ADVANCED WEB SEARCHING FOR THE INFORMATION PROFESSIONAL

Kristen L. Metzger
Library & Information Center
Harbor Branch Oceanographic Institution
5600 U.S. 1 North
Fort Pierce FL 34946 USA
metzger@hboi.edu

ABSTRACT: Information professionals don't "surf the net" since this implies that one is just skimming over the surface of a sea of information or worse, drowning in it. To quickly extract specific, relevant information from the Internet, the serious searcher must be familiar with the structure, functionality, strengths, weaknesses and special features of the most efficient search engines. We'll examine the differences between free-text and index-based search engines, multi-search engines, web directories, metasites and intelligent agents.

KEYWORDS: Web Search Engines, World Wide Web, Internet Searching, Electronic Information Resource Searching

Everything's on the Internet! Librarians know that this is not so, but when you're searching on the web, the volume of information might convince you that it *is* true. How do we, as information professionals serving the scientific community, locate relevant sites on the Internet when the sites we want are so overwhelmed by those catering to the great-unwashed masses. We don't want sites devoted to Pokemon, Britney Spears, the Backstreet Boys and naked women. In general, we are searching for information, not for entertainment.

SEARCHING BASICS

As always in reference work, you have to find out what the patron is really after. Successful reference interviewing is essential. I regularly have scientists arrive with a little scrap of paper that they shove under my nose. Invariably it contains a list of keywords. The first thing I say is "Tell me in a sentence what it is you want to know." Often I follow that with 3 or 4 more questions getting more detail until I arrive at what, I hope, is the question they are actually asking.

Don't forget that you were able to provide reference help long before there were computers. Sometimes it's just easier and faster to use an almanac, dictionary or a phone book. Information that would have taken 15 minutes to turn up on the web, I've often found in the World Almanac in a nanosecond.

Knowledge of Boolean logic is essential. George Boole, a mathematician from the 19th century, developed a simple way to define logical relationships between terms in a search query by using AND, OR and NOT. Keep in mind that some of the search engines allowing for Boolean searching use AND NOT instead of NOT. Use of Boolean operators focuses your search by telling the search engine how your search terms relate to one another instead of just finding the terms at random. On web search engines, it is best to capitalize Boolean operators. It is generally smarter still to use "simplified Boolean"; the plus (+) sign before a word indicates AND and the minus (-) sign indicates NOT. When simplified Boolean is used, a search engine's ranking algorithm is not generally overridden. However, with AND, OR, NOT, sometimes it is.

Successful web searching requires the searcher to understand that computers look for words, not concepts. Narrower specific searches are better than broad ones. I regularly watch in awe as highly educated individuals interested in something as specific as "Calanoid Copepods" will enter a search for "Marine Invertebrates". One researcher told me that he thought that a search on *fish farming* would "automatically" pick up *aquaculture* sites. Computers don't do anything automatically.

Effective searching is pretty much limited to people who know how to spell since computers are quite unforgiving if you misspell words. Occasionally misspelled words, such as the scientific names of marine species, will allow you to arrive at a web site produced by - you guessed it – someone who can't spell.

JUNK ON THE WEB

While it is impossible to cover the vast subject of web searching in one short paper, I hope to point you to some of the best metasites, search engines, metasearch engines and intelligent agents. However, one caveat: while I hope to illustrate ways to turn up relevant information on the web, all bets are off once you've arrived at the "relevant" web site. There is definitely no quality control on web sites. If the information isn't flat out incorrect or out of date, the design of the site drives you up a wall. You will very likely encounter sites that should be banned from cyberspace - those filled with erroneous and/or outdated information, irritating animations and scrolling text pages, pages with ridiculously long download times and poorly organized information that no logical person could navigate, aggravated further by complex URLs that you couldn't repeat to a colleague even if you wanted to.

Poorly designed, useless and/or moronic web sites could easily be the topic of a lengthy paper on its own. A number of web sites are devoted to this subject. Ironically, many of these arbiters of design and content are guilty of creating sites with any number of egregious design flaws. Should this subject interest you, here are a few sites to check out:

http://www.worstoftheweb.com/
http://www.webpagesthatsuck.com

WEB SEARCHING

Unfortunately, web search engines are developed with the casual searcher in mind. The individual who has searched extensively on sophisticated online services such as Dialog, Data-Star, and Lexis-Nexis is undoubtedly disappointed by the lack of search features and reliability exhibited by web search engines. On the up side, use of search engines is free and the development of these search engines in the past seven years has revolutionized the search for information on the Internet.

Your odds of turning up useful information are vastly improved if you understand the strengths, weaknesses and advanced features of the major search engines. Indispensable charts comparing the search features of major search engines are compiled and updated by Greg Notess at http://searchengineshowdown.com/features. The information is presented in a different and equally useful format at http://searchengineshowdown.com/features/byfeature.shtml

I couldn't possibly present a comprehensive overview of search engines in a thirty-minute presentation or in a short proceedings article. The search engines I've chosen to concentrate on are my personal favorites.

INDEX BASED SEARCH ENGINES, WEB DIRECTORIES, PORTALS

Index based search engines arrange information in a structured pattern, using headings and subheadings going from the general to the specific. For better or worse, these subject headings are created by people. On the up side, a person is more likely to think like you than a computer. Conversely, the success of a web directory depends on the searcher thinking in the same way as the creator of the index.

About
http://www.about.com

This site is an excellent web directory sorted by subject, but it is definitely a site with value added. It distinguishes itself from other subject directory sites by including original articles written by its "guides." About.com, formerly The Mining Company, hires real live, thinking people, called guides, to organize individual subject areas on the site. The 700 guides gather the resources in their areas of expertise and are encouraged to interact with searchers via newsletters, email, discussion and chat groups. I like this site well enough to overlook its messy home page.

Infomine
http://infomine.ucr.edu

A very selective, research-oriented web directory, Infomine is a collection of quality sites likely to be useful to students, faculty and researchers at the university level. It is compiled by librarians at the University of California at Riverside. A recently added

feature is their "New Resources Alert Service." By registering your email address and choosing general subject areas, you will be informed of new resources as they are added to Infomine.

Looksmart
http://www.looksmart.com

Looksmart is one of the largest web directories. It has a cascading menu that is very easy to use and has about a 1.4 million URL's presented in more than 70,000 categories. It is really more consumer-oriented than research-oriented, but not all of our reference questions are scientific in nature.

Open Directory
http://www.dmoz.org

This site is growing rapidly and now competes with Looksmart as the largest web directory. It is "Open" in that other services can lift the entire directory or relevant subject areas and put them on their own servers.

The Open Directory is compiled by an army of volunteer editors, and understandably, the quality of editors varies considerably. The volunteer editors organize a small portion of the web, removing the bad and useless sites. More search engines, such as Hotbot, Lycos and Altavista are using Open as their associated directory than any other.

The Open Directory for the more visually oriented can be found at http://www.webbrain.com , where the categories are presented spatially.

Yahoo
http://www.yahoo.com

Yahoo is probably the best known of the web directories. It's a very well organized site that allows you to search within a specific category or to search the entire Yahoo database. Yahoo will accept the Boolean AND and OR and is case sensitive, allowing for searches on AIDS that won't turn up a billion sites on instructional aids or heaven forbid, marital aids.

The "Advanced Search" feature allows you to move beyond the Yahoo directory by clicking on "Web Sites." This runs your query into the Google search engine. The primary weakness of Yahoo is that only a small portion of the web is catalogued by Yahoo and it's quite possible that only "paid" sites will get listed in a timely fashion. A $100 fee will move your site up the priority list for an evaluation, but it still doesn't guarantee that your site will be included.

FREE TEXT SEARCH ENGINES

A number of the so called "free text search engines" listed below have morphed into "portals", offering email, web directories, stock quotations, maps and any number of other bells and whistles, but these were all born as search engines.

AltaVista
http://www.altavista.com

AltaVista, launched in late 1995, was the first intense search engine that penetrated below the top level of web sites and was able to pull up terms buried down a level or two. It offers both simple and advanced modes, allows for Boolean, proximity and phrase searching, plus truncation. Its Advanced Search option mode accepts the proximity operator, NEAR.

AltaVista offers a sophisticated text analysis technique called REFINE on the results page. The REFINE option acts a bit like a thesaurus. It displays a list of terms that might be appropriate to your search, then allows you to identify related terms, add additional concepts and eliminate terms that aren't relevant to your search.

In AltaVista, you can search in more than 20 languages and it has a translation link that will translate either way between English and French, German, Italian, Portugese, Russian and Spanish. Recently added language options include two way translations between French and German.

Several months ago, AltaVista created one of the largest directories on the Internet by combining the full content of the Open Directory and the LookSmart Directory. The records from both databases were merged, the duplicates removed and the entries in each section are now sorted using AltaVista's relevance ranking.

Another feature of AltaVista that I particularly like is ability to search for images. As my institution's Intranet content manager, I'm always searching for clipart or animated gifs to add interest to the news and this is a great way to find them.

Raging http://raging.com provides the powerful AltaVista search engine on a nice clean screen with no graphics, banner ads or surrounding portal content. It appears to be a copy of Google, but with one major difference. Link popularity is the primary relevance-ranking criterion at Google, whereas at AltaVista and Raging, it is only one factor in its relevance-ranking algorithm.

Fast (formerly AlltheWeb)
http://www.alltheweb.com

Launched in May of 1999, Fast is just that – fast. This is a great search engine to use for a quick and dirty search. It offers simple and advanced search options. The Advanced

Search option allows you to set language preferences, filter domains and restrict results by number and/or whether or not the content is offensive. Boolean is limited to the use of the plus and minus signs. Fast is not case sensitive and doesn't allow truncation. Despite those limitations, it is huge and fast and definitely worth a look. It appears that its relevance-ranking algorithm may be superior to that of other major search engines.

Google
http://www.google.com

A popularity engine, Google ranks records based on their popularity; that is, how often other sites refer, or link, to a site. It gives some consideration to the proximity of search terms within a record. Google has been praised for both its high rate of relevancy and the simplicity of its home page. You enter your terms in the search box and then click on either "Google Search" or "I Feel Lucky". "I Feel Lucky" will take you to the one web site that would have appeared first in your results had you run a "Google Search."

Google automatically ANDS all your search terms. To NOT a term, use the minus sign. Google doesn't allow truncation and is not case sensitive. It doesn't search on "stop" words, but if you have entered a stop word that it is ignoring, that information is reflected at the top of your results page.

A unique feature of Google is that it "caches" the web pages it indexes. If a web site has disappeared or changed its URL between the last time Google indexed it and the time you searched for it, you can take a look at the "cached" site.

HotBot
http://www.hotbot.com

One of my favorite search engines for some time now, HotBot was created by the same organization that published *Wired* magazine. Its size, functionality, search options and quality of retrieval have made it a favorite among serious searchers.

If the garish colors and clutter on the home page annoy you, click on *Help,* go to the bottom of the screen and click on *Text-only version.*

Hotbot's Advanced Search option allows narrowing of searches by language, word filters, date, domain, page depth and type of media. It is case sensitive, accepts full Boolean, phrase searching, and, as far as I know, is the only search engine accepting both right *and left* truncation. If I have to find a negative about HotBot, it would be its use of stop words.

Northern Light
http://www.northernlight.com

Launched in 1997, Northern Light is a web search engine with a couple of very unique features: (1) Results are organized into "custom search folders", bringing together sites that have common characteristics such as type of document, subject category, etc. Using a controlled vocabulary, it automatically classifies every document. (2) More than 6,900 "premium" sources are covered. These include magazines, newsletters, books, newspapers, etc. Abstracts can be viewed and the full document can be purchased for a nominal charge. Titles covered can be viewed by clicking on the "What is Special Collection" link on the home page and then downloading the list.

Northern Light accepts both full and simple Boolean, phrase searching, truncation, embedded wildcards (%) and field searching. It is not case sensitive and does not use stop words. Their method for determining the relevance of web sites is more complex than that of many other search engines. Relevance is based on query term frequency, the presence of query terms in document titles, the document date and link popularity. By offering a bibliographic database search service and document delivery, Northern Light is much more than another web search engine.

SPECIALTY SEARCH ENGINES

Askjeeves
http://www.aj.com

I won't claim to be a regular user of Ask Jeeves, but I am intrigued by its rather unique approach to searching. Instead of using search terms, you ask a plain English question. Ask Jeeves responds by displaying a list of closely matching questions that it understands. Following the list of questions, results are displayed from other search engines that it has probed.

The Jeeves Peek option http://www.ask.com/docs/peek/ displays the questions being posted to Askjeeves in real time and the page refreshes every 30 seconds. Some might find this depressing, since it often illustrates the shocking stupidity of the human race.

Biolinks
http://www.biolinks.com/

Referred to as an "Internet Search Engine Designed by Scientists for Scientists", Biolinks provides a search engine and a very useful index format that includes categories such as meetings, careers, companies and organizations, all scientific in nature.

Euroferret
http://www.euroferret.com

Euroferret is an index of 35 million European web pages. Searches can be narrowed to individual countries and/or languages. Ironically, euroferret.com does not index any sites in the .com domain. Euroferret turns up more European domain documents than can be retrieved by either AltaVista or HotBot. Euroferret has recently been absorbed by Webtop.com, which does include .com sites. However, searching the older version of Euroferret is still an option.

Search Adobe PDF Online
http://searchpdf.adobe.com

Web search engines ignore PDF files and unfortunately there is a lot of information for the serious researcher that is only in PDF format. This search engine by Adobe® contains searchable summaries from more than a million PDF documents across the web. After viewing the summary, you may choose to download or view the entire document. This is a great site for turning up obscure technical reports and other gray literature.

Dejanews
http://www.deja.com/usenet

Dejanews is now the sole provider of a searchable database of the contents of Usenet groups. Altavista used to allow newsgroup searching via RemarQ.com and RemarQ.com could be searched directly at their own site. Unfortunately, RemarQ has been acquired by Critical Path and they have no plans to offer free web based access to newsgroups. HotBot and Metacrawler both include Usenet searching as a feature, but Deja.com powers it in both cases.

METASEARCH ENGINES

Metasearch engines send queries simultaneously to multiple web search engines or directories. They differ widely as to which engines they cover, how they display results, whether they eliminate duplicates, the limits they place on how many records can be retrieved from each search engine and whether or not they convey your full query syntax to the targeted search engines. These metasearch engines are most useful for single word or phrase queries. The more obscure the topic, the more likely that a metasearch engine will provide the results you want.

Dogpile
http://www.dogpile.com

Dogpile is a favorite of many searchers, maybe because it replaces the "Search" button with "Fetch". Dogpile runs your search through 14 search engines. By clicking on "Change search order", you can customize your search by specifying the order in which

queries are sent to the search engines or you can delete search engines from the list. Dogpile transmits Boolean operators to search engines that accept them and is capable of changing NEAR to AND if the receiving engine doesn't support the NEAR operator.

Ixquick
http://www.ixquick.com

Launched in October 1999, Ixquick searches 14 search engines at once more intelligently than most other metasearch engines. It knows which search engines can handle Boolean logic, truncation, phrasing or field searching and translates your search into each engine's syntax. For instance, if you place your search terms in quotes to signify a phrase search, Ixquick will omit the search engines that ignore quotes. Duplicates are removed and results are ranked according to how each individual search engine ranked them, awarding one star for each search engine that placed a site in its top ten.

Mamma
http://www.mamma.com

"The mother of all search engines", Mamma was started in 1996. It searches 10 major search engines, formatting the words and syntax for each source being queried. Mamma organizes the results into a uniform format and presents them by relevance and source. The home page and search box is very simple and uncluttered. Mamma also has a "Power Search" option.

Metacrawler
http://www.metacrawler.com

Metacrawler, purchased by Go2Net, Inc. from the University of Washington, also offers a home page simple search box and "power search" options. Both search options provide limited Boolean capabilities by allowing phrase searching or allowing the choice of "any" or "all." A new feature of interest to voyeurs is "Metaspy", which shows ten real-time search topics. The "Metaspy" list automatically refreshes every 15 seconds. This feature is very similar to Jeeves Peek at Ask Jeeves.

MetaIQ
http://www.metaIQ.com

MetaIQ allows you to choose from 15 major search engines or to target other search engines by picking from their index of subject specialty engines. Additionally, a search box is provided for retrieving up to the minute news from more than 1500 news sources.

Profusion
http://www.profusion.com

Designed by the University of Kansas, Profusion was purchased last April by Intelliseek and its speed and accessibility has improved dramatically. It is a very popular metasearch engine that offers options in search engine selection. You may choose the best three, the fastest three, all 9 or any combination of the available engines. Profusion allows Boolean and phrase searching, eliminates duplicates and sorts results according to a relevance ranking score.

METASITES

Metasites are specialized directories related to a particular topic. The purpose of a metasite is to direct you to other sites on the web.

Argus Clearinghouse
http://www.clearinghouse.net

Clearinghouse is a directory of reviewed metasites, essentially a metasite of metasites. The Argus Clearinghouse identifies, describes and evaluates Internet based information resources. It has a hierarchical organization that begins with 13 top level categories. You may have to click through several sublevels to get to your topic of interest.

Digital Librarian
http://www.digital-librarian.com

Called "A librarian's choice of the best of the web", this is an excellent, eclectic collection of web sites. I question the wisdom of offering 90 different categories on the home page of the site, but it is still workable. Obviously reflecting the interests of librarian, Margaret Vail Anderson, the categories include topics like "central New York" and "yurts and tipis". However, most of the categories are more mainstream than that and the "marine sciences" collection of sites is certainly worth a look.

Homework Helper
http://www.bjpinchbeck.com

This is the first place I send kids when they ask for help. Thirteen-year old B.J. has compiled an impressive, well-organized collection of more than 600 useful sites and generally adds a one-sentence evaluation of each.

Internet Public Library
http://www.ipl.org

The IPL originated in a graduate seminar in the School of Information and Library Studies at the University of Michigan in 1995. It contains more than 27,000 carefully selected and cataloged sites and serves as an excellent electronic ready reference collection.

Librarians Index to the Internet
http://www.lii.org

This site originated as the gopher bookmark file of Carol Leita, a Berkeley, California public librarian. It contains quality information that has been classified and annotated by 94 librarians into 1,347 subject areas. (McDermott, 2000)

Library catalogs
http://sunsite.berkeley.edu/Libweb/
http://lcweb.loc.gov./z3950/gateway.html

Both of these sites provide extensive access to individual library catalogs. If you are involved in interlibrary loan, this is a great way to find out if a book is really currently available or is already checked out. By looking at a library's serial records, you can determine if you're the only library in the world that still hasn't received a particular issue of a journal.

Search Engines Worldwide
http://www.twics.com/~takakuwa/search/search.html

Search Engine Colossus
http://www.searchenginecolossus.com

Search Engine Directory
http://www.allsearchengines.com/

Beaucoup
http://www.beaucoup.com

All four of these sites are metasites of numerous search engines. The first two are indexed by country, the last two by subject area.

Bibliographic Databases in Oceanography, Earth Science, and Related Areas
http://scilib.ucsd.edu/sio/guide/bibliographies.html

Careers in Marine Science
http://scilib.ucsd.edu/sio/guide/career.html

Tides and Tide Prediction
http://scilib.ucsd.edu/sio/tide/

All three of these wonderful web pages are compiled by Peter Brueggeman of Scripps Institution of Oceanography. The first directory is an exhaustive list of free bibliographic databases. Databases requiring payment and/or site licenses are not included. The Careers in Marine Science and the Tides and Tide Prediction sites are equally comprehensive.

Virtual Technical Reports
http://www.lib.umd.edu/UMCP/ENGIN/TechReports/Virtual-TechReports.html

Virtual TechReports has an extensive list of institutions that provide either full-text reports or searchable extended abstracts. It provides links to preprints, technical reports, dissertations and theses.

INTELLIGENT AGENTS (Desktop Portals, Desktop Browser Searchbots, Browsing Companions)

Regardless of what you choose to call them, these are software programs downloaded to your computer that run separately from your browser or Internet programs. Two standalone search agents are BullsEye and Copernic, both offering free versions and upgraded versions for a fee.

BullsEye 2.5
http://info.intelliseek.com/prod/bullseye.htm

Copernic
http://www.copernic.com

BullsEye and Copernic are quite similar. I prefer Copernic so I will focus on its features here. The Copernic web site is available in 6 different languages and both PC and Mac versions of the software are available on this site.

The Copernic software is closely integrated with Internet Explorer. After downloading Copernic, clicking on the *Search* button in Internet Explorer will place a Copernic search box on the left side of the screen where your "Favorites" normally appear.

Copernic simultaneously searches the web, newsgroups or email using numerous information sources. It supports term searching, phrase searching and natural language queries. While search results are gathered, a pop-up window shows each search engine being contacted and the progress on results. A search can be further refined after searching, dead links can be removed and the results can be emailed in either HTML or text formats. Searches are automatically stored and easily retrieved.

SUMMARY

Search engines change continuously. To keep on top of these changes, I highly recommend the following print and electronic materials:

The CyberSkeptic's Guide to Internet Research
Online: the leading magazine for information professionals
Searcher: the magazine for database professionals
http://searchengineshowdown.com/
http://www.searchenginewatch.com

REFERENCES

Basch, Reva and Mary Ellen Bates. 2000. *Researching Online for Dummies.* 2nd ed. IDG Books Worldwide, Foster City, California.

Bradley, Phil. 1999. *Internet Power Searching: the Advanced Manual.* Neal-Shuman Publishers Inc., New York, New York.

Hock, Ran. "*Brief profile: Fast.*" [Online.] Available: http://onstrat.com/engines/fastprofile.htm [December 7, 2000].

Hock, Ran. "*Brief profile: Google.*" [Online.] Available: http://onstrat.com/engines/google.htm [December 7, 2000].

Hock, Ran. 2000. Raging search – simplicity revisited. *CyberSkeptic's Guide to Internet Research* July/August 2000, 7.

Hock, Randolph. 1999. *The Extreme Searcher's Guide to Web Search Engines; a Handbook for the Serious Searcher.* CyberAge Books, Medford, New Jersey.

Holler, Suzi. *Tangled in the Web.* [Online.] Available: http://pegasus.cc.ucf.edu/~s-holler/tangled.html [December 7, 2000].

Kennedy, Shirley Duglin. 2000. Trapped in a web of bad design. *Information Today* 17(4): 32-34.

King, David. 2000. Specialized search engines; alternatives to the big guys. *Online* 24(3): 67-74.

Mandelbaum, Judith. 2000. Ixquick: a new metasearch engine. *CyberSkeptic's Guide to Internet Research* July/August 2000, 6.

McDermott, Irene E. 2000. Classified with class: superior subject sites. *Searcher* 8(4): 10-18.

McDermott, Irene E. 2000. Gotta catch 'em all: metasearching the web. *Searcher* 8(3): 24-28.

Notess, Greg. *"Search engine showdown."* [Online.] Available: http://searchengineshowdown.com/ [December 7, 2000].