For Science. FINAL.

Building a "GenBank" of the published literature

Richard J. Roberts, New England Biolabs, Beverly, MA, USA
Harold E. Varmus, Memorial Sloan Kettering Cancer Center, New York, USA
Michael Ashburner, Department of Genetics, University of Cambridge, UK and EMBL - European Bioinformatics Institute, Cambridge, UK
Patrick O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, USA
Michael B. Eisen, Life Sciences Division, Lawrence Orlando Berkeley National Lab and Department of Molecular and Cell Biology, University of California, Berkeley, USA
Chaitan Khosla, Departments of Chemistry and Chemical Engineering, Stanford University, USA
Marc Kirshner, Department of Cell Biology, Harvard Medical School
Matthew Scott, Departments of Developmental Biology and Genetics and Howard Hughes Medical Institute, Stanford University School of Medicine, USA
Barbara Wold, Biology Division, Cal Tech, USA

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 and more scientific journals produce online editions to supplement or replace printed versions. In this letter 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.

Anyone who has spent time in a library searching for a key paper, result or method will immediately see one of the benefits of such comprehensive repositories. Those gems of information that are often buried within papers, but 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, it is crucial to understand the important difference between material that is freely accessible, on a controlled basis, one paper at a time, at a journal's web site and material that is freely accessible in a single comprehensive collection. The latter can be efficiently indexed, searched, and linked to, while 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 collections of sequences each of which could only be accessed one at a time through a genome center's website. Only by creating repositories with uniform, explicitly defined and structured formats, can a dynamic digital archive of the life science research literature become possible. Unimpeded open distribution of the material in these archives will enable researchers to begin to take on the challenge of integrating and interconnecting the fantastically rich but extremely fragmented and chaotic scientific literature.

How can we ensure that complete public scientific archives become a fully workable reality? Clearly, the necessary infrastructure must be constructed. The National Institutes of Health has taken an important step by creating PubMedCentral (PMC; http://www.pubmedcentral.nih.gov) 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 tthe 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 PNAS, the British Medical Journal, Nucleic Acids Research, Molecular Biology of the Cell

and the BioMedCentral journals, have already agreed to deposit their content with PMC, following at most a short delay after print publication. Publishers now have a wonderful opportunity to reinforce their longstanding and productive partnership with the scientific community by acting to support 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 websites to make the same content available in electronic archives. For other journals, 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 progressively 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 the journals to bear the cost of the digital conversion of their printed archives. Indeed, efforts to raise the necessary funds are already underway so that digital conversion of archival volumes can proceed rapidly once publishers agree to allow the digitized articles to be freely distributed in public electronic archives.

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 sites that provide access to the comprehensive archives will help ensure ready access for users around the world and guarantee that no single government or institution can control access to any part of our common scientific heritage. This diversity will also foster innovation in the ways the material in the archives is used and allow scientists to apply their creativity and energy toward making this huge information resource more valuable and accessible.

We feel sure that if journal editors and publishers were to poll their customers, the authors and readers, they would find overwhelmingly support for comprehensive open literature archives. The strength of this support is demonstrated by the growing list of scientists who have signed an open letter (see http://www.publiclibraryofscience.org for the text of the letter and a list of signatures) that advocates the free and unrestricted distribution of scientific literature six months after publication. We urge our colleagues, especially students and the younger members of the scientific community, to make your voices and your views heard. If these efforts are successful, in ten years everybody's ability to do science will have been greatly enriched by ready access to the full record of the world's scientific research, and we will all wonder how it was possible to work without it.