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.

Open Access and the Future of the Scientific Research Article

Matthew J. Cockerill and Vitek Tracz
BioMed Central, London W1T 4LB, United Kingdom

Open access to the scientific literature remains a controversial area. To adequately summarize the different arguments and opinions on the matter could easily fill an entire book (Willinsky, 2005). In this commentary, we present the perspective of an open access publisher.

Since BioMed Central launched its pioneering open access journals in 2000, massive progress has been made toward opening up the scientific literature. In the last year, the pace of change has noticeably accelerated, as practical evidence demonstrating the strength of the open access model accumulates.

In the early days of the open access movement, critics expressed concern about the likely quality of peer review under an open access model. The Thomson-ISI statistics (such as the *Genome Biology* 2005 impact factor of 9.71) have mostly addressed such criticisms, however. Impact factors are not a perfect measure of journal quality, but they are, by far, the most widely used objective metric for research assessment, and it seems clear that the open access model is entirely compatible with high standards of peer review.

Another frequently raised concern is that open access might not be economically viable or affordable by the scientific community. However, two reports on the economics of science publishing, commissioned by the Wellcome Trust (2003) and by the European Commission (European Commission Directorate-General for Research, 2004), have both concluded that open access publishing can be expected to be more efficient than the traditional model and so should, in fact, be more affordable for the scientific community. The recent news article in *Nature* on the finances of Public Library of Science (Butler, 2006), which suggested that open access publishing might not be economically sustainable, attracted a strongly worded rebuttal from P. Peters (Peters, 2006) of the Hindawi Publishing Corporation, which already operates a profitable commercial open access publishing business.

An important factor in the recent growth of open access is active support from research funders. Major research funders are no longer willing to let publishers tell them what they can and cannot do with their own research, and this has resulted not just in words but in concrete actions by funding agencies that are determined to maximize access to the research that they have funded.

The National Institutes of Health (NIH) Public Access Policy (http://publicaccess.nih.gov/), launched in February 2005, was an important starting point, although the decision of the NIH to "request" rather than to "require" authors to comply with its policy has resulted in a very low compliance rate of $<4\%$, as of January 2006 (National Institutes of Health, 2006). In contrast, the Wellcome Trust was the first major biomedical funder to require funded researchers to deposit resulting research in an open access archive as a condition of their grant (Wellcome Trust, 2005). Several of the United Kingdom Research Councils have since followed suit (Research Councils UK, 2006).

This explicit support from funders has been an important factor in the continued growth of open access, because it means that authors increasingly have the confidence to publish in open access journals, knowing that their funder encourages such publication and will recognize the value of such publications when it comes to evaluating future grant proposals.

One of the most direct demonstrations of the progress of open access in recent years is the sustained and rapid growth in submissions to the BioMed Central open access journals over the 6 year period since launch (Fig. 1). This growth has continued even as the number of open access choices available for authors has increased dramatically. More and more well established publishers are recognizing that the rules of the game have changed and that they will have to adapt. For example, in first 2 weeks of August 2006, three major publishers (the BMJ Publishing Group, Cambridge University Press, and Wiley) all introduced new open access options for authors publishing in their journals.

In young but rapidly growing fields such as bioinformatics, genomics, and systems biology, open access publications are increasingly the norm rather than the exception. More broadly, many would argue that open access is clearly on the way to becoming a reality. It is therefore worth revisiting why open access is important.

There are many important benefits of open access that have helped to drive its adoption. Immediate barrier-free access to previous publications certainly makes the research process more efficient. The lack of barriers is especially important in facilitating work that spans multiple disciplines (for example, computer scientists and mathematicians need easy access to the latest biological research if they are to

Received Aug. 15, 2006; accepted Aug. 16, 2006.

Correspondence should be addressed to Dr. Matthew J. Cockerill, BioMed Central, 34-42 Cleveland Street, London W1T 4LB, UK. E-mail: matthew.cockerill@biomedcentral.com.

DOI:10.1523/JNEUROSCI.3534-06.2006

 0270-6474/06/2610079-03$15.00/0

be able to effectively spot opportunities for collaboration with biologists in areas such as systems biology). Open access also particularly helps those at less well funded institutions, and in developing countries, whose access under the traditional model is especially constrained. The transparency of the "article processing charge" also promises to deliver a more efficient marketplace for scientific publications, keeping costs down.

There is another reason, however, why open access is not just desirable but critical to the future of biomedical research. The rate at which biomedical knowledge is being generated is exploding as a result of increased investment and the use of high-throughput genomic and proteomic technologies. This already creates a huge problem of information overload for researchers, who cannot possibly keep up with all of the research articles being published in their field. Similarly, curators of molecular function databases such as Uniprot are increasingly finding that it is difficult or impossible to keep up with the published research results that need to be captured and translated into database annotations.

The only feasible solution to this problem is to develop better systems to help researchers work with the literature. Ideally, the current state of biological knowledge, as reported in peer-reviewed research articles, needs to be captured by automated tools that will allow researchers to easily identify relevant facts, conflicts, or correlations, wherever they may be hidden.

An important consequence is that, in the future, the readership of research articles will include not only humans but also the many computerized systems that will be scanning the literature to add the relevant material to their knowledgebase. Open access to raw data and to original research articles is critically important to the development of such tools (Table 1), which is why BioMed Central makes its entire full text XML corpus of >18,000 articles freely available for download and mining (http://www.biomedcentral.com/info/about/datamining).

Text mining is often used to describe the process of taking unstructured text and attempting to infer structure and computer-readable knowledge from it. Unfortunately, accurately disambiguating what an author meant when writing an article turns out to be an immensely difficult task. As a trivial example, when genetics researchers use the term "hedgehog," they may conceivably be referring to the proverbial hedgehog (compared with the proverbial fox). More likely, they are referring to one of the *hedgehog* gene families, but it is hard for an automated system to know for sure (Mons, 2005). Similarly, many automated systems would be confused by a statement such as, "We suspected that protein A bound to protein B, but subsequent investigation showed this not to be the case." This type of problem makes natural language processing into an interesting academic challenge, but there seems to be no immediate prospect of a fully satisfactory solution.

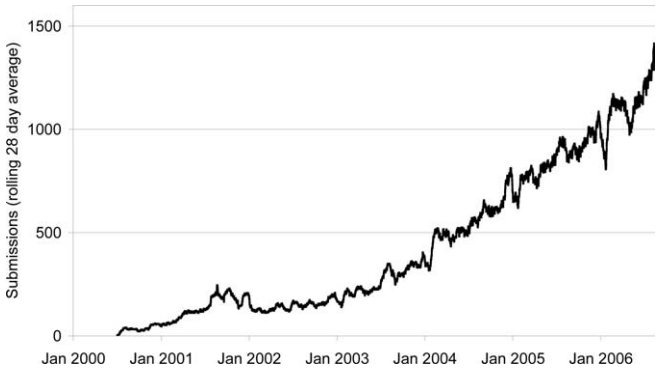

Figure 1. Growth in manuscript submissions to the BioMed Central open access journals.

Table 1. Publications, data, and the role of open access

Publications and data are a continuum:
Publications refer to data
Publications include data
For text mining purposes, publications are data
To make sense of the flood of publications and data resulting from post-genomic research, the community needs:
The best possible tools for mining this material[a]
The widest possible availability of material to be mined[a]

[a]These two factors are linked, because open access stimulates innovation in the creation of tools: open access means that anyone with a good idea for mining the literature can immediately put it into practice.

But perhaps human authors can meet computer systems half way. Perhaps in the future, an important part of publishing a scientific paper will be to ensure that the pertinent new "facts" reported in the paper are expressed in an unambiguous way so as to facilitate their interpretation by automated systems, which will then allow richer and more accurate tools to be created to mine data and synthesize knowledge. BioMed Central is actively working, both with the academic community and with other publishers, to develop standards and tools for the semantic enrichment of the scientific literature. One example of this is the Neurocommons project (http://www.neurocommons.org/), which aims to define best practices for semantic enrichment by using a combination of automated tools and manual annotation to add semantic tags to a small example corpus of neuroscience research articles.

The challenges involved in persuading authors to add structure to their research articles should not be underestimated, however. Whereas adding structure to a manuscript may deliver long-term benefits for future researchers, it is likely to involve an extra short-term effort on the part of authors. How can authors be persuaded that this effort is worthwhile?

There are a few key principles that may help. First and most importantly, the process of adding structured information to an article needs to be made as easy as possible. Ideally, the process of adding structure should actually be streamlined to the point that it can actually provide a net saving of effort for the author. Smart autocompletion of complex terms (such as protein, chemical, or gene names), enforcing standard formatting and spelling, while allowing disambiguation, may be one way to achieve this. Second, the payoff needs to be made as immediate as possible. If authors go to extra effort during manuscript submission, they need to receive some kind of reward immediately. For example, if an author submits clinical trial data in a highly structured way designed to facilitate meta-analysis, it would make sense to provide the author a provisionally updated meta-analysis incorporating their as yet unpublished data, following submission. Finally, the problem of capturing structured knowledge is far too big to address all at once. It needs to be approached in manageable chunks, first identifying a small amount of structured information that authors can provide reasonably easily, and gradually building from there to capture

more structured information as the tools improve and the benefits become increasingly evident.

The importance of such tools was highlighted recently by C. Sander, head of the Computational Biology Center at the Memorial Sloan-Kettering Cancer Center, who used his keynote at the ISMB bioinformatics conference in Brazil to announce the "Cashew prize," which will be awarded to the creator of the best tool for assisting authors to conveniently express the pertinent facts reported in their article, in an unambiguous computer-readable form. This is a challenging goal, but a vital one.

References

Butler D (2006) Open-access journal hits rocky times. Nature 441:914.

European Commission Directorate-General for Research (2006) Study on the economic and technical evolution of the scientific publication markets in Europe. Retrieved August 30, 2006, from http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf

Mons B (2005) Which gene did you mean? BMC Bioinformatics 6:142.

National Institutes of Health (2006) Progress report on the NIH Public Access Policy. Retrieved August 30, 2006, from http://publicaccess.nih.gov/Final_Report_20060201.pdf

Peters P (2006) Comment on: Open-access journal hits rocky times. Nature News Blog. Retrieved August 30, 2006, from http://blogs.nature.com/news/blog/2006/06/openaccess_journal_hits_rocky.html#comment-17436

Research Councils UK (2006) Research Councils UK statement on open access. Retrieved August 30, 2006, from http://www.rcuk.ac.uk/access/2006statement.pdf

Wellcome Trust (2003) Costs and business models in scientific research publishing. Retrieved August 30, 2006, from http://www.wellcome.ac.uk/doc_WTD003185.html

Wellcome Trust (2005) Wellcome Trust open access policy announcement. Retrieved August 30, 2006, from http://www.wellcome.ac.uk/doc_WTX025191.html

Willinsky J (2005) The access principle: the case for open access to research and scholarship. Cambridge, MA: MIT.