

## Access to scientific literature

Steve Lawrence

NEC Research Institute, 4 Independence Way, Princeton, NJ 08540

lawrence@research.nj.nec.com

The greatest impact of the web does not come from the availability of more information. A lot of information has long been available, for example in libraries or by asking the right person. Rather, the greatest impact is due to the dramatic improvement in the convenience of accessing information. Scientists now have almost instant access to a large and rapidly increasing amount of information that previously required trips to the library, inter-library loan delays, or substantial effort in locating the source. Evidence shows that usage increases when information access is more convenient<sup>7</sup>, and surely we want to maximize the use of the scientific record.

Many scientists have free access to a significant subset of the literature through libraries, however such access is degrading in value over time. The fraction of scientists with access to research libraries is decreasing, as is the fraction of the literature available in the libraries<sup>7</sup>. Perhaps more importantly, accessing printed literature in libraries is becoming less attractive as more literature becomes freely available on the web, and the reward for expending the extra effort to visit a library diminishes. An increasing percentage of researchers (particularly students) rarely use libraries, preferring to access research that is instantly available online<sup>5</sup>. Moreover, libraries focus on journals, and research is increasingly being made available on the web long before appearing in journals (in preprints, technical reports, and conference papers, for example). Additionally, few individuals can afford subscription access to much more than top-tier journals such as *Nature* and *Science*. What is troubling is that a significant fraction of research in higher impact journals is unavailable for free, meaning that researchers are increasingly becoming biased against locating some of the higher quality research.

There have been many proposals and attempts to provide widespread access to scientific literature on the web. For example, Cameron proposed a universal bibliographic and citation database that would make all published research available and searchable by anyone with web access, complete with citation links between papers<sup>3</sup>. Cameron's proposal requires authors or institutions to submit article and citation metadata in a specific format, and is yet to be implemented. While Cameron's vision remains on the drawing board, others have built systems. However, most have experienced limited success to date. The most successful traditional service has been the Los Alamos e-print archive, now called arXiv (*arxiv.org*). arXiv contains about 140,000 free full-text research papers, of which over 50% are in Astrophysics or High Energy Physics. Similar services, providing access to far fewer articles, include BioMed Central (biology and medicine), PubMed Central (life sciences, refereed journal articles only), RePEc (economics), CoRR (computing), and CogPrints (cognitive science).

Although posting of research papers on the web varies by discipline, there are many more research papers freely available on the web than those contained in arXiv and similar services (on the homepages of authors or institutions, for example). The ResearchIndex service at NEC (*researchindex.org*) is the world's largest free full-text archive of scientific literature. ResearchIndex contains over 400,000 papers, currently focusing on computer science and related literature that is freely available on the web. Why then do the other online archives not contain many more papers? Two probable factors are the requirement for substantial effort on the part of authors or institutions, and the lack of sufficient incentives for authors to make their articles

available in the archives.

In general, overhead limits usage, with excessive overhead effectively preventing usage. An interesting analogy is the web itself, arguably one big ad hoc, decentralized mess, littered with dead links, and lacking built-in support for features such as content indexing and access payments. These deficiencies are in principle solvable, and indeed proposals for hypertext systems without these deficiencies existed long before the web (e.g., Xanadu<sup>6</sup>). However, the reality of designing, implementing, and participating in more idealized hypertext systems, namely greater overhead for designers and participants, has prevented the widespread success of such systems.

Most online research archives have made more idealized design choices at the expense of greater overhead for participants. Authors are required to enter large amounts of information and submit articles in specific formats, and features such as reference linking are only performed if the author formats references as required by the system. In contrast, ResearchIndex aims for a more practical system with minimal overhead for participants. Indeed, while many authors submit articles to the system, this is not even necessary – ResearchIndex automatically locates papers archived on the publicly indexable web. ResearchIndex does not require authors to enter metadata for their articles, automatically extracting title and author information, citations, and citation context (and allowing authors to make corrections or enter additional information).

In addition to imposing too much overhead, most online archives lack many incentives that can promote growth. Although many articles are already freely available on the homepages of authors or institutions, the archives start from scratch, forcing authors to duplicate effort and manually submit articles. By indexing papers already on the web, ResearchIndex has a better chance of obtaining a critical mass of freely available articles, which leads to greater use, and in turn leads to a greater incentive for authors to have their articles available in the system.

Other incentives for use of online archives include autonomous citation indexing, reference linking (e.g., Crossref<sup>1</sup>), easy access to the context of citations (the sentences and paragraphs within articles where citations are made), online discussion and debate forums (cf. *Nature* online), advanced navigation and retrieval functions, and evaluation based on quality. ResearchIndex provides all of these incentives. The algorithms and software that make up ResearchIndex are available for use in other services, and hopefully similar functionality will be made available in other online archives.

Perhaps one of the biggest incentives to make quality research available online will be the realization that the impact of such research will be greater and occur faster if available online. Additionally, movement towards evaluation of science based on the quality of the research instead of the quantity of papers will provide further incentive for authors to make their research as widely available as possible. By making the context of citations easily and quickly browsable, and by indexing preprints, conference papers, and other literature that is often available earlier than journal articles, ResearchIndex can help to evaluate the importance of individual contributions more accurately and quickly, as well as provide more timely and accessible links and feedback to aid the scientific process. As services with the features of ResearchIndex become more common, evaluation of science may increasingly focus on quality.

Changes in access to scientific literature do more than just improve convenience of access. They change the entire scientific process, and the effect of such changes should be carefully considered. We must be concerned, for example, with a bias toward access to non-peer-reviewed medical information. Currently, such information is easier to access than the peer-reviewed medical literature, and thus more likely to be read by consumers.

Another issue is equal access. In addition to differences in access to the web, not much appears to be equal on the web. For example, the distribution of traffic and links to sites is extremely skewed and approximates a power law<sup>2</sup>. This can be seen as a trend towards a “winners take all” effect, where a disproportionate share of traffic and links goes to a small number of very popular sites. The web may have the effect of creating a more common experience when locating information. Whereas researchers may have used a variety of

commercial databases, attendance at meetings, browsing of journals, and discussions with colleagues to locate research in the past, greater use of the web may increasingly bias access towards research that is more highly linked, and research that is more likely to be returned amongst the top results listed by search engines. ResearchIndex attempts to minimize any such effect for access to research papers by providing multiple ways of ranking search results, by weighting citation ranks to account for the age of articles (giving recent articles a better chance to be seen), and by focusing on query-independent methods of locating articles, such as citation linking and a variety of algorithms for locating related documents.

Finally, the existence of an increasing percentage of scientific knowledge in linked form on the web is a development that has advantages beyond the commonly stated improvements to information access. The potential for analysis of interests and relationships within science and society are great. For example, recent research<sup>4</sup> provides efficient methods for objectively analyzing communities and their relationships within the network of scientific literature. This and other techniques may allow more effective analysis of the scientific process, and the progress of science.

1. Publishers agree on deal to link journals on the web. *Nature*, 402:226, 2000.
2. Albert-László Barabasi and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
3. Robert D. Cameron. A universal citation database as a catalyst for reform in scholarly communication. Technical Report CMPT TR 95-07, School of Computer Science, Simon Fraser University, 1995.
4. Gary Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August 20–23 2000.
5. Michael Lesk. The organization of digital libraries, 1999.
6. Theodor Nelson. *Literary machines*. Mindful Press, Sausalito, CA, 1993.
7. A. Odlyzko. The rapid evolution of scholarly communication. *Learned Publishing*, 2001. to appear.