7-1-2002

The Web: Searchable, Hidden, and Deceitful.

Carol Tenopir
University of Tennessee - Knoxville

BY CAROL TENOPIR

The Web: Searchable, Hidden, and Deceitful

THE BLURRING OF LINES among special librarians and other information professionals makes the tracks that some conferences use seem artificial. After all, many librarians work as online experts, knowledge managers, and builders of electronic libraries—often all in one day. This became clear at the recent Information Today (IT) meetings in New York in May, which combined three previously separate shows: National Online, KnowledgeNets, and E-Libraries (see also InfoTech, *LJ* 6/15/02, p. 23). While there were plenty of redundancies, there were several useful sessions on web searching and electronic publishing.

Web searching

One of the few well-attended sessions was led by Ran Hock, principal of Online Strategies and author of *The Extreme Searcher's Guide to Web Search Engines*. Hock provided seven practical searching tips: 1) no search engine covers everything; 2) different search engines miss different things; 3) retrieving large numbers of results is not necessarily bad; 4) all search engines offer advanced search techniques to improve results; 5) meta–search engines are not the same as search engines; 6) Google is great but not the only search engine; and 7) be prepared for changes in all search engines.

Hock recommended Search Engine Watch (www.searchenginewatch.com) and Search Engine Showdown (www.notess.com) for making comparisons. Studies show that the number and depth of web sites indexed by each search engine varies. Every engine may retrieve unique, relevant information for a specific search. Meta–search engines like Vivisimo, DogPile, Meta Crawler, and Search.com provide only the top ten to 20 hits from any one search engine. For a search to be comprehensive, use several different engines and go directly to the search engines themselves.

In addition to indexing selectively, all of the most popular search engines have different proprietary ranking methods. But since all are relevance-ranking search engines, the number of items they retrieve is not nearly as important as the ranking order. The most relevant items should be in the first few screens. Hock got the biggest laugh when he reminded the audience that they are never obligated to look at every single item retrieved. All of these search engines have advanced search features, including Boolean logic, truncation, field searching, and phrase searching.

Hock acknowledged that Google is the favorite search engine because it allows searching for multiple formats (including PDF), indexes a large number of web sites, and factors popularity in its relevance ranking. Not even Google, however, will retrieve everything for a search—for example, its current news coverage is not as good as AllTheWeb. Google also does not offer features like truncation and the NEAR proximity operator, available on AltaVista. Finally, no search engine stays the same for long. The interface, ranking algorithms, advanced features, and depth of indexing may change without notice.

The hidden web

Chris Sherman and Gary Price, coauthors of *The Invisible Web*, revealed secrets of the hidden web. Expert searchers should not only search multiple search engines, they should also remember that valuable material is often not indexed by any search engine and available only if you know the URL.

Materials may be hidden because the producer has barred crawlers or failed to include sufficient metadata or for other reasons. Some examples of these so-called hidden or opaque web resources include materials from the U.S. House of Representatives and Congressional Research Service, videos from sources like PBS, and pages that are dynamically created from a customized search at a web site (such as the Census Bureau site).

Deception on the web

An eye-opening talk for all web searchers was "Lies, Damn Lies, and the Internet" by Anne Mintz, director of knowledge management, Forbes, Inc. Mintz is the editor of *Web of Deception: Misinformation on the Internet*, available in September from CyberAge Books. Mintz showed web sites that purposely mislead. A Martin Luther King site, for example, appears legitimate but is actually the product of a white supremacist hate group. A site that mirrors the World Trade Organization (WTO) web site is actually put together by opponents of the WTO and includes purposely erroneous information.

Mintz gave tips to help searchers protect themselves from deceptive web sites. She recommends checking register.com to determine who owns the domain name; reviewing documented hoaxes or urban legends on www.snopes2.com; scrutinizing the language on a site for inappropriate words, vulgar language, or grammatical errors; and verifying everything in multiple sources. Clearly, the reference librarian's old rule of verifying with three sources is more important than ever in the web environment.

The future of publishing

No one had a definite answer, but several speakers speculated on how long print journals can survive in their present form. David Goodman, research librarian and biological sciences bibliographer at Princeton University Library, questioned whether the present STM (scientific, technical, and medical) journal system can continue.

Goodman built a predictive model by identifying all of the variables that go into answering this question. Depending on the strength of several factors, Goodman believes the answer to how long the present system will last may be anywhere between two years to a decade, as e-print article archives replace traditional published journals. Factors that will hasten

Carol Tenopir
(ctenopir@utk.edu)
is Professor at the School of Information Sciences, University of Tennessee, Knoxville

change include user desire for e-print archives in lieu of purchased journals, general economic conditions that may prove disastrous to library budgets, desire for change in the academic world, and publishers' policies. Flexible policies and satisfaction with the current system will allow journals to continue longer. Goodman believes that at this slow rate of change, only about 40 percent of current journals will survive in a decade.

Michael Eisen of the Public Library of Science (PLS) and I participated in a session, billed as a debate, on the future of scholarly publishing. We ended up agreeing on too many points to have much of a debate. We both acknowledged the value of the editorial and peer-review processes in traditional journals, including the value of a well-respected journal brand name. Eisen and other PLS signators, however, want all journals to allow their articles to be placed on freely available sites after six months; I questioned the wisdom of a single policy for all journals in all disciplines without acknowledging the potential impact on publishers or archiving.

Schroeder weighs in

The final plenary session was by Patricia Schroeder. A former member of Congress and president and CEO of the Association of American Publishers, Schroeder used the opportunity to present her take on the present and future of publishing.

Schroeder's perspective is that of a book publisher, not a publisher of scholarly journals or magazines, although many of her comments, as well as audience questions, focused on periodical publishing. She described a moderate growth in the publishing industry but emphasized the need of growing the market by building readers and changing the culture to one where "what you know" is more valued than "who you know."

Schroeder described the major problems facing commercial publishers, including strengthening copyright laws and overcoming the idea that free speech implies free downloading for all. Of particular relevance to electronic journal publishers is the importance of the Digital Object Identifier (DOI) standard. DOI allows publishers to work together to provide links across journal titles. Schroeder described both linking and branding as "critical." This last point wasn't too far removed from the other sessions on scholarly journal publishing; all recognized that scholarly publishers need to add more value and make their services obvious if they are to survive, let alone thrive.

When you have to choose just one meeting, it is more likely to be the annual conference of the Special Libraries Association or the American Library Association

Do these meetings have a future?

Attendance at Information Today 2002 was down from previous years—a phenomenon experienced by almost all meetings since September 11, 2001. Confusion over what track to attend could have contributed. But it is more likely that soft travel budgets made the nearly $600 price tag for the combined Information Today meetings out of reach for many.

IT's President Tom Hogan estimated an eight to ten percent downturn in paid attendees. Hogan also estimated a decrease of 20 percent in exhibitors, but the decrease in number of exhibit booths made it appear to be much greater. The normally two-floor exhibit space was reduced to just one floor, and many important online exhibitors were absent. LexisNexis, Factiva, and CSA, among others, all skipped this year's meeting.

This might just be an anomaly, or it could indicate that the days of specialized meetings are coming to an end. When you have to choose just one meeting, it is more likely to be the annual conference of the Special Libraries Association or the American Library Association, despite Information Today's promise of three meetings in one.