ERIC

ERIC Digest

OCTOBER 2001

ERIC CLEARINGHOUSE ON INFORMATION & TECHNOLOGY
SYRACUSE UNIVERSITY
621 SKYTOP ROAD, SUITE 160
SYRACUSE, NEW YORK 13244
EDO-IR-2001-02

Uncovering the Hidden Web, Part I: Finding What the Search Engines Don't

Marcia Mardis

Currently, the World Wide Web contains an estimated 7.4 million sites (OCLC, 2001). Yet even the most experienced searcher, using the most robust search engines, can access only about 16% of these pages (Dahn, 2001). The other 84% of the publicly available information on the Web is referred to as the "hidden," "invisible," or "deep" Web.

Despite the explosion in Web content, commonly used search processes have not changed significantly since the Web's inception. Information is commonly found now as it was ten years ago, with directories and search engines. But the ever-quickening pace of the World Wide Web's growth demands an expanded set of search tools and skills. This article provides tips on augmenting traditional search techniques with knowledge of the hidden Web, helping readers to access some of the Web's most valuable content.

The Wrath of the Math

Recent studies estimate the size of the hidden Web to be about 500 times larger than the size of the known "surface" Web indexed by search engines. There are billions of documents obscured in databases, written in non-HTML formats, and hosted through non-http means. According to experts (Bergman, 2000), the hidden Web is conprised of:

- Nearly 550 billion individual documents
- The largest growing category of new information on the Internet
- Content that is highly relevant to every information need, market and domain
- More focused content than surface Web sites
- Total quality content that is up to 2,000 times greater than that of the surface Web
- 95% publicly accessible information not subject to fees or subscriptions.

What do all of these characterizations mean in terms of content? The hidden Web contains current news articles, image collections from museums, and specialized databases full of discipline specific research and reports (ERIC documents being only one example of thousands), U.S. Census information, and so on. Tools to access more of the Web are nascent, but they are growing.

The Way We Are Now

Directories like Yahoo (http://www.yahoo.com) and About.com (http://www.about.com) are human-mediated collections of reviewed and categorized links. Users browse through categories by clicking ever-narrower subject lists. Since a directory's staff can only review and classify a finite number of sites, directories simply cannot keep pace with the explosion of Web content.

AllTheWeb (http://www.alltheweb.com) and Google (http://www.google.com) are examples of traditional search engines that use spidering programs. When the spider program executes, it starts at a specified Web page, indexes that page's content, and follows any hyperlinks on that page. The process is repeated at the destination of each of the hyperlinks. In this way, the program crawls and indexes a web of hyperlinked pages.

When a user enters terms into the engine's search box, those terms are matched in the engine's index; the terms are not found on the "live" Web. Therefore, search engines allow users to go beyond the classification preferences of directory editors to gain term level control over search results. Metasearch tools like Ixquick (http://www.ixquick.com) and MetaCrawler (http://www.metacrawler.com) extend the search engine principle by allowing users to run a query in multiple search engines simultaneously.

While Web directories are obviously constrained by human limits, search engines fail because they primarily index documents written in HTML. Spiders cannot index pages generated dynamically like those in Microsoft's Searchable Knowledge Base and documents written using methods like Adobe Acrobat, Active Server Pages, or Cold Fusion. Likewise, database contents are excluded from the indexing process; spiders cannot transform search terms in database queries or complete a login process. And, in many instances, protocols other than HTTP (e.g., FTP, gopher) are excluded.

Finding the Hidden Web

The first step to accessing the hidden Web is much like that of other search processes: use familiar and reliable resources. Although directories offer limitations as primary search tools, directory categories often contain hidden Web databases. Also, professional journals and magazines provide a wealth of current knowledge; look for reviews of new reference tools and subject directories. In addition to these basic steps, Web-based and desktop solutions are available to access the hidden Web.

With over 7,000 topic-specific databases, there is no way to access every hidden Web resource. But, Web-based gateways, collections, and desktop tools point to specialized databases. These tools are most effective when a few of them are used regularly and integrated into an overall search strategy.

A Smattering of Solutions

- Around the Web in 80 Sites: The Best of the Invisible Web (http://websearch.about.com/library/blow2000.htm)
  The search gurus at About.com created this list of hidden Web resources strong in categorization and expert selection.

- Direct Search (http://gwis2.circ.gwu.edu/~gprice/direct.htm)
  Provides access to the search interfaces of resources that are not easily located with major search engines. This resource is considered by many librarians to be *the* key hidden Web resource.

- Infomine (http://infomine.ucr.edu/)
  A virtual library and reference tool containing highly useful Internet resources including databases, electronic journals, electronic books and many other types of information in a broad range of subject areas and reading levels.

- LexiBot (http://www.lexibot.com/)
  Desktop software that is able to make dozens of queries simultaneously. Surface and hidden Web results are tested for dead links and presented in a format that allows previewing or Web browser viewing. Made for PCs only, this tool is free to try.

- Searchability: Guides To Specialized Search Engines (http://www.searchability.com)
  A gateway site with an annotated list of thousands of search engines covering hundreds of subjects. Descriptions include size, specificity, and some aspects of collection quality.

- SearchEngineGuide.Com: The Guide to Search Engines, Portals, and Directories (http://searchengineguide.com/)
  Currently indexes almost four thousand search engines. Browse for search engines by category or use the keyword search feature. Each entry provides a brief summary.

Quality Should Be Job One

The claim that the hidden Web surpasses the quality of the surface Web is justified by compelling arguments. First, the hidden Web is primarily composed of databases. A site that employs a searchable database is probably current, since Web-accessible databases are fairly new phenomena. Also, a site that puts effort toward collecting and publishing information in a database is usually vested in the topic area.

For example, the Researching Librarian (http://www2.msstate.edu/~kerjsmit/trl/) lists many sites that contain information of interest to information scientists; the most valuable and current information can be found in the sites listed in the database section.

The second argument for the hidden Web's superior quality is that traditional search engines overwhelmingly favor sites in the burgeoning commercial domain (O'Leary, 2000). Since search engines can only find sites that have links to them from other pages, users tend to put links on their pages to popular and well-known commercial sites. Also, sites produced by nonprofit and educational entities do not receive the same advertising and brand name recognition that commercial sites enjoy.

Commercial sites are by no means consistently unreliable. However, educational and nonprofit entities that conduct research in certain disciplines are often excluded in traditional searches. The best source of information is an expert; hidden Web databases point to specialized and authoritative resources.

Wanted: Magic Bullet

Although the Web is often the first place to look, it is not necessarily the best place to look. The hidden Web and other Web-based information resources should be seen as part of an information retrieval process that includes books, serials, and subscription databases. The frontier of the World Wide Web yields a range of quality, currency, authority, and stability levels, so quality issues should be a priority in discussions of information retrieval and in searching instruction.

After using the hidden Web sites, many searchers are disappointed by the need to search each database individually. But search tools have not evolved to the point where the power of a search engine can be seamlessly combined with the quality and depth of the hidden Web. There is no magic bullet; research is a process of carefully uncovering obscured information, not exposing the obvious.

References and Further Reading

Bergman, M. (2000, n.d.). *The deep web: Surfacing hidden value*. BrightPlanet.com LLC. Retrieved August 15, 2001, from the World Wide Web: http://www.completeplanet.com/Tutorials/DeepWeb/contents04.asp

Dahn, M. (2000, January/February). Counting angels on a pinhead: Critically interpreting web size estimates. *Online*, 35-40.

Diaz, K. (2000). The invisible Web: Navigating the Web outside traditional search engines. *Reference & User Services Quarterly*, *40* (2), 131-134.

Ensor, P. (2001, June 14). *Toolkit for the expert web searcher*. Library Information Technology Association. Retrieved August 15, 2001, from the World Wide Web: http://www.lita.org/committe/toptech/toolkit.htm

OCLC (Online Computer Library Center). (2001, July 13). *Statistics*. Online Computer Library Center, Inc. Retrieved August 15, 2001, from the World Wide Web: http://wcp.oclc.org/stats.html

O'Leary, M. (2000, January). Invisible Web uncovers hidden treasures. *Information Today*, 16-18.

Price, G., & Sherman, C. (2001, July/August). Exploring the invisible Web. *Online*, 32-34.

Price, G. & Sherman, C. (2001). *The invisible Web: Uncovering information sources search engines can't see*. CyberAge Books.

Sherman, C. (2000, n.d.). *Worth a look: Searching the invisible Web*. About.com. Retrieved August 15, 2001, from the World Wide Web:http://websearch.about.com/library/searchwiz/bl_invisibleweb_apra.htm

Sherman, C., & Price, G. (2001). The invisible Web. *Searcher*, 9 (6), 62-74.

Snow, B. (2000, May). The Internet's hidden content and how to find it. *Online*, 24. (EJ 613 396).

About the Author

Marcia Mardis, MILS, a former K-12 media specialist, is Program Coordinator and Internet Media Specialist at the Center to Support Technology in Education at Merit Network, Inc. She presents on Web searching issues at conferences around the country and writes frequently on K-12 use of the Internet.

This publication is funded in part with Federal funds from the U.S. Department of Education under contract number ED-99-CO-0005. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government. Visit the Department of Education's Web site at http://www.ed.gov