The Invisible Web

Searching the hidden parts of the web

Ken Wiseman, Apple Distinguished Educator
Technology Coordinator
High School District 214
Arlington Heights, IL

Search engines are pretty good, considering the vast quantity of web pages: more than 800 million pages currently posted on the web, encompassing 15 terabytes of information and about 180 million images. Search engines are often used with the expectation that they are similar to a library card catalog. To find the information that you want on the web, it is just a matter of entering the appropriate words into the search engine and it will return the exact information that you require, right? Those of us growing up in the card catalog days remember that librarians carefully organized information contained by the library for easy access by patrons. Our familiarity with this wonderful service provided by the library community has translated to our use of the web. Unfortunately, the similar user-friendly front end of most search engines hides chaos and information anarchy on the other side.

You and thousands of other professional educators today are learning how to search the web for reliable and valid information using web search tools. You've seen demonstrations and dabbled a bit in digital information retrieval. You know the names Yahoo, Excite, AltaVista, and HotBot. You might have even tried some of the newer search engines, such as Goggle, Inference, and Searchopolis. They all seem to be similar, and yet their results are often confusing. But you persevere, and believe that as your experience and training increases, these search engines will become as valuable as the old, faithful library card catalog.

Sorry to say, this expectation is not likely to be met anytime in the near future. If you are ready for an "ah ha" experience in the Internet realm, let me take a few paragraphs to explain three basic problems that the web has as we try to make it become the information resource of the next century.

Do you need to know these esoteric workings of the web just to do simple searches? The answer is an emphatic YES. Professional educators need to be aware of the limitations of what is destined to become a basic tool of their classrooms and their lives in the next century. Because if you don't, you run the risk of unintentionally excluding more than half of the web from your searches. Would you want to exclude much of the 12 million documents contained in the Library of Congress? Or the U.S. Census Bureau? Would you want to overlook galleries of fine art held by some of the leading art museums

in the world? I could list thousands of similar examples (this is not in the least bit exaggerated) of highly useful information available on the web, but unreachable via the usual list of search tools.

As a typical reader, you might be thinking that I am promoting a new "super search engine" or some service of mine that will overcome the limitations of standard search tools. Let me assure you, in the explanation that follows, that I will not advocate any product, solution, or resource. What I am about to describe is the invisible web that all current general search tools miss. I will offer a couple of potential solutions, but nothing that is comprehensive or final. This is a problem that needs your attention, understanding, and appropriate action.

How web search tools work

First, a bit of background on how search tools work. There are two basic types of search tools available on the web: directories and search engines.[1] Each has its place and is valuable, but they should not be confused and used interchangeably.

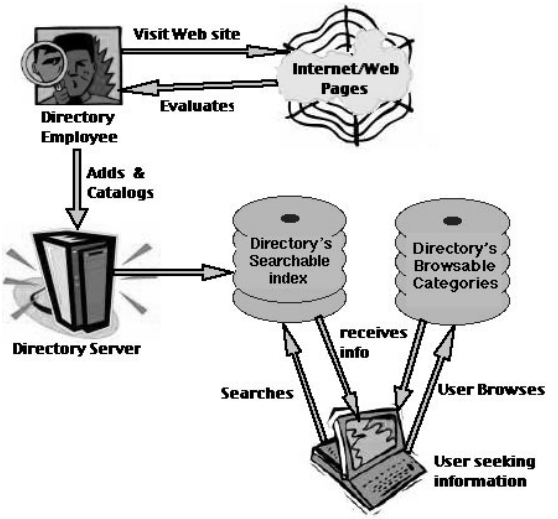

Figure 1. How Internet directories work.

[1] Notice that I have stopped using the term "search engine" in the general way it is often used today. From this point on, I will use "search engine" to describe a particular type of web search tool.

Everyone has used a directory at some point. Conceptually, directories are hierarchical menus with broad categories at the top and buttons to take the user deeper into the organization until the specific information link is reached (Figure 1). Yahoo is the most famous of the directories. Its opening page has broad categories of information listed that the user clicks through to find the exact information desired. Have you asked yourself how web sites get listed in a category? An army of web surfers hired by Yahoo categorizes web sites. Thus, a strength of the directory approach is that a real live person has looked at the site and tried to categorize it in a way that would help you. The limitation of this approach is that with more than 800 million web pages available, this surfing army cannot come close to keeping up with this number of web pages. Also, the accuracy of searches depends on this army thinking and categorizing the way you think and categorize. There is no categorization standard such as the Dewey Decimal System or the Library of Congress cataloging system. Directories are ad hoc, created on the fly by each organization to meet their needs and specifications.

This limitation is not a Yahoo-specific limitation; rather, it is a limitation of the directory approach in general. Yahoo indexes only about six million sites (and it has recently started accepting payment for listing). Using a directory-style search tool, you could unintentionally exclude more than 750,000 million web pages from your searches.

So how about using a search engine? Search engines (such as AltaVista, HotBot, or NorthernLight) are automated robots or spiders that systematically comb the web for servers and web pages. Once a page is found, the robot reads the words on the web pages and adds them to its database for later recovery when queried by a user such as yourself (Figure 2). Doing a search with a search engine is simply querying its database of words (be sure that you understand this concept, as it will come into play later). The search engine does NOT scan the web on your behalf when you type in words to be searched. The search engine merely checks its database of words found by the "bot" on the web. It returns to you URLs containing those words.

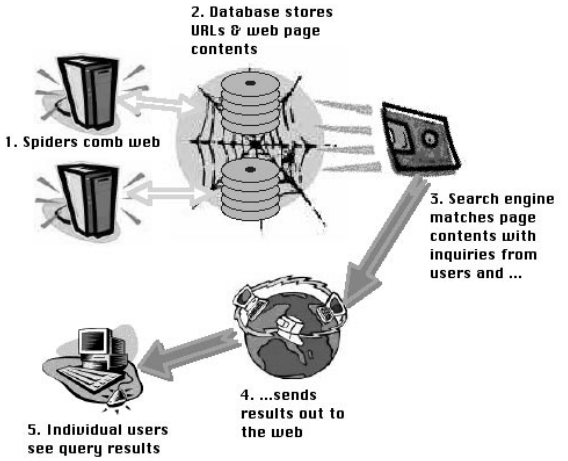

Figure 2. How a search engine works.

This process is all well and good until you realize that search engines suffer from a similar limitation as directories. In a study of 11 search engines conducted by the NEC Research Institute, it was estimated that as of February 1999, the searchable web consisted of 800 million pages containing more than 6 trillion characters [Lawrence and Giles, 1999]. Their previous December 1997 survey put the number of pages at about 320 million (Figure 3). By comparison, the 532 miles of shelves in the Library of Congress contain an estimated 20 trillion characters. Even the most robust web search engine (at this time, NorthernLight) has indexed only 155 million web pages– only one-sixth of the web! Although substantially better than any directory, users are still excluding more than two-thirds of the information on the web. That is down from one-third for the best engine a year and a half ago. The NorthernLight robots have examined about 16 percent of the web, and the other search engines (Snap, AltaVista, HotBot, and so on) have examined even less. (HotBot, which led the 1997 survey with 34 percent coverage, was down to 11 percent in the 1999 study.) The problem isn't that these search engines have lazy robots, but rather that the web is growing so fast that even these 24-hour-a-day, 7-day-a-week robots are losing ground! The 1999 study also found that it takes more than six months on average for a new web page to make it into a search engine's listings.

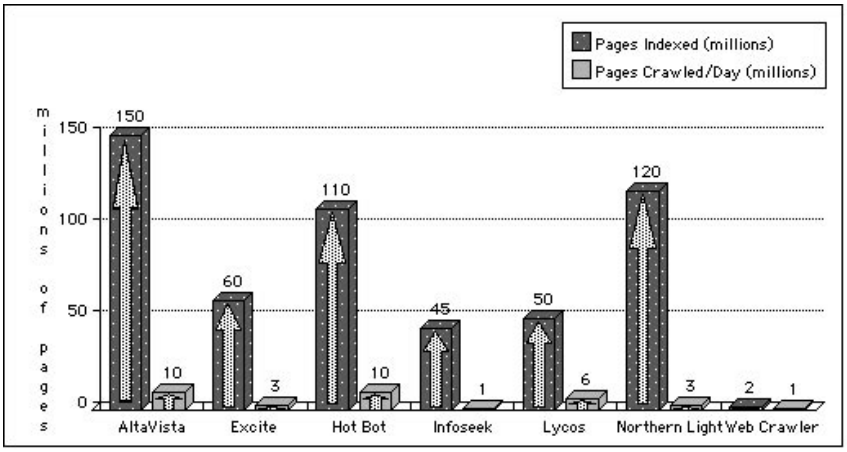

Figure 3. How search engines compare.

Armed with this knowledge, you can become a more sophisticated searcher. All search tools are not created equally. Systematically try search tools to find one that suits your needs. Has it indexed the part of the web that contains the information you seek? Does its database contain pages that represent your needs? These are important questions that most users have never asked. They instead focus on the search tool's ease of use or quick response, assuming that all of the databases were similar.

Other interesting and little-known search engine weaknesses:

- Search engines are more likely to index sites that have more links to them (more "popular" sites).
- They are more likely to index U.S. sites than non-U.S. sites.
- Search sites are more likely to index commercial sites than educational sites.
- Indexing of new or modified pages by just one of the major search engines can take months.

Watch the new search tools as the web evolves. The web has caused major changes in information access and itself is causing the evolution of new and revised search tools. AltaVista promises to index between 400 and 500 million pages within a year. Other search tools such as Google are working to make the relevance of hits higher through intuitive link popularity criteria. Another up-and-coming search site is Netscape's revised search, which relies on both Google and their populist Open Directory search site. However, at this time no one has a viable solution to this perplexing problem, so being aware and willing to try various search tools is the best strategy.

The dynamically generated unseen web

The second unseen aspect of the invisible web is a bit more obscure, but of greater importance to educators. Search engines are designed to read flat web pages. You are familiar with flat web pages if you have done any web page construction. You've spent the time to create a nice-looking web page using a WYSIWYG authoring tool, and your web page was subsequently mounted on a server for the world to see. It has a URL based on the host server. Eventually the words on the page will be read by the search "bots" and added to the database of the search tools.

As the web evolves, however, it is becoming difficult to create these individual pages quickly and in sufficient quantity to house the information contained in some collections. The Library of Congress is a good example. Its web site contains about 12 million documents. This would be an astounding number of multipaged web documents to manually create, link, and assign URLs. The Library of Congress and many other information-rich sites use databases to create web pages on the fly when requested by a user. The database contains the information, which is inserted into a web page template on demand. Thus no flat page is ever created. As a consequence, there is no page for the "bot" to index, and thus no listing in the search engine database.

If your search tools can't see these database-driven, dynamically constructed web pages (and most current search tools can't), you are unintentionally excluding from your web searches:

- 12 million documents from the Library of Congress
- Most data from the U.S. Census Bureau
- ERIC databases
- Most daily newspapers
- Vast collections of fine art owned by important museums
- More than 1,700 other information-rich databases

As I explained above, search engines are databases. One database cannot (without special programming) search another database. When your favorite search engine is confronted with a search entry box (a request to enter information), it is stopped dead in its tracks. The search engine does not dig into the targeted database beyond the initial "enter-search-string" dialog. Remember when doing an Internet search that you are actually querying a search engine's database, not performing a live search of the web.

And even if the search engine robots could get into the databases used by dynamically generated web sites, most dynamically created web pages have changing and variable URLs. Thus, a search engine could not rely on the URL found to be accurate on the next search. Therefore it would not be entered into the search tool database with any reliability.

For clarity, let's take an example on a very simple level. I have typed my name as a query into a specialized web database called AnyWho (www.anywho.com). This database returns my home address, phone number, and information about my neighbors gleaned from the phone book records. As a test I typed my name into a standard search engine. It returned seven hits on my name. None of these were from AnyWho. None contained any personal information. None contained my phone number. The reason, of course, is that the search engine was unable to enter the AnyWho database to retrieve my personal information. Although my personal information is available on the free web, it was inaccessible to the search engine.

On a more complex and comprehensive scale, this is why much of the web is invisible. The AnyWho database created a dynamic web page containing the information requested. The general search engine was not able to enter the text "Ken Wiseman" into the AnyWho database. Thus no information was retrieved.

No one really has a handle on the scope of this problem, but suffice it to say that the problem is vast. There are currently more than 7,000 specialized databases on the web. Each creates dynamic web pages on demand based on user input. Hundreds of other web sites use databases to create their web content on the fly. However, don't be too dismayed; there is some help available.

Two resources, Beaucoup and Lycos, help users manage this dilemma. Beaucoup's sole purpose is to provide an index of searchable databases. Users can search for a database containing the appropriate information they seek. Often these databases contain information that cannot be found elsewhere on the web.

Lycos, the general search engine, has recently (in July 1999) recognized the problem that I have described and is offering a partial solution. The "Invisible Web Catalog" provides links to more than 7,000 specialty search resources. Users can browse listings, or Lycos will suggest appropriate databases within its own search results. For instance, say you searched for "cancer." You'll notice in the search results that there's a link to "Reference>Searchable Databases>Health>Diseases>Cancer." If you click through, you'll discover some important cancer-related sites listed, with links that lead straight to their search

pages. You can then select a resource and try a specific search there. So to get the most out of the Invisible Web Catalog, change your search strategy at Lycos. If you see a searchable database link in the results, consider clicking through to explore the resources there.

You can also browse the Invisible Web Catalog's listings by going to its home page (listed at the end of this article). Once there, you can also choose to search just within the catalog for databases of interest. IntelliSeek, the creator of Lycos (and others), indicates that at some point in the future there may be a solution for this "database meets database" interface problem, but for the present we are left to our own manual searches of multiple databases.

The obvious solution to these problems would be meta-searchers. These are popular search tools that allow searching of multiple search engines simultaneously. Although these seem like a great idea, in practice, searching like search engines (AltaVista, HotBot, Lycos, Excite) all at once simply adds to the clutter and web noise. These all yield similar results because of their similar strategies for combing the web. You are allowed to specify which search tools to query, but these are not the special databases mentioned above.

Apple's Sherlock: One way to search the invisible web

Apple Computer offers an interesting solution to the invisible web problem in the latest version of the Macintosh operating system. Sherlock, an integrated part of the Mac OS, offers the ability to search virtually any database through the use of plug-ins. Through simple programming, the plug-in can teach the Sherlock engine how to query outside databases and make sense of the returns provided by the database. This offers a true revolution in the ability to search the web with a single query. Sherlock has been seen as another meta-search tool (similar to DogPile, MetaCrawler, or SavvySearch), but in light of the invisible web and its plug-in structure, Sherlock offers much more potential than any other general search tool to date. At this time there are hundreds of plug-ins available for Sherlock. These are free and ready for downloading from many web sites. The Apple Donuts site has more than 400 plug-ins at this time. Also, many individual sites have their own plug-ins available for download.

Sherlock is installed as a standard feature of Mac OS 8.5 and later. Once the appropriate plug-in is put into the Internet Search Sites folder in the System Folder, users can click to activate the search of a specific site when a search is requested. Thus, depending on how many plug-ins are activated, Sherlock will query that many databases simultaneously for the requested information (Figure 4). This is a very powerful feature of Sherlock. It can be considered a meta-search tool (it has no database of web sites of its own) with the added function of customization.

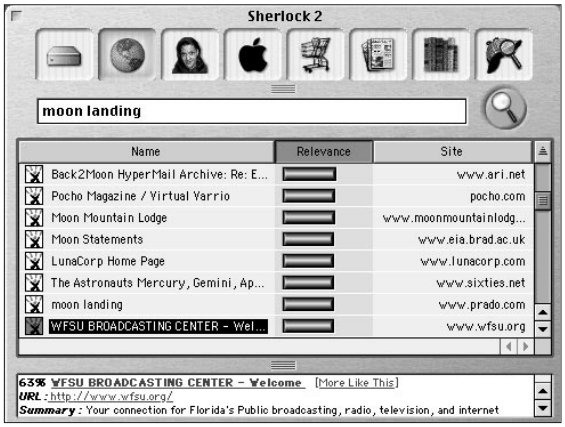

Figure 4. Sherlock search results.

Customization comes by way of the plug-ins, which can be created either by the user or the webmaster via a simple text-based scripting language. Thus, a school district with a web-based library automation system (electronic card catalog) can write a plug-in that allows students to search their own local holdings (print or other media) while making more general web searches [Simmons, 1999].

With the saved sets of search tools feature, Sherlock can be used to search specialized databases—with the emphasis on the plural—with a single entry. Users can have a list of medical databases to be searched when a medical search term is encountered, for example. This feature will obviously save valuable time and effort with no downside trade-off. This is an important feature for both the search professional and the casual searcher.

Unfortunately Sherlock is not presently a cross-platform tool and I see no indications that it will become the QuickTime for Internet searching in the future. This does not, however, leave Windows users completely out of the picture. Browser-based search and indexing functions are incorporated into both Internet Explorer 5 and Netscape 4.5. This smart browser feature brings up a streamlined index of popular sites in response to search queries. Similar Sherlock-like utilities for Windows come in the form of Mata Hari and Internet EZ Search.

I can safely predict that the invisible portion of the web will continue to grow exponentially before the tools to uncover it are ready for general use. Until then, educators and trainers must make an effort to understand and explain the problem to novices. They must be given the tools and motivation to go beyond the simple "search" button in their browser to seek all of the relevant

sources of information. Of course, articles like this one can be quickly outdated. Even as I write this, new resources are becoming available. New natural-language search tools, XML searching, and improved context searching are just on the horizon. We all must become active searchers of the search tools to make our queries more productive.

URLs

Invisible Web resource page	www3.dist214.k12.il.us/invisible/default.html
Alta Vista	www.altavista.com/
AnyWho	www.anywho.com
Apple Donuts	www.apple-donuts.com/
Ask Jeeves	www.askjeeves.com/
Beaucoup	www.beaucoup.com
Ditto	www.ditto.com
DogPile	www.dogpile.com/
Excite	www.excite.com/
Google	www.google.com
GoTo	www.goto.com
HotBot	www.hotbot.com/
Inference	www.infind.com
Lycos	www.lycos.com/
Lycos Invisible Web Catalog	dir.Lycos.com/Reference/Searchable_Databases/
Mata Hari	www.thewebtools.com/
MetaCrawler	www.go2net.com/search.html
NorthernLight	www.northernlight.com/
SavvySearch	www.savvysearch.com/
Searchopolis	www.searchopolis.com
Sherlock	www.apple.com/sherlock/
Snap	www.snap.com
Yahoo	www.yahoo.com

References

Simmons, Mark. "How to Make a Sherlock Plug-in," *MacAddict,* Volume 4, Number 4 (April 1999), pp. 72–76.

Lawrence, Steve, and Giles, C. Lee. "Accessibility of Information on the Web," *Nature,* 8 July 1999, pp. 107–109.

About the author

Ken Wiseman (kwiseman@dist214.k12.il.us), is the High School District 214 Technology Coordinator, Arlington Heights, Illinois. He has spent 22 years teaching in the schools of northern Illinois and more than 10 years as a full-time technology professional working for school districts and the Illinois State Board of Education. He holds a Masters degree in Educational Technology from Northern Illinois University.

Since 1994, he has been District Technology Coordinator for the second-largest high school district in Illinois, District 214. This district is in the midst of an extensive technology renovation plan to provide the infrastructure that will be needed by educators as they move into the next century.

Ken's perspective is teacher based, with a strong orientation toward infusing new technology to improve education. At the forefront of this orientation is the importance of educating students for their future and applying the tools of technology where appropriate. He knows all too well that boxes and wires installed in a classroom do not create systemic change. It takes teachers working with students along with district support to create change.