/berryfin.htm.

Ginsparg P (1996) Winners and losers in the global research village. In: Proceedings of the ICSU Press-UNESCO Conference on Electronic Publishing in Science (Elliot R, Shaw D, eds). Retrieved August 16, 2006, from http://www.icsu.org/5_abouticsu/CDSI_web/EPS1/ginsparg.htm.

Ginsparg P (2001) Creating a global knowledge network. In: Proceedings of the second ICSU Press-UNESCO Conference on Electronic Publishing in Science (Elliot R, Shaw D, eds). Retrieved August 16, 2006, from http://www.icsu.org/5_abouticsu/CDSI_web/EPS2/ginspargfin.htm.

Ginsparg P (2002) Can peer review be better focused? Sci Technol Libr 22:5–18.

Hajjem C, Harnad S, Gingras Y (2005) Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. IEEE Data Eng Bull 28:39–47.

Henneken EA, Kurtz MJ, Eichhorn G, Accomazzi A, Grant C, Thompson D, Murray SS (2006)

Effect of e-printing on citation rates in astronomy and physics. Retrieved August 16, 2006, from http://arXiv.org/abs/cs.DL/0604061.

Kurtz MJ, Eichhorn G, Accomazzi A, Grant C, Demleitner M, Henneken EA, Murray SS (2005) The effect of use and access on citations. Inform Process Manag 41:1395–1402.

Metcalfe TS (2005) The rise and citation impact of astro-ph in major journals. Bull Am Astron Soc 37:555–557.

Metcalfe TS (2006) The citation impact of digital preprint archives for solar physics papers. Retrieved August 16, 2006, from http://arXiv.org/abs/astro-ph/0607079.

Rogers S, Hurt C (1989) How scholarly communication should work in the 21st century. The Chronicle of Higher Education 36 (October 18):A56.

Wren JD (2005) Open access and openly accessible: a study of scientific publications shared via the internet. BMJ 330:1128.