icsti forum
Quarterly Newsletter of the International Council for Scientific and Technical Information
N°26, November 1997

The Role of A&I Services in Facilitating Access to the E-Archive of Science

by Maureen C. Kelly
BIOSIS

Fast or slow, observed or unnoticed, change is ubiquitous and persistent. It just happens. It does not ask permission. It is only marginally responsive to planning. Most of the time it leaves us scrambling in its wake, trying to discern and manage the challenges and opportunities it presents.

The change from printed to electronic dissemination of scientific information is typical. This has not been an orderly progression. As so often occurs, we have sought first to take advantage of the opportunities, leaving for later the resolution of difficulties. One such difficulty that now confronts us is how to ensure the availability and accessibility of a reliable archive for the growing corpus of electronic information.

This challenge requires more than an arithmetic progression of the effort employed in maintaining an archive for printed information. The difficulties associated with maintaining an accessible e-archive lie in part, but only in part, in the rate of growth of scientific information. That difficulty is compounded by the rate of change for both electronic content and medium. Electronic information does not stay put in the same way we have come to expect of print. Computer systems and software change, requiring changes in the databases they support. In addition, the content in these databases can be revised, amended, and amplified by the scientist in ways unknown to the print world. But the difficulty does not end there; problems of volume and volatility are not sufficient to describe the challenges ahead. The changes that are underway affect the very nature of the scientific research and communication process and the artifacts by which it is documented. These changes have major consequences for both the archive of scientific information and the organizations that build, maintain, and facilitate access to that archive.

The Importance of the Archive

It is widely agreed that the archive of science is essential to the progress of science. The tradition of collecting and storing written information in a centralized repository is over 2000 years old. An archive makes it possible for new research to build on the research of the past. But for the archive to function successfully, it must be both stable and accessible. As observed by Dr. Robert Hayes (*Science and Technology Libraries*, Summer 1992): "We acknowledge our debt to the past by citation to it. By doing so, we assure that our sources can

be checked, verified, validated. But that implies that the material so referenced, so cited must be available for checking, verifying, validating." Dr. Hayes goes on to observe, "What happens if the source data [is electronic and] has been erased or, worse yet, altered since it was used? The entire structure of scholarly progress would collapse."

Scientific Communication in a Print-Only World

The traditional archive for published scientific information has been the university library. In the print-only world, these libraries held collections of journals and books in a network of distributed, overlapping resources. Union lists, card catalogs, and interlibrary loan arrangements helped the user locate and acquire specific articles. But these tools were not particularly useful for identifying which articles pertained to the research question at hand. This is where abstracting and indexing publications were used. By translating a query into the vocabulary of the A&I index, the users could navigate large aggregations of articles in their subject domain. The *lingua franca* of these two resources (the library tools and the A&I tools) was the bibliographic citation - a standardized expression intended to uniquely identify each journal article or book chapter.

Within this print-only world, communication was essentially one-way, from the author to the reader of the article. Comments from the reader might take the form of letters that could appear in subsequent issues of the journal, but the original publication did not change to reflect these comments; its contents were stable. The basic tokens of information were words and, to a lesser extent, illustrations. The primary information containers were journal articles and books. And the identifiers for these containers were bibliographic citations. While the standard format for a bibliographic citation might vary across disciplines or publishers, its contents were sufficiently consistent to serve as a reliable identifier for the article.

The Transition to Electronic Information

A&I tools were among the first information resources to become available in electronic form. Some 25 years before the Internet became commonplace in scholarly communication, A&I publishers began to distribute their citations and indexes as databases that could be searched by computer. For most of that time, however, the primary scientific literature remained in print form only. Gradually (and cautiously), electronic versions of printed journals began to appear. In limited cases, publishers made entire collections of their journals available electronically in image form.

The user could post a query to an electronic database supplied by an A&I service, a primary publisher, or a document supplier. The result of such a query might be either a citation and abstract or a copy of the full article (printed from an electronic image). In many cases, the user would still rely on the library to retrieve a printed version of the articles of interest. Libraries remained the primary archival resource for scientific information, and their collections reflected this role. Where a library might be compelled to drop a subscription, it would likely be backed up by another library in its interlibrary loan network.

During this transition phase, communication remained one-way and content remained stable. While users might interact with the databases, they did not contribute directly or dynamically

to the contents. Words and illustrations continued as the basic tokens of information; journal articles and books remained the primary information containers; and bibliographic citations continued to be the identifiers for these containers. It is noteworthy that, during this period, attention was given to developing standard identification codes for these citations. The quest for a unique, coded identifier for bibliographic citations continues today, with work underway on both proprietary and open solutions.

Scientific Communication in an Internet-Centric World

With the advent of the Internet, scholarly communication has embarked on a period of turmoil and change. Eli Noam (*Science*, 13 October 1995) characterized scholarly activity as consisting primarily of three elements: 1) the creation of knowledge and evaluation of its validity; 2) the preservation of information; and 3) the transmission of this information to others. All three elements are now undergoing fundamental changes.

The changes that the Internet has enabled in the transmission of scientific information have not been especially orderly, but they give evidence of a pent-up desire for rapid exchange, dynamic content, and 2-way interaction (all attributes of an earlier, oral tradition). No longer does the user need to go to the library for information; rather, with the Internet, the information comes to the user. There is, of course, a price to be paid for this convenience. Whereas the printed word was reliably fixed to the page, the content of the Internet is volatile. Whereas archives of printed journals and books could tolerate a degree of benign neglect, an archive of Internet content will require more active attention. And whereas libraries have been constructed and funded to archive printed information, they are not so well equipped to function as archives for scientific information on the Internet.

The Internet web page and web site are elegant in their simplicity: relatively easy to construct and very easy to access. The hypertext link has proven to be of particular value to the users of scientific information. It allows a user to continue a path of exploration without stopping to locate the needed resources. One click on the magic blue words, and you can be transported to a site around the globe. Combine hypertext linking with the various available search engines and web crawlers, and it becomes possible to satisfy a whole array of information needs right from your desktop.

Unlike earlier forms of electronic communication, the Internet is affecting more than just the mode of distribution. It is also affecting the content. The basic tokens of information content have expanded beyond words to include images and sounds, data and metadata. This information is packaged for distribution and access in the form of web pages and web sites.

Information on the Internet is identified by the "site" on which it resides rather than the "cite" that describes its contents. There is more than a letter's difference between these two approaches. The URL that appears at the top of a Web page (or behind a hypertext link) identifies the location of that page rather than its contents. This is a break with the long-standing print tradition of identifiers that describe content rather than location. Scientists have an expectation that a citation to a printed journal article will yield the same contents whether the journal is retrieved from a library at Harvard or Princeton. The user of a bibliographic "cite" presumes that the content it identifies is stable and unchanging. This is

not a presumption that can be made of the content identified by web "sites". The problem goes beyond the need for URLs that are persistent over time. The Internet is in need of content identifiers: identifiers that are globally unique and independent of physical location and that ensure stability of contents. Lacking this, it will not be possible to build reliable bibliographies, archives, or A&I databases.

The Diversification of Scientific Information

Journal articles and books have long been the accepted containers for conveying scientific research results through the dimensions of space and time. Library archives have been built to house them. A&I databases have been designed to inventory, organize, and facilitate access to them. The academic reward system of promotion and tenure revolves around them. Simply put, they are important to the process and progress of science. Why, then, should we question their continued primacy in this new era of electronic communication?

For all their importance, neither print journal articles or books nor their electronic equivalents are sufficient to the task ahead. If we are to succeed in building an e-archive that serves science, we must look beyond the current, text-centric paradigm. The results of scientific research are increasingly complex and multi-dimensional. Bibliographic and textual forms of information are no longer adequate. Today's scientists require access to data in numerical, symbolic, and image form; they also need mechanisms for sharing computational models and simulations along with other collections of functional information. In their current forms, neither journal articles nor web sites are sufficient to this task; the future requires a more versatile container for publishing and archiving scientific information. This new container must accommodate a variety of forms and types of data and information, all the while retaining the full power and utility of the contents. This container must both convey knowledge and enable the discovery of new knowledge. Such containers can be thought of as knowledge objects or KNOBs, reflecting this focus on content and functionality.

A knowledge object might contain an entire database or extracts from several databases pertaining to a specific research topic. KNOBs could be used to contain such diverse contents as meta-analyses, research protocols, DNA sequences or gene maps, computer simulations of protein folding or forest dynamics, a century's worth of ocean temperature data linked to geospatial coordinates, or the electronic version of a journal article. A KNOB is an adaptable container for publishing and archiving information about the process and results of scientific research. Ideally, a KNOB has a conceptual integrity and research purpose. Each KNOB should have its own unique identifier and should carry structured metadata to characterize its contents. It should also carry (or point to) information about the rights and restrictions that govern its use.

While KNOBs, *per se*, may never come to be, they are useful for conceptualizing the challenges that lie ahead. They can help us understand what is needed for an e-archive to effectively and efficiently support the changing process of scientific research and communication. KNOBs can also help us visualize future roles for the organizations that build, maintain, and facilitate access to that archive.

Plan though we might, however, the e-archive will evolve as much in response to business opportunity as to the needs of science. After all, even the "dot-orgs" and "dot-edus" of the world have to be attentive to the bottom line. The e-archive will undoubtedly be distributed, much like the Internet that spawned it. Even though this cyberspace "library" will contain a bewildering variety of information types, it will have "shelves" that are bare or sparsely populated.

Fortunately, where commerce goes, standards eventually follow, and these standards will help to bring structure and interoperability to the collective resource. These standards take on special importance because it is improbable that the e-archive will be built as a cohesive whole following rigorous and consistent specifications. We need to be attentive so that the standards serve science as well as commerce.

Tomorrow's information users will navigate through varied suppliers of services, tools, and content en route to the information they seek. Without giving it much attention, they will make use of communication and transaction processors, rights managers, billing agents, and query processors, as well as concept locators, physical locators, and content warehouses.

It's risky to predict what sectors of today's information infrastructure will remain, or what roles the players will fill. A&I services are finding themselves in competition with their information suppliers (both authors and publishers) as well as with an ever-smarter collection of web crawlers and search engines. The need for conceptual ordering of information, the traditional role of the A&I service, will remain. However, that order may be achieved on-the-fly in response to a query, making use of knowledge tools rather than pre-ordered collections of information.

To secure a viable and vital place in the new information infrastructure will require shrewd analysis of core competencies, competitive strengths, and business opportunities. Every problem represents a business opportunity. Science will require knowledge organizers for its knowledge objects.

Ms. Maureen C. Kelly
Vice President, Document Analysis Division
BIOSIS
2100 Arch Street
Philadelphia, PA 19103-1399, USA
mckelly@mail.biosis.org

http://www.icsti.org/forum/26/index.html