VIEWPOINT: INFORMATION ACCESS

# Building A "GenBank" of the Published Literature

Richard J. Roberts,* Harold E. Varmus, Michael Ashburner, Patrick O. Brown,† Michael B. Eisen, Chaitan Khosla, Marc Kirschner, Roel Nusse,† Matthew Scott,† Barbara Wold

Since the time of the great library of Alexandria, scholars have recognized the value of central repositories of knowledge. As scientists, we are particularly dependent on ready and unimpeded access to our published literature, the only permanent record of our ideas, discoveries, and research results, upon which future scientific activity and progress are based. The growth of the Internet is changing the way we access this literature, as more scientific journals produce online editions to supplement or replace printed versions. We urge journal publishers, their editors, and all working scientists to join together to create public, electronic archives of the scientific literature, containing complete copies of all published scientific papers.

**To encourage community dialogue about proposals that affect all of us, we focus on a controversial development in publishing, the formation of freely accessible, public archives of scientific literature. Dr. Roberts and the cosigners above have written a letter that represents one view of the issues. By agreement with Dr. Roberts, we are following it with a response of our own.**

Anyone who has spent time in a library searching for a key paper, result, or method will immediately see one of the benefits of comprehensive repositories. Those gems of information that are often buried within papers, but are not referred to in the abstract or keywords, will become readily retrievable. You will be able to locate descriptions of methods or find the original data that underlie crucial conclusions. You will be able to trace connections between observations originally scattered among many papers in different journals and databases. However, the value of central archives goes well beyond facilitated searching and retrieval. Bringing all of the scientific literature together in a common format will encourage the development of new, more sophisticated, and valuable ways of using this information, much as GenBank has done for DNA sequences.

Some have argued that central repositories are of no additional value because many journals already make their online contents freely available after some delay through their own Web sites. However, material that is freely accessible, on a controlled basis, one paper at a time, at a journal's Web site differs from material that is freely accessible in a single comprehensive collection. The latter can be efficiently indexed, searched, and linked to, whereas the former cannot. Imagine how much less useful DNA sequences would be if instead of GenBank and other global repositories, we had dozens of smaller sequence collections that could only be accessed one at a time through a genome center's Web site. Only by creating repositories with uniform, explicitly defined, and structured formats, can a dynamic digital archive of life science research literature become possible. Unimpeded access to these archives and open distribution of their contents will enable researchers to take on the challenge of integrating and interconnecting the fantastically rich, but extremely fragmented and chaotic, scientific literature.

To ensure that complete public scientific archives become a fully workable reality, the necessary infrastructure must be constructed. The National Institutes of Health has taken an important step by creating PubMed Central (PMC) (*1*) with the goal of storing the life sciences literature in digital form and providing free and convenient access, linked to the popular bibliographical database, PubMed. We envision PMC as only the first of many public archives. However, such archives will not realize their potential until they are populated. This requires that journal publishers allow their digital content to be distributed and used through online public archives. Several journals, including the *Proceedings of the National Academy of Sciences*, the *British Medical Journal*, *Nucleic Acids Research*, *Molecular Biology of the Cell*, and the BioMed Central (*2*) journals, have already agreed to deposit their content with PMC, following, at most, a short delay

VIEWPOINT: *SCIENCE*'S RESPONSE

# Is a Government Archive the Best Option?

Rich Roberts and his colleagues have constructed a thoughtful argument for an online archive of published science. A seamless way of getting access to the scientific literature is an objective many scientists have sought, and the version outlined in the Roberts piece is being pursued with vigor and understandable passion by its advocates. We admire the goal, and suspect that evolutionary forces may be moving us toward it. We have decided to make our own back research reports and articles freely available after 12 months—at our own Web site—later this year.

The specific proposal of Roberts *et al.* goes further. It urges our readers to sign a petition that "advocates the free and unrestricted distribution of scientific literature 6 months after publication." Actually, the petition does quite a bit more than that. It urges an economic boycott: signers agree not to submit papers to, review for, or subscribe to journals that do not submit to the petition's proposals. To begin a conversation among scholars with a threat of economic boycott is unfortunate.

However, we would rather focus on the qualities that Roberts *et al.* believe are essential to the archive they advocate. It should include all scientific papers and the content should be in a common format that allows for advanced search capabilities. Content should be free and "open distribution" should be allowed. PubMed Central (PMC) is given as the model of an archive that will meet these criteria. We believe other alternatives exist that can meet most of these goals faster and more effectively without putting nonprofit scholarly publishing at risk.

There already are multiple-journal sites—for example, the nonprofit HighWire Press (HWP), which archives over 230 journals, including biological, physical and interdisciplinary papers. More than 200,000 articles are freely available at this site. By comparison, there are only about a dozen journals at PMC, limited currently to biology.

Advocates of PMC argue that sites in which each journal is archived separately are insufficiently integrated. But searching across multi-journal, full-text repositories is already possible at sites such as HWP. In addition, 60% of this content is in a common format already. Why not begin with the already populated venue

after print publication. Publishers now have a wonderful opportunity to reinforce their partnership with the scientific community by supporting extant archives like PMC and by allowing archival material to be freely used and distributed, and we strongly urge them to do so. It would be natural and simple for journals that have already decided to make their back issues freely accessible at their own Web sites to make the same content available in electronic archives. The costs of participating in open archives would be minimal and would be more than offset by the benefits their participation would bring to the scientific community.

Historically, publishers have left the job of archiving to the libraries. Library archives have become more accessible as we have moved from indexed abstract books to rapidly updated online abstract searching tools. Public online archives should be viewed as the logical continuation of this tradition and, thus, as a complement to the publisher's normal activities. For electronic archives to assume this role fully, decades of volumes that currently exist only in printed form will need to be digitized. We do not expect journals to bear the cost of the digital conversion of their printed archives. Indeed, efforts to raise the necessary funds are under way, so that digital conversion of archival volumes can proceed rapidly.

It is important not only that PMC succeed, but also that other institutions be encouraged to provide independent online sites for the distribution and use of the same comprehensive archives. Multiple independent online sites will help ensure ready access for users around the world and will guarantee that no single government or institution can control access to our common scientific heritage. This diversity will also foster innovation in the ways the material in the archives is used.

We feel that if journal editors and publishers were to poll their authors and readers, they would find overwhelming support for such archives. The strength of this support is demonstrated by the growing list of scientists who have signed an open letter (*3*) advocating free and unrestricted distribution of scientific literature 6 months after publication. We urge our colleagues, especially students and the younger members of the scientific community, to make your views heard. If these efforts are successful, in 10 years, everyone's ability to do science will have been greatly enriched, and we will all wonder how it was possible to work without such archives.

**References and Notes**

1. www.pubmedcentral.nih.gov
2. www.biomedcentral.com
3. www.publiclibraryofscience.org

R. J. Roberts, New England Biolabs, Beverly, MA 01915, USA. H. E. Varmus, Memorial Sloan-Kettering Cancer Center, New York, NY 10021, USA. M. Ashburner, University of Cambridge, CB2 3EH, UK, and EMBL–European Bioinformatics Institute, Cambridge, CB10 1SD, UK. P. O. Brown, Stanford University School of Medicine, Stanford, CA 94305, USA. M. B. Eisen, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, and University of California, Berkeley, CA 94720, USA. C. Khosla, Stanford University, Stanford, CA 94305, USA. M. Kirschner, Harvard Medical School, Boston, MA 02115, USA. R. Nusse and M. Scott, Stanford University School of Medicine, B. Wold, Biology Division, California Institute of Technology, Pasadena, CA 91125, USA.

*To whom correspondence should be addressed. E-mail: roberts@neb.com

†Howard Hughes Medical Institute investigator.

and add the integration, rather than the other way around? Why not use taxpayer dollars to promote innovative search technologies that do not require taking control of services provided by the private sector?

The proposition of Roberts *et al.* raises problems for *Science*, and for other journals. First, it will reroute an economically important source of online traffic for journals that offer content and other products on their sites. Second, unlimited redistribution of content could lead to misuse of content and loss of quality control. Third, it may expose users to risks historically associated with monopoly suppliers. For example, recently PubMed—on which PMC will depend—unexpectedly failed to process new content for over a month, inconveniencing authors and publishers.

We also wonder whether enough attention has been given to some of the economic issues. Experience shows that demand for scientific papers drops to about 1/10th within 4 to 5 months, but then continues at a low level for years. We plan to track our experience with free back issues carefully, but in the meanwhile, we take little comfort from the assurance that "costs of participation in open archives will be minimal." Subscription and advertising revenue will be at some risk and transferring primary access to someone else's site may expose us to further losses. The value we add—through peer review, perspective and context-setting analysis of research, and good news coverage—requires revenue support from advertising. Moreover, *Science* supports other activities of AAAS—including science and public policy, kindergarten through 12th-grade education, a career-mentoring Web site for young scientists, and innovative "knowledge environments." These benefit scientists from all fields. Posting our back content on a site that primarily serves biomedical scientists would confer a benefit on one group by taking benefits away from another—creating, in effect, a transfer payment from the sciences in general to biology in particular. That bothers us.

We worry, too, about another group of journals that will be entering a riskier environment. Our association is an umbrella organization, including many specialized scientific societies as affiliates. Their more focused journals must remain viable to ensure continued publishing options in highly specialized fields and for younger scientists. In most cases, academic library subscriptions provide the economic "floor" that guarantees financial sustainability. If papers from specialized journals were to become available on the PMC site, budget-conscious library directors would be tempted to cancel subscriptions. Some of the signers of the petition are scientists who belong to those very societies. Have they considered that their initiative will put PMC in competition with their own journals? When tax-exempt organizations go into competition with commercial entities they must pay unrelated-business income tax. When tax-supported organizations compete with commercial entities and nonprofits, the public has usually raised strong objections.

There are also questions about whether the proposed location for PMC—the National Library of Medicine, part of the National Institutes of Health—is the right one. NIH already sponsors, through its extramural programs, much of the biomedical research PMC will archive. It regulates the conduct of that research, controls much of the training of the next generation of researchers, and archives primary data. It now proposes that the results of the research it funds be given over by publishers and authors to a server subject to its exclusive control. The Congress or the President can eliminate support for certain kinds of science and have done so in the past. Would PMC then be able to archive papers on those subjects? Concentrating this kind of womb-to-tomb control in a single federal agency has risks, and we should ask whether we are entirely comfortable with a state-run, centrally managed economy in biomedicine.

Proponents of this plan include scientists of high reputation: Nobel laureates, leaders of institutions, and others whom we all admire. Nonetheless, we think its potential consequences require careful analysis and policy debate. We at *Science* are determined to participate in a constructive spirit.

**THE EDITORS**