nature webdebates

Tailoring access to the source: preprints, grey literature and journal articles

Walter Warnick
Director
The Office of Scientific and Technical Information (OSTI)
US Department of Energy

Any discussion of free on-line access must give separate consideration to the three main types of primary scientific and technical literature: preprints, 'grey' literature and journals. Each has specific characteristics that strongly shape how they can be accessed. Preprints are posted directly by their authors, who presumably welcome free full-text access, although many of the sites do not provide metadata to help identify, describe and retrieve preprints. Grey literature is usually produced by research institutions and posted on sites that compile full-text databases, and provide detailed metadata and search tools. Accessing journal literature requires dealing with the copyright claims of publishers.

Because government plays a major role in sponsoring the R&D that generates much of the scientific and technical literature, it has special responsibilities to collect, preserve and disseminate this literature in relevant disciplines. Broad responsibilities have been codified in the laws under which the US federal R&D agencies operate. Federal agencies are well positioned to develop tools that facilitate discovery in the primary literature and, to the extent that such tools are successful, scientific communication is already being greatly improved, even revolutionized.

Preprints

There are around 7,000 scientific and technical preprint sites, although most of these lack formal data structures, such as metadata, to aid in the identification, description and retrieval of preprints. Perhaps the best known of the sites is Paul Ginsparg's Los Alamos e-Print Archive. The PrePRINT Network search engine can be used to cross-search 4,000 preprint sites and the remaining 3,000 will be included by the end of 2001. Altogether, around 375,000 scientific and technical preprints can be searched via the PrePRINT Network and users can freely access and search the full text of this type of primary literature. The universe of preprint sites continues to expand, but the pace of growth is manageable.

The PrePRINT Network can crawl preprint sites or pulse the search engines of repositories without placing any meaningful burden on the owners. In contrast, the Open Archives Initiative (OAI) relies on standards that site owners must comply with, and is therefore better suited to formal repositories. If OAI is widely adopted by site owners, the PrePRINT Network would also have to adopt this standard or be supplanted. The extra work required for compliance would pay dividends in increased power in searches across preprint sites.

Grey literature

Currently, over 100,000 recent scientific and technical reports by the US Department of Energy, the Department of Defense, the Environmental Protection Agency and the National Aeronautics and Space Administration have been integrated into a virtual single product called the GrayLIT Network. To my knowledge, this is one of the few Internet tools that gives users on-line full text and searchable access to a virtual compilation of this type of primary literature.

Further grey literature collections, once identified, can be added to the GrayLIT Network with the investment of a few days of effort. Before grey literature collections can be considered comprehensive, however, the major obstacle of vast quantities of legacy material that reside in non-digitized formats at federal agencies must be overcome. I am not aware of any agency that is systematically digitizing its legacy collection, so for the foreseeable future only metadata about such legacy collections, but not the full text, will be available on-line.

Journal literature

It is technologically feasible to build a system whereby students and researchers can quickly conduct a journal literature review via the Internet. As shared knowledge is the enabler of scientific progress, this would be a tremendous boon. Two necessary technological milestones have already been largely overcome. First, journals must be available in full-text electronic formats. Many publishers of traditional paper journals have made great progress in the transition to electronic formats, with some also making notable progress with their legacy collections. In addition, new electronic-only journals are emerging.

Second, a system must be available that allows users to identify articles of interest using cross-publisher searches, preferably of abstracts, and that system must include hyperlinks to the full text. New systems have been created and are evolving rapidly to allow cross-publisher searching and hyperlinking to full text. A thousand flowers are blooming: some systems target the individual student or researcher, whereas others are geared to institutional subscribers; some are free to the user, whereas others require that the institutional subscriber pays start-up fees of hundreds of thousands of dollars; some are narrow in the scope of disciplines covered, whereas others are broad; some are comprehensive within their disciplines, whereas others are less so; some offer any of a wide variety of value-added features, e.g., citation counting, targeted to specific user communities; some host abstracts, whereas others focus on authors and titles; some re-host the full text, whereas others link to publishers' servers.

A description of this wide array of systems available is beyond the scope of this article. As the topic of this forum is 'free' on-line access to literature, it is appropriate to focus on three of the many US systems that begin to meet this criterion. These systems also differ in many other ways from other systems.

For medical and biological sciences, the National Library of Medicine (NLM) hosts two products, PubMed and PubMed Central. PubMed hosts citation information, including abstracts, and provides hyperlinks for collaborating publishers that takes the patron to the full text at the publisher's server at no cost to the user. PubMed Central does this, too, except the publisher must agree to submit the full text to NLM within one year and that the full text will be made freely available by that time at no cost to the user.

For the physical sciences, the Department of Energy hosts PubSCIENCE, which is similar in design to PubMed. The hyperlinks in PubMed and PubSCIENCE are automatically live if the user or his institution has an electronic subscription or site license to the journal. A full-text product for the physical sciences comparable to PubMed Central does not exist. Integration of PubMed, PubMed Central and PubSCIENCE, as well as any other cross-publisher search product, into a single tool that would operate via a single query, an easy task with a directed query engine, would seem to be a significant opportunity to use the Internet for the benefit of science.

Avoiding copyright problems

Achieving all of the technological milestones is not enough. Institutional arrangements with primary publishers who own the content copyright are also needed, but are incomplete. Two types of content need to be considered separately: abstracts and full text.

Considerable progress has been made with institutional arrangements whereby publishers make abstracts freely available for use by PubMed or PubSCIENCE. Although the hyperlinks in PubMed or PubSCIENCE are not comprehensive, PubMed now has hyperlinks for approximately 1,800 journals, and PubSCIENCE has hyperlinks for approximately 1,000 journals, including *Nature* and a number of other leading titles. Similar arrangements with publishers are beginning to materialize for 'free' full text for PubMed Central.

Thus, freely available life science products have advanced farther than physical science products. The Open Letter of the Public Library of Science (PLS) initiative focuses on full-text material in medicine and the life sciences and publishers who own copyright are being asked to grant distribution rights to full text in these disciplines. In contrast, much progress could be made in the physical sciences without raising issues about copyright.

The most immediate challenge to making literature searches in the physical sciences more available on the Internet is to complete collaborations with publishers for citations, including abstracts. In important instances, consummating such arrangements has been problematic for PubSCIENCE. Complicating the situation are the prior arrangements that some primary publishers have with secondary publishers. A key to meeting the challenge may be to devise a way to work constructively with secondary publishers; for example, hyperlinks in PubSCIENCE could deliver the scientifically attentive citizen to the doorsteps of secondary publishers. This may stop short of the PLS goal of having the complete text of scientific articles freely available in searchable interlinked formats, but at least citizens would have access to more article titles, authors and perhaps abstracts than they have now.

To be more precise, an immediate goal in the physical sciences is access to a comprehensive searchable index of abstracts. We do not need the abstracts themselves, only the ability to create an index of abstracts together with hyperlinks to the abstracts, hosted by primary or secondary publishers. That is an immediate challenge to using the Internet to benefit the physical sciences.