nature webdebates

Digital archives: how we can provide access to 'old' biomedical information

Richard R. Rowe
Chairman, RoweCom
rowe@rowe.com

The debate over free scientific information often revolves around access to current information that is central to human welfare such as human genome discoveries, AIDS research and genetically modified organisms. However, archival services, providing unfettered access to the terabytes of 'old' literature that constitute the foundations of today's knowledge, increasingly present their own challenges.

Archiving information cost-effectively has become a major burden for those institutions that have shouldered this important and mammoth responsibility. As the volume of information continues to rise, the sheer cost of archiving and providing access to those archives is becoming unsustainable.

The challenge is even greater with regard to e-publications for which there are no established archiving traditions, where the formats themselves are unstable and likely to change over time and where the resource itself seems less tangible. The increasing quantity and velocity of information gushing from the digital fire hose is beyond any one institution's ability to capture, sort out and store, let alone make accessible. In addition, current copyright laws continue to limit legal access to archival literature.

Our current systems of knowledge preservation no longer satisfy our needs or fit our resources. In order for digital information to have lasting usefulness, we must address two basic questions: What should we keep? How can we make it accessible? Technical challenges abound regarding the preservation of and access to digital archives that increasingly involve multiple media. But technical solutions necessarily rest upon the professional, economic and public policies that affect the archiving of information. We must respond creatively to the opportunity presented by these policy challenges.

There are several models designed to address the problem of what biomedical information should be stored and how to make it economically accessible:

Publisher-based solutions.

Historically, most publishers have not archived their material in an accessible manner – that has been the role of libraries. The additional cost to publishers of archiving their publications would constitute a significant financial burden that would have to be paid by someone and, finally, there is no guarantee that any publisher will exist in perpetuity. In at least one case a publisher has asked a university to maintain its archives. However, if one accepts the proposition that knowledge should be organized and preserved by discipline rather than by publisher, a maximally useful archive will have content from multiple publishers. Thus a publisher-based solution has limited value.

Megacentres.

There is much discussion about building large, centralized digital archives. This is a logical extension of the 'national library' concept. The International Congress of Scientific and Technical Information (http://www.icsti.org) has explored digital archiving issues (http://www.icsti.org/icsti.forum/33 and http://www.icsti.org/icsti.forum/35) and has given particular attention to this model. JStor (http://www.jstor.org), an ambitious and popular non-profit journal-archiving initiative, is another effort to create a 'megacentre'. The US National Library of Medicine (NLM) is in many respects a de facto biomedical megacentre. This model is often deeply dependent upon government or foundation funding. Because the value of information is almost purely 'contextual', this model has serious limitations as a truly comprehensive archive for any particular biomedical discipline, particularly in the digital world. What is important for one field of inquiry may be irrelevant to another and vice versa. The decision to 'keep' or 'throw away' should be done within the framework of a particular body of knowledge. Furthermore, the absolute quantity of information that today is being generated on a global basis will soon be far beyond the capacity of any one organization, even the NLM, to maintain effectively. We must learn how to share the responsibility of maintaining and providing access to the world's knowledge.

Networked collaborative model.

This model relies upon libraries, who from their beginning have maintained print archives, to extend their archiving functions to digital resources. It calls for libraries to share the task of maintaining accessible archives of important knowledge in a collaborative global network. Library centres of excellence throughout the world currently have in-depth archives in their fields of specialization. In many cases they could readily provide these archival services globally. There could be multiple, redundant library centres for the most important areas of biomedical knowledge. These services could be available through the Internet on a paid subscription basis, thus enabling the archival centres to offset the costs of maintaining high-quality services on behalf of the world. Such services would provide access to both the print and the digital resources of a particular discipline.

A global biomedical archives consortium (BMAC) could have several co-ordinating functions. The consortium could:

- establish technical and organizational standards for global biomedical archive centres;
- identify, certify and de-certify specialized centres of excellence;
- maintain 'optimal redundancy' in the global network of biomedical archives;
- provide a meta-catalogue of global biomedical knowledge resources.

The BMAC would be an open consortium supportive of the numerous existing and planned initiatives for providing widespread online access to digital biomedical archives. It should be global in scope and could be an independent body, or could work under the auspices of an existing international body such as the World Health Organization.

With BMAC in place, institutions could provide their members with unfettered access to the fundamental bodies of biomedical literature by subscribing to those centres that maintain the archives to which the institution needs access. Medical schools, research centres and hospitals would no longer need to use valuable space and time storing – at considerable cost – old publications 'just in case' they might need access to them. The subscription revenues generated by this model would enable the centres to improve their services, including using the most advanced technologies to provide hyperlinks to related information. In some cases, public subsidies would be needed in order to increase the quality and quantity of the service and to provide unfettered access to users. However, this model would, to a significant degree, reflect the supply-and-demand characteristics of biomedical research and practice.

In this fast-moving world, information ages quickly and the aggregate demand for 'old' information

is relatively small. It therefore seems reasonable to pool resources in order to preserve our growing knowledge archives. To do so, we must modify our rules and practices concerning "old" intellectual property, and allow ageing knowledge to devolve into the public domain. Authors and publishers should be able to agree upon a limited period for copyright of biomedical research literature, after which scientific information would become a public resource.

There is increasing support for this concept. Recently, interest has focused on what is an appropriate period for copyright protection of scientific knowledge. The Public Library of Science has suggested six months. Some publishers have argued that one fixed time period is not appropriate because the "half-life" of scientific articles varies substantially from one field to another. The issue is primarily an economic one. My impression is that today virtually all of the revenue of publishers comes from the original sale of their journals. The argument that many institutions would simply wait six months to get access to a biomedical article rather than continuing to purchase journals as they do now is questionable. Timeliness is far too important a variable these days. It would be useful to analyse by discipline the use patterns over time of digitally published articles. Some facts need not be debated. Our common goals should be to make scientific information readily accessible to all those who need it within a reasonable period after it has been published and to do so in a manner that does not seriously undermine the financial base of scientific publishing.

No matter what the length of the original copyright may be, we need to begin using the copyright expiration date, as a standard eight-digit XML tag attached to each scientific article, to enable us to know when any particular article is unfettered by copyright. Authors could attach a simple "copyright expiration provision" as a condition of their assignment of copyright to publishers, in order to ensure that their articles would be in the public domain after the agreed-upon period.

An open meeting exploring the collaborative model for biomedical archives described above, with representatives of professional societies, publishers and libraries participating, is being held on Sunday 19 August (2"5 pm) at the MIT Press in Cambridge, Massachusetts, coinciding with the International Federation of Library Associations meetings that week in Boston. Go to www.biomedarchives.org for more information on this meeting.

Richard R. Rowe, is chairman and chief executive officer of RoweCom, Inc., an online international library subscription service. Formerly president and chief executive officer of the Faxon Company, Rowe was associate dean at the Harvard Graduate School of Education and was a member of the Massachusetts State Board of Education. He currently serves on the management board of the MIT Press, the board of visitors of the School of Information Sciences at the University of Pittsburgh and is chairman of the Massachusetts Business Alliance for Education.