2004

Internet searching: Moving from clueless to competent

Robert I. Berkman
The University of Montana

INTERNET SEARCHING:

MOVING FROM CLUELESS TO COMPETENT

by

Robert I. Berkman

B.A. The University of Virginia, 1980

presented in partial fulfillment of the requirements

for the degree of

Masters in Journalism

The University of Montana

May 2004

Approved by:

Chairperson

Dean, Graduate School

5-27-04

Date

Berkman, Robert I., M.A., May 2004 Journalism

Internet Searching: Moving from Clueless to Competent

Chairperson: Dennis Swibold DS

Defense Completed and Approved: May 13, 1998; Final Copies Delivered: May 2004

Journalists are moving quickly to the Internet to perform their research. Reporters gather information online on topics ranging from technical data, government reports, newspaper archives, international statistics, and much more. However, few reporters have been trained in Internet research techniques, and in order to be effective searchers, it is important to know the basic strategies in Web research.

There are a few factors that can improve reporters' searching. These include knowing when to use the Internet as a source, when to use a different resource, and choosing the right search tool. Not all search tools are alike and it is important to choose the best one for specified tasks. For example, hierarchical indexes like Yahoo! are best for broad searches and standard search engines like AltaVista are best for narrower searches. It's also important to be able to know how to create an effective keyword search, including using Boolean operators and field searching when appropriate.

Other important points that reporters need to keep in mind when going on the Internet to do their research is to use the expertise of an in-house librarian, and to know how to evaluate the quality and credibility of the Web pages retrieved. A few basic guidelines include basic journalistic rules like ensuring that the story is from the original source, verifying the data with other sources, asking the source questions and using one's own judgment.

One emerging problem faced by reporters is information overload. There are specific strategies that journalists can employ to help overcome this problem.

Internet Searching: Moving from Clueless to Competent

Table of Contents

Reporters Migrating to the Web 1

Where to Go? 3

What's Not on the Menu 4

News on the Net 4

Starting a Search 7

Creating a Search Statement 9

Remember Your Librarian 16

SIDEBAR: Searching AltaVista 17

Special AltaVista Search Tips 19

But is it True? 20

Journalism 101 22

Information Overload 24

Further Resources 26

Attachment 27

Updated published piece in the *Chronicle of Higher Education*

INTERNET SEARCHING: MOVING FROM CLUELESS TO COMPETENT

As Steve Woodward, a veteran reporter at the *Oregonian,* tells it, his newspaper's science reporter took a vacation at a most inopportune time.

In just two days Comet Hyakutake was expected to become visible in the Northwest sky. Readers would be expecting some help from their daily paper: where in the sky to look, when precisely the comet would appear, and how to best view it. But with the science reporter gone, the assignment was given to Woodward, who had no knowledge of astronomy.

But he did know the Internet. By doing a few quick searches, Woodward located star maps, links to astronomy publications, postings to astronomy newsgroups, releases from the Smithsonian, and a wide range of helpful pages created by passionate amateur astronomers. Woodward says that the information he found even allowed him to double-check some viewing information provided to the paper by a local source—who turned out to be incorrect.

Meanwhile, across the country, at *The Cape Cod Times,* a regional daily based in Hyannis, reporter Susan Milton was also turning to the Internet for help in pursuing her story: the likely impact of the forthcoming deregulation of electricity on the Cape. The utilities claimed that prices would decrease. Milton knew that deregulation had already come to the area, so she searched for and found a regional newspaper online. She did some reading and discovered that the price of power in Providence had risen, not dropped!

Reporters Migrating to the Web

Internet success stories like these are becoming increasingly common, as reporters continue to flock to the Internet. A September 1997 survey of 2,500 managing and business editors of magazines and daily newspapers conducted by Columbia University's Graduate School of Journalism and Middleburg

Associates, a public relations firm, found that almost half said that they or their staff go online every day. That's up from 34 percent in 1996 and 23 percent from 1995. Furthermore, 93 percent of the respondents reported that they or their staff use online services at least occasionally (See: Middleberg/Ross Media in Cyberspace Study <http://www.mediasource.com>).

In what seems like no time, the ocean of information called the Internet has, for many reporters, evolved from a murky and mysterious world to a familiar destination for scooping up buckets of data. J.R. Wilson, a freelance writer based in California who covers the aerospace and defense industries says that he can find something relevant and of interest on the Internet for about any story he undertakes.

But although the Internet is evolving into a standard research source, many reporters have little training and knowledge in how to do effective research on the Internet. Says Woodward, most reporters are actually "pretty clueless" about Web searching. Adds Richard Dresden, a news librarian and accomplished searcher at *The Washington Post,* "many are fairly ignorant" about how to go about creating a skilled search statement.

Some of this is understandable. After all, until just a few years ago, the somewhat mysterious art and science of online searching was the special province of trained librarians, information specialists, and scientists. These people were taught the necessary and somewhat arcane skills of constructing complex searches, modifying their searches, and sorting and evaluating results. And since these were conducted on very expensive and complex online systems, searchers had to be very careful in making sure that their searches were efficient and accurate. But the Internet has now made online searching available to anyone with a PC—but almost nobody has been trained in how to actually go about doing effective searches.

It's true that searching for information on the Internet can be pretty simple. "Even a two year old can do a basic keyword search" says Peter Basofin, chief librarian at the *Sacramento Bee's* newsroom. Typing a couple words into a search engine is not exactly brain surgery. But, as in most endeavors, there's a big difference in being able to "get by" and to be proficient. To be a good researcher, you need to know the tricks and strategies for doing the best possible searching on the Net. Let's take a look at some:

Where to Go?

One of the first things you'll want to know is when it makes sense to turn to the Net. Is it, in fact, like "Alice's Restaurant", where you can get anything you want? Well, yes and no. While it's true that there is an incredible amount of information, in some instances the Internet is a less suitable source.

Let's look first at where the Net does fulfill its potential. At the top of the list would have to be government information. Whether it's statistical data, economic releases, regulatory data, or consumer advisories, public agency information is abundant on the Web. Much of this comes from the federal government, but there are mountains of information from the state and local governments too.

Other countries' governments have put information on the Net as well. J.R. Wilson needed to do research on the aerospace industry in Latin America, and he says that some of his best sources came from embassies that posted statistics and information on their Web sites.

In a similar vein, when Gary Dehlson, a reporter at the *Sacramento Bee* needed background on current political issues in Israel, he found mountains of useful data from Web sites created by the country's Prime Minister's office, as well as from the Palestinian National Authority's site. He even found the full text of Israel's treaties.

Governmental information isn't the only area where the Net delivers. You can find all sorts of business and industry information, including company home pages, annual reports, corporate press releases and financial filings with the SEC. You'll also find high-tech news and market research data, health and medical information on everything from the common cold to rare diseases, alternative and out-of-the-mainstream political and social viewpoints. There's also an enormous amount of reference materials. You'll find access to library catalogs, national and international phone books, dictionaries, and quotation books. The Web is also a prime source for locating international information such as country profiles and trade and exporting statistics. There's also quick access to current news headlines; a path to hard to find and obscure information; and leads to all sorts of experts.

Of particular benefit to journalists has been the explosion of media sources on the Web: newspapers, journals, directories and so on. According to the Editor & Publisher Medialnfo Site <http://www.mediainfo.com>, the Web contains more than 3,500 newspapers from around the world

What's Not on the Menu

At the same time, though, there remains some notable gaps. Remember that most of the world's knowledge still remains in print. Almost all "older" information which in the Web world means before the early 90s resides in paper or microform and can't be accessed via the Internet.

Few books exist on the Web, and while thousands of newspapers and journals that are now accessible on the Web, even more are not. So the Net has certainly not—at least yet—eliminated the need to visit your good old library or to hunt down the office almanac.

News on the Net

What about newspaper archives? Here the Net is something of a mixed bag. On the one hand, there certainly is a lot available. Out of the 3,500 plus newspapers from around the globe that have launched online versions, only a certain percentage also allow archive searching.

But there are problems with the newspaper archives on the Net. Most only go back a couple weeks or a few months, compared with the 10 to 15 years' worth of articles available from a traditional online service like Dialog or Nexis. Also newspaper archive searching on the Web is not always free. Many publishers, as a means to support their Web sites, charge for downloading articles. But perhaps the biggest drawback is that in most cases you can only search one paper's archive at a time. Again, with a Dialog or Nexis you can search dozens of papers from multiple publishers simultaneously. (It's true that both Dialog and Nexis are now or will soon be making their databases available over the Internet; however, they remain expensive subscription only services, and a search engine does not index their pages).

Some of these limitations may be evaporating, though. Recently, a few sites have emerged that allow some level of multiple newspaper archive searching. TotalNews <http://www.totalnews.com> and NewsHub <http://www.newshub.com> provide access to the archives of dozens of newspapers as well as other news services such as wire services and Web-based news outlets like C/Net). However, the extent of the archives on these sites varies wildly based on the individual paper. Some go back just a couple days, others many months or a year or more.

Another noteworthy news archive site is American City Business Journals <http://www.amcity.com>, a publisher of business journals that permits free and extensive simultaneous archive searching on its 37 regional business publications. And finally, a site called NewsLibrary <http://www.infi.net/newslibrary>, allows one to search the archives of about 50 large and mid size Knight-Ridder dailies, all the way back to the mid 1980s. You can search for free, but retrieval costs $1-$2.50 per article.

Another problem with searching newspaper archives on the Internet is that Web-based newspapers often do not have the same content as their print counterparts. What's posted on a Web version of a newspaper is different than what you'd find in that paper's print copy, or even available on an online service like Dialog or Nexis which is actually a digitized version of the print edition.

On the plus side, a Web newspaper typically contains more high tech stories than its companion print version, and may be updated much faster, but, notes Eric Meyer, a journalist, who manages AJR/NewsLink <http://www.newslink.org>, a Web site that collects information on online media, a Web newspaper site usually omits certain items found in print. This may include classifieds, births, deaths, real estate, accidents, crimes, awards and promotions and smaller stories.

There are other content differences between Web and print newspapers. Local information is still in short supply on the Web, (though recent efforts by America Online's Digital Cities, Yahoo!, and Advance Publications Internet sites like New Jersey Online are filling some holes). And Steve Weinberg, author of *The Reporters Handbook* and past president of the Investigative Reporters and Editors Association, finds a lack of good investigative journalism on the Web. (See "Can content-providers be investigative journalists? The worries of a veteran reporter." *Columbia Journalism Review,* v35, n4, p36(3) Nov-Dec, 1996).

For these reasons you can't just assume that if you search or link to Web papers, you won't need to read the original print edition. To see what's contained in the print newspapers, you still either need to go to the library, or ask your news librarian to conduct a search of print newspapers online.

Keep in mind too that just because some information might be able to be found on the Net, that this doesn't mean that it's the best place to turn. It many cases a standard print reference is quicker and more convenient. For example, if you need the capital of Uruguay, or need to know the number of physicians in the U.S., why not check a copy of the *U.S. Statistical Abstract* that's likely sitting on a shelf nearby? And an encyclopedia or a dictionary are still usually the best tools for looking up quick facts ranging from spelling questions to a succinct description of some matter. Many popular directories are still found primarily in print as well—some of the most useful ones include the *Encyclopedia of Associations,* and the various newspapers and magazine directories, such *as Editor & Publisher Yearbook, The Standard Periodical Directory* and *Ulrich's International Periodical Directory.*

Starting a Search

Once you know when it makes sense to go to the Internet and are aware of its limitations, your next challenge is deciding how to go about performing a search. The place to begin any research on the Net is one of the popular search engines, but which one should you choose?

According to the Media in Cyberspace survey, reporters do have some favorites. The most popular entry points are Yahoo! and AltaVista, which together represent about one half of all sites listed by the respondents (43 percent selected Yahoo! and 19 percent chose AltaVista). Sarah Cohen, a trainer for NICAR, the National Institute for Computer Assisted Reporting corroborates that from her experience with her students, journalists seem to most frequently use Yahoo!, AltaVista (or Excite) for searching the Web.

But all search engines are not all alike, and few reporters select a particular search engine based on some rational criteria. Brant Houston, executive director of the Investigative Reporters and Editors Association (IRE) located at the University of Missouri, says few journalists really know ahead of time how search engines differ or which they should be using. A reporter at *The Cape Cod Times*, one of the most knowledgeable Internet users on her paper, admits that she just uses "whatever is up on the screen", which most of the time is Yahoo!.

But the decision to pick one search engine over the other is critical because the choice will affect the results.

Not all search engines operate the same way. In fact, some Web search tools, such as Yahoo!, are technically not even search engines though they often get lumped together with them. Here's a quick tutorial:

On the Net, there are really two major categories of Web research tools: hierarchical indexes and robots. Yahoo! is the most well known of the hierarchical indexes—which provides users with categories of subjects (e.g., business, computing, entertainment); and then within these broad categories, there are subcategories (e.g., under business, a subcategory is company information) and then within those subcategories there are smaller sub-subcategories (e.g., under company information, you would see international company information) and so on. Searchers look for sites in Yahoo! by "drilling down", that is, first selecting a broad category, viewing the narrower subcategories, choosing one of these and clicking and so forth, until the desired sites are located. Another, newer site that works the same way is called LookSmart <http://www.looksmart.com>.

There are two key differences between a site like Yahoo! or LookSmart and a true search engine. First, Yahoo!'s listings are created by people, professional indexers, editors and researchers who surf the Web to find sites and then categorize them under Yahoo!'s headings. Second, Yahoo!'s listings are very selective. Contrary to what some may assume, Yahoo! is not, by any means, a comprehensive index of sites on the Web. It represents subjective, but informed selections.

Then there are the true, search engines, called "spiders" or "robots" that roam the Web, looking for information. The big names in this category are AltaVista, HotBot, InfoSeek, Excite and a newer one called Northern Light. These spiders are actually software programs that scour the Web automatically, locate Web pages, and then index the words on those Web pages. This is all performed without human intervention.

When should you use which kind of search tool? If your subject is very obscure or arcane, use a spider—they search and index, by far, the greatest number of Web pages—literally tens of millions. But if your subject is very broad and you are worried about getting swamped with sites, you should use a reviewing service or a hierarchical index like Yahoo! or LookSmart. Although you won't be performing anything near a comprehensive search of the Web, you will at least be getting some human assistance in selecting what are hopefully, the most valuable sources.

Creating a Search Statement

It's the rare journalist who is well versed in the ins and outs of conducting proficient keyword searches on a search engine but the basic rules and guidelines can be followed by anyone, especially a quick reporter. Here are six basic rules to follow.

1. If you don't find what you need, try another search engine.

Although each of the major search engines claim to do the most comprehensive job in scouring the Web, the fact is each one can only do a partial job. Furthermore each one uses its own proprietary algorithm for determining which Web sites are most relevant to your search (e.g., the formula it uses to examine how your keywords match up to the words on the site). If a search on one engine doesn't turn up what you're looking for, its quite common that the same search on another search engine will.

2. Use synonyms.

If you suspect that the information you seek is on the Web, but your searches aren't turning anything up, think of a few synonyms for your keywords. Say, for example, you're looking for statistics on crimes committed by juveniles. And, let's say that you keyed in the words: "crimes committed by juveniles", but didn't come up with useful sites. You should then think of other ways that those concepts could be expressed: e.g., instead of the word "juveniles", perhaps you should try words like "youth" or "teenagers" or "children." Even the word crime could be substituted with "offenses" or "law breaking."

3. Know when and how to use "Boolean operators."

Most search engines are set up so that you can just enter a string of keywords that best describe your topic, and those words will be used by the search engine to determine the relevancy of Web sites. If you chose good keywords, this relevancy search should work fine. But sometimes you need to perform a more advanced search, one that utilizes Boolean operators. These allow you to construct a more precise search.

All the leading search engines allow users to employ Boolean operators as a way to construct a more precise search. While each search engine has its own protocols on how to employ a Boolean operator (see, for example, AltaVista's protocols below), the basic concepts are the same. The basic Boolean operators are the words AND, OR, and NOT. They are used to specify what specific type of operation should be applied to the words that immediately surround those operators.

Here's an example:

Say you are searching for Web sites that discuss leukemia in children. This means that you want to locate sites that discuss BOTH concepts—leukemia AND children. Your search statement, then, could be:

leukemia AND children

In another search, say for an article you were researching on heart disease operations, you might be interested in seeing sites that discussed bypass surgery or angioplasty. In that case, your search statement might be:

Bypass surgery OR angioplasty

And, in another example, say you were working on a piece on insomnia, but wanted to exclude all the sites that discussed snoring, you could enter:

Insomnia NOT snoring

Some search engines (such as AltaVista, below) allow you to string a long series of Boolean phrases by the use of parentheses, so you can construct more complex searches. So, for example, if you wanted to find Web pages that discussed elephants or lions in zoos or circuses in either Russia or China, you might have a search statement like:

(elephants OR lions) AND (zoos OR circuses) AND (Russia OR China)

The parentheses serve to tell the search engine which part of the statement to "work on" first to avoid confusion. One caution, though, when using parentheses is that most Web search engines appear to do a better job with simpler searches. The more complex and involved they become, the less effective the results.

Keep in mind that although Boolean searching does permit you to create more precise searching, you can't assume that an advanced Boolean search will always automatically retrieve better results than simply entering several keywords in a plain simple search. The reason is that entering Boolean operators can "override" a search engine's relevancy ranking algorithm—that's the formula that the search engine applies to Web pages to determine how relevant they are for your search. So when you want to be comprehensive or are looking for hard to find information, you may wish to perform both a simple keyword search as well as a Boolean search, since your results will likely be different—even on the same search engine!

4. Know how to perform phrase searching and wildcard searching.

In addition to understanding how to use Boolean operators, the other vital bit of search knowledge is knowing how to construct phrase searching and wildcard searching.

Phrase searching is simply instructing the search engine to treat two or more words as a single phrase (e.g., "doctor assisted suicide", "chronic fatigue syndrome", "budget bill", etc.). If you're looking for information on frozen yogurt, you don't want to find Web pages that happen to contain the word frozen and the word yogurt somewhere on the page, since they may have nothing to do with each other. Many search engines (such as AltaVista, see below) let you indicate phrases by enclosing those words in quotation marks.

Wildcard searching allows you to search not just for the exact word that you've entered, but for variations of that word. For instance, if you instructed a search engine to perform a search on the word: "sleep", you'd be instructing the engine to search also for pages that include, say, "sleeps", "sleeping", "sleeper", and "sleepy."

Search engines vary in how they handle wildcard searching. Some will automatically assume you want it, others won't let you perform it at all, and the most flexible ones, like AltaVista, let you specify when you want to use a wildcard by entering a specific symbol following the root word (the asterisk in this case—so you'd enter: sleep*).

5. Know how to search "fields".

In the information world, units of related information are called "records." A record can be an article, an abstract of an article, a bibliographic citation, a Web page, and so on. A field" is a specified portion of a record. For example, for Web pages, fields may include:

* The title of the Web page
*The date that Web page was last modified
*The Web page's URL (the Web address)
* The meta tags of the Web page (this is the textual description of the page that's written by the creator of the page to inform search engines what the page is about, but the text is invisible to users)

Some search engines allow you to instruct the engine to look for your keywords ONLY in a specified field. When would you want to do this kind of more limited field search?

Say you're doing a search on Vietnam. If you did a keyword search just on that word, you'd be swamped with sites, and, to make matters worse, for many of those sites, the word Vietnam would probably just be mentioned in passing and would not be the primary focus of the page. But, by restricting your search to the title field, you'd help ensure that that keyword is of primary focus of those Web pages returned.

If it's recent information you need, some search engines like AltaVista let you search a date field, so that you can specify that you only want Web pages that were recently updated.

Sometimes you may only want to see Web sites that were produced by a specific organization, or by a particular type of institution. In these cases, you can search the URL field. This is the part of the Web site that contains the page's address <e.g., http://www.website.com>.

There are at least two types of situations where you might wish to search the URL field.

1. If you want to retrieve Web pages from a specific firm or organization. Say, for instance, that you're doing a story on Sony Corp., and you want to locate Web pages produced by that company. Entering "Sony", though, in a search engine will retrieve any Web page that contains that word, not just pages created by the firm. But you can zero in on Sony's own sites by searching for the word Sony only in the URL. The URL normally contains the name of the institution that created the page.

2. If you want to see sites that are from a particular type of institution: i.e., a commercial or business site, a governmental one, a nonprofit or an academic site. URL's are coded so that the last "piece" of the address designates the type of organization: academic sites end in "edu"; business and commercial sites end in "com"; organizational/non-profits end in "org"; and government created sites in "gov". So, for example, if you were looking for statistics on aircraft deaths, but only wanted the data from official governmental sites, you could instruct the search engine to only return sites that end in "gov".

6. Modify your searches when needed.

As any librarian knows, getting good results from an online search often means tinkering with the initial search statement. If an initial search turns up no Web pages, too many, or irrelevant ones, it's a good idea to examine and possibly alter that keyword search statement. If your first search doesn't work well, you can do this by trying the following:

* Check the spelling of the words you've entered. Since computers are nothing if not literal, any misspelled words will destroy a search.
* If you are searching on some unfamiliar term or in an unfamiliar field, make sure that the words you've used are standard terminology. For example, if you were searching for Web pages on "static electricity" it would be important to know that the technical term for that phenomena is "electrostatics".
* Think of synonyms for the words you've used (as discussed above).
* Double check your syntax if you've used Boolean operators.
* If you did use Boolean operators:
...and you didn't get enough hits, take out any AND operators you entered.
...and you got too many hits, either add an AND operator to make your search more restrictive, or eliminate one of the keywords you entered, or limit your keywords to a title field.

Even if you've done a perfectly good search, you still can't always assume that the Web pages that the search engines returned to you are really the most relevant. Why? Some Web site creators are manipulating their pages and the keywords on their sites with the specific goal of getting high rankings by the search engines.

According to Danny Sullivan who publishes an e-mail newsletter called *Search Engine News* <http://www.searchenginewatch.com> some sites do this by developing what are called "bridge" pages. A bridge page is a page that is created specifically for the purpose of receiving a high ranking from a search engine. When that page is retrieved, and you click on that decoy-like page, you are routed to the site's actual home page.

In his newsletter, Sullivan gave an example of how one large company uses bridge pages to receive top rankings. Sullivan found that State Farm Insurance submitted several pages to a variety of search engines, and each page was different and was specifically written to rank well for each search engine's ranking algorithms. Furthermore, several pages were submitted to each search engine; each page was written to "capture" a particular sought-after insurance-related keyword search. For example, some of State Farm's bridge pages were created to rank highly for a search on "auto insurance," another for "boat insurance," and so on. Furthermore, several different types of pages were created for each of those subtopics, again to further increase the chance of high placement. When a searcher clicked on one of those dummy pages, he or she would be routed immediately to State Farm Insurance's actual home page.

Another way that Web site owners are attempting to improve their position on search engines is by contracting the services of consulting firms that claim they can help Web site creators get higher rankings. These companies take their clients' Web pages, run them against the major search engines, analyze placement of the page, and then provide specific strategies on how to improve rankings. One of these firms, called Webposition <http://www.webposition.com> also provides some free advice on its site on how to improve rankings on a search engine.

Things are not, then, always what they appear to be in the brave new world of the Web! On a traditional online service, like Dialog, you could assume that if you conducted a good search, you'd pretty much get what you'd expect. But on the Web, it now appears that the intention of the creator of the data is another new force that impacts what you'll retrieve.

Also, although popular search engines like to claim that their robots scour the entire Web, the fact of the matter is, no search engine can actually index all the pages out there. In fact, a recent study published in the April 2nd 1998 issue of *Science* magazine reported on an extensive study by Drs. Steve Lawrence and C. Lee Giles, of the NEC Research Institute, (the article was titled: "Millions of Web Pages Overwhelm Search Engines"). The researchers posed over 550 scientific search questions to the five largest search engines, and discovered that there are about 320 million Web pages—but that search engines only index about 40% of these!

Remember your Librarian

In your enthusiasm for doing research on the Net, don't forget to take advantage of your in-house librarian. If your newsroom is lucky enough to have a news librarian, get to know this person and know how to use his or her expertise. While you can likely learn enough about Web searching to do many straightforward searches yourself, certain types of research projects are still best left to the pros.

When should you go to a librarian for help? Basofin advises going to the librarian for extensive newspaper archive searching, or simply when you're uncomfortable with performing a certain search. Sam Meddis, the Online Editor for *USA Today's* Web site <http://www.usatoday.com> says he likes to use his librarians as fact checkers—he turns to them when he finds something on the Web that looks interesting, but wants further verification.

Other research situations where you may wish to get help from your librarian include long-term projects that involve tracking some issue over time, very complex search questions, for searching tricky scientific and technical databases (e.g., patent, chemical databases), or if you need to dig up older data.

Remember too to involve your in-house librarian when creating an initial research plan. When you begin researching a new story, it's worth sitting down with your librarian and brainstorming potential sources and information finding strategies. These people are trained information experts, and they can be invaluable founts of knowledge.

The librarian profession is actually in the throes of great change, as a result of the rise of the Internet and new technologies. Explains the *Post's* Drezden, "We're getting away from the 'go fetch' type of role and evolving into information trainers, facilitators and consultants." News librarians today, he says, are now in the position to "assist reporters in learning how to use the Web, identify possible sources, get them up to speed on using search engines, and help them evaluate the quality of information that is available on the Net."

Basofin likens this new role of the news librarian to that of a "coach, trainer, and technology facilitator." So although it's important that you have the skills do be able to do your own research in most cases, don't feel you have to always go it alone.

SIDEBAR: Searching AltaVista

According to surveys conducted by MediaSource and the University of Miami, the Web search engine used most often by journalists is AltaVista. If this is your search engine of choice, or you want to make it so, here are the most important tips and procedures to keep in mind for doing effective research:

1. Know the difference between a "simple search" and an "advanced search"

* Simple Search

When you call up the AltaVista Web site (http://www.altavista.digital.com), you are automatically linked to its "simple search" page. To conduct a simple search, you just enter keywords that best describe your research subject (e.g., Children Prozac) without Boolean operators. The simple search engine works via a "relevancy" engine, meaning that it turns up sites for you that it thinks are most relevant. To determine the relevancy of a particular Web page to your search, AltaVista looks at various elements on the Web pages, such as the rarity of the words, how close they are found together, and how close they are to the beginning of the document. The most relevant documents, as determined by AltaVista, are placed at the top of your returned list.

* Advanced Search

In the "advanced search" option, you can use the operators: AND, OR, NOT, and NEAR to establish a relationship between your keywords. To call up the "advanced search" screen, you just need to click on the "advanced search" box displayed in the upper left hand portion of the simple search page.

Earlier I discussed the use of AND and OR, but AltaVista also gives you another Boolean type operator that it calls NEAR, which can be even more useful than "AND". The NEAR operator will only return sites where your keywords are separated by no more than 10 words. It is a particularly powerful tool for doing real precision searching. So, for instance, if you entered Children NEAR Prozac, the sites returned would be those where the word "Children" and "Prozac" were within 10 words of each other, and would establish the relationship between the words is likely to be very close.

(Note, one quirk in AltaVista is that to use the NOT operator, you actually need to enter the words "AND NOT"; Example: Computers AND NOT IBM.)

* Phrase Searching

To search for a phrase, enclose these words in quotation marks. For example, to search for Web pages about chronic fatigue syndrome, you'd enter: "Chronic Fatigue Syndrome".

* Wildcards

As described above, wildcard searching lets you search for word variations. On AltaVista, you designate the wildcard with an asterisk. So, a search of the word: "bank*" would also retrieve "banks", "banker", "banking", etc.

Tip: be careful when using the wild card--you can get irrelevant results too (For instance, you wouldn't want to do the search "car" to find references to "car" and "cars", since you'd also retrieve words like "carry", "carpet", "carp", etc.).

* Fields

Two very useful fields in AltaVista are the "title" field, and "url" field. These are useful for limiting your searches, as described above. On AltaVista, you specify that you want to do a field search by preceding your keywords with the field name and then a colon. For example, to search for Web pages that had the phrase "welfare reform" just in the title field, you'd enter:

title: "welfare reform"

Special AltaVista Search Tips

* Words keyed in lower case will retrieve pages with words in either lower or upper case. But if you enter a word in upper case, AltaVista will restrict its search to only pages where the word appears in upper case.

* In "simple search," since the most relevant pages are placed at the beginning of the list, you normally don't need to read past the first 2-3 screens worth. If you don't find what you need there, modify your search or try a different search engine.

* To find only the most recent documents, go to the Advanced Search page. Underneath the search statement box there are two boxes where you can enter "start" and "end" dates. By putting a date in the "start" box you can limit the search so that only Web pages that have been modified during the latest month, several months, year, etc. are returned. Caution: the amount of time it takes for a newly created or modified Web page to be indexed on a search engine can vary quite a bit. Often, the newest Web sites indexed on AltaVista are already at least a month old.

* When you use AltaVista's Advanced Search box, there is a "rank results" box below the search screen. You can use this box to tell the search engine which Web pages it should put towards the beginning of the list. This is an important tool for ensuring that the most relevant sites appear at or near the top.

* Very common words: words like it, the, but, with, etc. are ignored by AltaVista. However you can bypass this by using quotation marks. In other words, if you entered the slogan just do it you'd get nothing back since AltaVista would ignore those very common words; instead, you'd need to enter: "just do it".

<u>But is it True?</u>

Even if you've done your search correctly, and retrieved a nice neat list of relevant hits, you still have one more critical job to do before you can consider your search complete: you need to determine whether the information you've uncovered is trustworthy.

Bad information on the Net is an issue that's gotten a lot of press recently, and deservedly so. There is a great deal of deceptive, bogus and junk information out there, ranging from scholarly looking but discredited Holocaust "revisionist" sites to conspiracist ravings from the far right to the far left. But the fact of the matter is, how you verify information from the Net is really not much different than how you'd verify any other data.

The key, as it always is and has been, is to first find out about the source. Who are they? Do you know them? What is their reputation? What is their agenda? You can first make a broad determination about the source by examining the last three letters of the Internet address (the "url"). Government sites end in "gov", educational institutions end in "edu", nonprofit organizations in "org" , and for-profit commercial sites in "com". That at least gives you a start. Then it's up to you (or your news organization) to decide how much credibility you're going to put into the particular types of sources.

For example, some newspapers have a policy of relying only on "official" sites—typically that means governmental sites. These are the public sites like the FAA, the SEC, a state's environmental department, The British Embassy, and so forth that release official data. Don Burgess, a reporter at the *Bermuda Sun* newspaper says his paper has "an unwritten policy" of using only official sites as accredited sources. He says, for instance, that in a story his paper did last year on a U.S. military cleanup of a base on the island that the article picked up quotes from senators off the Web, but only as reported on the official senate Web site.

Other newspapers and journalists will use information from commercial sites, but only those that have a recognized "brand name." For instance, Tim Malloy, who edits the publication *Internet Newsroom* (Glen Echo, MD) says that he trusts what he finds on CNN's Web site; Drezden feels comfortable using information that he gleans from other newspaper's Web sites.

It's helpful to collect and bookmark your favorite and most relied upon Web sites. David Crumm, religion reporter at the *Detroit Free Press* has his set of trusted sites bookmarked and grouped into subject folders; e.g., "Buddhist Resources." Sarah Cohen, a trainer at NICAR believes that having a set of trustworthy bookmarks is, in fact, a more effective way to use the Net than trying to perform a search for each new story, since Web searching can be time consuming, and, you're never too sure about the quality of what you eventually come across.

Journalism 101

But what do you do if you stumble across what seems to be a good tip from an unofficial or unknown source? Well, again, you'd do just what you'd do if you got a tip from any other unfamiliar source, whether she called you up on the phone or walked into your office. You'd need to rely on good old fashioned legwork and reporter instinct.

For research on the Net, this means first making sure that you're getting the story straight from the original source. So, for example, say you were working on a story on automobile emission pollution control technologies and linked to an anti-automobile activist's Web site, and that site displayed a chart with the latest U.S. Department of Transportation statistics on new car emission violations. You'd then want to link to the DOT's own Web site directly to verify that data.

If the site is not an official one, you'd contact the source directly. Says Burgess, "we would never quote from an unofficial site unless we could talk with the owner. A simple procedure would be to e-mail the person first requesting a phone interview. I did this with an owner of an asbestos waste disposal company who has a procedure for turning asbestos into paving for roads." (Bermuda, being only 21 miles by 1.5 miles, has no place to dump asbestos and Greenpeace threatened demonstrations if the government dumped it at sea.)

Once you're satisfied that the person or source is who they say they are, and are credible you'd then do as you would do with any other uncertain information, verify that data with a second source. Again, it's really Journalism 101.

This isn't to say that the Internet doesn't pose any new challenges for you. The Internet has, in fact, changed the way information is created and disseminated. Before the Net, in order to get a message out, one would need access to a printing press and some means of distribution, like delivery trucks or a satellite. But the Web has demolished both of those barriers of entry to be a publisher and has opened up the information floodgates. "Don't believe those are the lyrics of Jimi Hendrix unless you know his songs" warns Paul D'Ambrosio, database editor at the *Asbury Park Press* (Neptune, New Jersey). The millions of Web pages created by individuals represent an enormous source of anecdotal and personal material never available before—at least not so much and not so easily. Anyone can be a publisher now.

Karen Koek, project manager for Gale Research's Cyberhound Internet directory service, believes that this phenomena means that as a society "we've peaked in getting information just from the scholarly". In other words, today a knowledgeable source may not have to have a PhD after his or her name, or been published. When evaluating a Web site that contains information based on an individual's own experiences, she maintains what she calls "an open mind but a healthy skepticism." Koek looks at whether the person's reasoning is sound. She also finds out whether their opinions were based on a single incident, or something that occurred over a long period of time.

Ultimately, you just may need to fall back on your instincts as a reporter. Woodward says that by reading through the material on a personal Web page, you can get a sense of the person's background, and whether they know what they're saying. You can also find out their credentials by contacting them and asking questions, and judging how well and openly they respond to queries.

It also helps, says Crumm, to know something about a subject area yourself. He says that there are tons of Catholic-oriented sites on the Web, and some look pretty "spiffy" but don't always have great content—but because he knows a good deal about the religion, he can sort the good from the bad.

Ultimately, all this information surely can give you more headaches, but the bottom line is that it provides you with more information and more sources. Meddis says that journalists should be viewing all of this new information on the Net as a bonanza by journalists, and "stop whining" about questionable information.

Not all of the burden of assuring information quality should fall to you as an individual reporter, though. These issues should also be addressed by your news organization. There should be some kind of overall policy on the use of information from the Net (at the *Asbury Park Press*, for example, there's a rule that the paper never reports any item as "according to such and such Web site"); and a policy on how information will be verified. And there should be some kind of Internet search training available. Such training could be done either in-house by the news librarian or by bringing in outside trainers.

Despite its reputation for questionable information, data on the Net could even be used as a means to verify other data. Burgess recounted a story where a local bicyclist was in a regional race and told him that he finished in the top ten. He checked the official Web site of the race organizers, and discovered that his actual standing was 64th. The red faced athlete was forced to admit that he did not tell the truth.

Information Overload

The other great problem you face today is information overload. The monster shows its ugly head most clearly when a search turns up some enormous number of hits (e.g., "your search returned 28,983 Web pages"). Crumm says that the problem is most acute when he needs to do research on a new area where he doesn't already have good reliable bookmarks and where he isn't yet knowledgeable. "I was doing a story on new translations of the Bible," he recounted, "and I had to wade through a lot of stuff, the sheer weight of it all" was frustrating.

While this is a problem that has yet to be solved, a few salient tips can be provided to help battle the ogre:

* Keep in mind that when a search engine returns Web pages, only the first "batch" of hits are likely to be truly relevant. This means that you shouldn't bother reading past the first 20-40 sites or so, or the equivalent of about 2-4 screens.

* To reduce information overload, you can choose to search a human created index like Yahoo! or LookSmart. While these directories index a much smaller portion of the Web than a true search engine, the sites are screened by human beings and were selected as being substantive and worthwhile.

* If you feel overwhelmed by what you get back from AltaVista or one of the other well known major search engines, try a newer search engine called Northern Light <http://www.nlsearch.com>. Northern Light does an admirable job in making a dent in information overload by grouping the sites it retrieves into logical folders, with each folder containing related information (e.g., all government pages grouped together, all educational pages, all pages from the same commercial site, etc.).

* Don't' forget that the best filter is not a search engine, but other people. If you can find an expert—your librarian or any other knowledgeable source—that person's smarts can cut through the information fog. Data is not information and information is not knowledge.

A search on the Net is a means to the end—and a great source for leads and tips. And like all tools the more effective you are in using it and knowing its limitations, the more you can get out of it. Just keep in mind the words of the president of one major search service, who acknowledges that the best search engine remains "the one between your ears."

Further Resources

There are several organizations and individuals that have tried to make life on the Web easier for journalists by compiling a list of reliable and useful links on the Web of particular interest to reporters. Some of these group the links into popular topic areas (e.g., "crime, economy, politics, etc.). Here are a few favorites:

<http://npc.press.org/library/reporter.htm>

<http://www.stlouisspj.org/resources.html>

<http://www.newslink.org/gref.html>

<http://www.cais.com/makulow/vlj.html>

Attachment:

Updated information on search engines. Article by author, titled, Searching for the Right Search Engine, published in the January 21, 2000 edition of the *Chronicle of Higher Education*.

Searching for the Right Search Engine

By Robert Berkman

RESEARCHERS now have it all on the World Wide Web: facts on virtually any topic, available from the far corners of the globe, unfiltered by reporters, editors, or publishers, and usually free. But sometimes we feel that we have too much information—often way too much—and that it may not be correct.

Despite the latest flurry of prime-time ads by search-engine vendors boasting that they can find anything you want online, search engines can't distinguish among Web pages based on their contents. The only way researchers can pinpoint information on the Web is if they learn how to do efficient Web searches, and which engines are best for which purposes.

One important lesson is to understand the range of search tools now available. Many researchers don't realize that they can use hierarchical indexes, standard search engines, alternative search engines, meta search engines, and databases—and that those tools are not all the same.

In a hierarchical index—probably the best known is Yahoo (http://www.yahoo.com)—people trained to categorize information, such as librarians and indexers, examine Web sites and put them in categories and subcategories. Thus, when you do a search on a hierarchical index, it is much more likely that what you find will be relevant to what you are looking for.

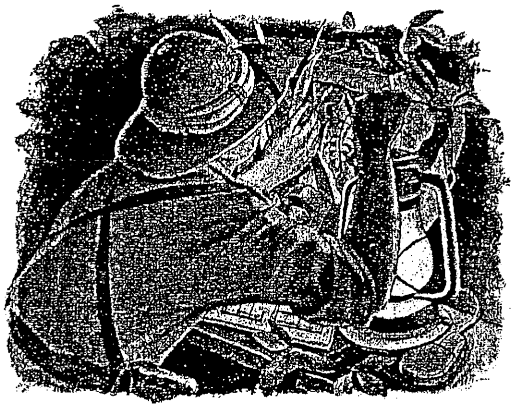
The drawback to hierarchical indexes is that they are extremely selective. Because they are created by human beings rather than by computers, they can include only a tiny portion of what is available on the Web. Of course, in these days of abundant information, that may not be such a bad thing.

Yahoo uses a standard search engine as well. For that reason, the results of a search on Yahoo are split into several sections. "Category matches" inform you if your topic matches one of Yahoo's existing categories. "Site matches" are the sites that have been indexed and categorized. "Web pages" provide links to pages located by the search engine. Yahoo also groups results into two other sections: "related news," for any news item it locates on your subject, and "Net events," which are mostly chat sites.

Yahoo is by no means the only hierarchical index, and some of the many others are aimed specifically at academic users. The latter group includes: AlphaSearch (http://www.calvin.edu/library/as), BUBL Link (http://www.bubl.ac.uk/link), and Infomine (http://infomine.ucr.edu).

Then there are the standard search engines. Popular ones include AltaVista (http://www.altavista.com), Excite (http://www.excite.com), Go Network (http://infoseek.go.com), and HotBot (http://hotbot.lycos.com). Unlike hierarchical indexes, standard search engines send out software "robots" or "spiders" to search the Web and index the pages in each site they encounter. The engines then calculate mathematically how relevant the pages are to your search terms; each engine uses its own algorithm to rank pages. Factors in the calculation include the frequency and placement of your keywords on a page, and their occurrence in the descriptions that owners write of their pages, which are invisible to users. The search engine puts the pages that get the highest score at the top of the list of results.

Savvy researchers will avoid standard search engines when they have a very broad subject. Instead, they will use a hierarchical index, to find just a few relevant, well-cataloged sites.



TOM HERZBERG FOR THE CHRONICLE

ALTERNATIVE search engines, which take various approaches to ranking and sorting the pages that they find, are often more helpful than standard engines. Northern Light (http://www.northernlight.com), for instance, ranks Web pages as a standard search engine does. But instead of displaying all of its results in a single listing, it sorts pages into categories and groups the results into folders. As an example, a search for "alternative energy" creates folders with labels such as "solar power," "air pollution," and "National Technical Information Service," which includes documents from that agency. And the folders contain subfolders. Within the solar-power folder, for instance, are folders for "photovoltaic systems" and "government sites." That arrangement of material can help you determine which groups of pages are most likely to be relevant to your needs.

Ask Jeeves (http://www.askjeeves.com) takes an altogether different approach. You don't enter keywords, but type a question in plain English—perhaps "Is there evidence of life on Mars?" Ask Jeeves has recorded millions of questions that users have asked it, and has found Web sites that answer those questions.

The first thing that Ask Jeeves does after getting your query is to scan its database of questions and answers. It then gives you a list of questions that it "thinks" you want the answer to. If you select one of them, it lists sites that contain the answers. Ask Jeeves doesn't always work, but it can save you time, and it is fun to use.

Google (http://www.google.com) takes yet another tack. Like other search engines, it first matches up your keywords to the pages it has collected in its index. Then, however, it ranks each page based on how many other pages link to it—and how many link to those pages in turn. The pages you see at the top of your list of results are those with the highest number of links to other pages. The idea is that such popularity is meaningful, just as a diner that has many trucks parked in front probably serves better food than the diner whose parking lot is empty. The approach works. After several years of being a loyal AltaVista user, I am now a "googler."

Oingo (http://www.oingo.com) has an even more radical approach. The site's slogan is "We know what you mean," and Oingo conducts a "conceptual search" to make sure that it understands your request. Ask it to search for "china," for example, and it will ask you to choose "porcelain" or any of the various geographical Chinas. Once you make a selection, Oingo will display "directory hits" and "Web hits." The site combines a hierarchical index and a search engine (it uses AltaVista), although the conceptual search applies only to its directory results.

Search engines that search other engines are called meta search engines. Among the popular ones are Dogpile (http://www.dogpile.com), Inference Find (http://www.inferencefind.com), and MetaCrawler (http://www.metacrawler.com). The concept here is that because no single search engine indexes the entire Web, using a meta search engine allows a researcher to scan more sites. The downside is that such an engine needs to use a "lowest common denominator" search statement, so that all of the search engines that it searches understand the request. Therefore, meta search engines are not a very good choice for complex searches, involving, say, Boolean logic. (Dogpile does include some Boolean-search capabilities.)

A completely different strategy is to search a database on the Web. Hundreds of databases originally searchable on CD-ROM or through proprietary online dial-up services are now available on the Web, and new databases are continually being born there as well. That makes it possible to search rich databases with a standard Web browser, although in many cases, the researcher must pay a fee or be affiliated with a university that subscribes to the database. The fee-based sites typically filter the data they contain, increasing the likelihood that the results will be relevant to a search; many also offer superior search capabilities, so requests can be more precise.

The many new, free databases on the Web can also be helpful. A site that does an excellent job of identifying and sorting free databases is The BigHub (http://www.thebighub.com). Through its "specialty search categories," it allows you to search more than 1,500 databases on the Web, many of which are oriented toward academics.

WHAT new tools for searching the Web are on the horizon? At a recent conference, I heard about "vortals," vertical portals that provide information from only a designated slice of the Web. For example, a vortal might search only those sites and pages that have to do with health care. VerticalNet (http://www.verticalnet.com) offers portals to industries including communications and advanced technologies. Although the concept is a good one, the jury is still out on vortals' usefulness.

Farther down the road are visual representations of search results. Those search tools display their results graphically, allowing you to see at a glance which items are the most relevant. A service called NewsMaps (http://www.newsmaps.com), for example, displays the results of your search as a thematic map. Topographical markers indicate clusters of similar documents—the most similar ones are piled up into little hills. According to Cartia, the company behind the technology, the maps are created automatically by an algorithm that "reads documents, extracts the content, and organizes the collection into a map." You can view some sample maps at the site.

No matter which search tool you choose, you will get the best results if you know what information you need, know the advantages and disadvantages of the various ways to search the Web, and regularly practice doing research online. Despite technological innovation, the best research tool remains the human brain.

Robert Berkman is a member of the faculty of the graduate media-studies program at the New School University, and conducts workshops on searching the Internet. He is the author of Find it Fast: How to Uncover Expert Information on Any Subject, *the fifth edition of which will be published by HarperCollins in May.*