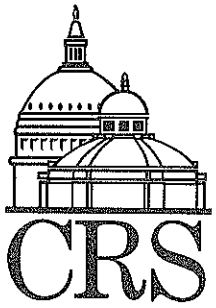
Order Code 97-556 C

# CRS Report for Congress

# Point & Click: Internet Search Engines, Subject Guides, and Searching Techniques

Updated October 24, 2000

Rita Tehan
Information Research Specialist
Information Research Division




Congressional Research Service • The Library of Congress

The Congressional Research Service works exclusively for the Congress, conducting research, analyzing legislation, and providing information at the request of committees, Members, and their staffs.

The Service makes such research available, without partisan bias, in many forms including studies, reports, compilations, digests, and background briefings. Upon request, CRS assists committees in analyzing legislative proposals and issues, and in assessing the possible effects of these proposals and their alternatives. The Service's senior specialists and subject analysts are also available for personal consultations in their respective fields of expertise.

# Point & Click: Internet Search Engines, Subject Guides, and Searching Techniques


## Summary

This report discusses criteria to consider when judging the quality of an Internet site and the best strategies for locating information on the World Wide Web (WWW). It includes a discussion of how to evaluate a Web site's caliber and merit.

There are two ways to search the Internet. The first is to use subject guides (e.g., Yahoo, Galaxy, or WWW Virtual Library), which are compiled by human indexers. These present an organized hierarchy of categories so a searcher can "drill down" through their links.

The second option is to use a search engine (e.g., AltaVista, Google, or Hotbot), an automated software robot which indexes Web pages and retrieves information based on relevancy-ranked algorithms. In addition, there are specialized search engines devoted to a particular topic (e.g., HealthFinder, LegalEngine, or GovBot). Some newly developed search engines (e.g., Oingo, SimpliFind, or WebTop) allow searchers to use natural language concepts in their searches.

In addition to discussing Internet searching techniques, this report describes how subject guides are compiled and how search engines index the WWW, as well as various features common to most search engines. In addition, the report suggests searching tips for retrieving the most precise information.

The report discusses Usenet news groups, e-mail discussion lists, gophers, and miscellaneous Web resources. This report will be updated from time to time.

## Contents


Challenges of Internet Searching . . . . . 1

Standards for Determining Information Quality . . . . . 3

Where to Start . . . . . 4

Subject Guides . . . . . 5

Specialized Subject Directories and Search Engines . . . . . 5

Search Engines: Spiders, Crawlers, Robots . . . . . 6
- New Generation Search Engines . . . . . 8
- Search Engine Features . . . . . 9
- Search Tips . . . . . 9
- Some Common Problems . . . . . 10

Usenet News Groups and E-mail Discussion Lists . . . . . 11

Gopher versus Web . . . . . 13

Miscellaneous Sources . . . . . 13
- Internet News . . . . . 13

Glossary of Selected Internet Terms . . . . . 14

# Point & Click:
# Internet Search Engines, Subject Guides, and Searching Techniques

## Challenges of Internet Searching

Finding information on the Internet can be challenging for even the most experienced searchers. Since the most popular means of accessing the Internet is through the World Wide Web (WWW), this report focuses on search strategies that locate Web information. Some search engines index gopher[1] and FTP (file transfer protocol)[2] sites as well as Web sites and Usenet newsgroups.[3] When the most comprehensive search is needed, it might be necessary to search gopher and FTP sites using the Archie and Veronica programs.[4]

If a searcher enters a simple query, such as "African elephant," into any of the top World Wide Web search engines, the resulting sets can range from 8,545 hits in AltaVista, to 2,412 in Infoseek, 13,239 in Northern Light, and 12,400 in Google. A quick review of the results shows some relevant hits near the top of each list, but retrieving so many items is usually counterproductive. Since there is no central catalog of Internet resources, a searcher must find other ways to retrieve more precise, relevant, and useful information. This report will suggest a number of strategies, tips, and techniques to use.

In May 2000, a study by NPD, a marketing information provider, reported that 82% of visitors find the information they are looking for at search engines most or all

---

[1] See Glossary for definition.

[2] See Glossary for definition.

[3] Usenet is a collection of e-mail messages on various subjects that are posted to servers on a worldwide network. Each subject collection of posted notes is known as a newsgroup. There are thousands of newsgroups.

[4] Archie helps find files available at file transfer protocol (FTP) hosts. When searching for a particular term, Archie searches the database and displays the name of each FTP host that has that file or directory and the exact path to that directory. See *Archie Services*, a gateway to Archie servers on the Web at: [http://archie.emnet.co.uk/].

Veronica is an indexer that can query every gopher on the gopher system to search for a keyword or phrase in a menu title and give the address of all menus with those key words. See: [gopher://munin.ub2.lu.se/11/resources/veronica].

of the time.[5] This is surprising because most searches yield many more results than could possibly be examined by the average searcher. In addition, relevancy rates fall off sharply after the first few dozen results. However, the NPD study explains that "search engine users would rather tinker with unsuccessful searches to find information in their favorite search engines than visit alternate sites. Many users think every search engine will provide the same information, so they stick with the search sites they know.

In addition, in a study published in the peer-reviewed journal *Education Policy Analysis Archives*, most people are conducting few searches, poorly formulating their questions, not using available tools, and are examining only a few potential resources.[6] The typical patron spends six minutes looking for information, composes two or three queries, and examines only three or four potentially relevant citations or "hits." The typical query is composed of only a word or phrase, with less than half the queries containing an "OR" to incorporate alternative terms. Even studies that noted a high level of user satisfaction observed that users rely on overly simple searches, make frequent errors, and fail to attain comprehensive results.

Search companies have long been aware that they are indexing less and less of the Web. There is a point of diminishing returns if a simple query retrieves thousands of hits. The question is not *how many* results are found, but which are the *most relevant* for the user. The interest in relevancy over comprehensiveness is cited in the literature as a leading reason why most of the search services have not made a bigger effort to substantially increase index size.

The NEC Research Institute computer scientists believe that search engine coverage will eventually equal the Web's growth because the rate of increase of computational resources is faster than the rate of increase of humans' production of information.

The tools that are available today are going to change, and there will be new and different ones a month or a week from now—or tomorrow. Ultimately, you will find a handful of useful sites by trial and error. Bookmark[7] these and return to them for future reference. Internet sites may change their uniform resource locator (URL)[8] addresses slightly, but usually only to move files from one directory to another. Significant Web sites seldom disappear completely. If it is a valuable resource, the organization that created the Web page has a stake in maintaining it. If the page moves, a responsible organization will provide a pointer URL to the new location.

---

[5] *NPD Study Shows Web Users See Improvements in Search Engine Sites*, NPD press release, May 9, 2000. [http://www.npd.com/corp/press/press_000509.htm].

[6] Hertzberg, Scott, and Lawrence Rudner. The Quality of Researchers' Searches of the ERIC Database. *Education Policy Analysis Archives*, August 25, 1999. [http://epaa.asu.edu/epaa/v7n25.html].

[7] See Glossary for definition.

[8] See Glossary for definition.

In addition, it is necessary to account for the "invisible Web" (databases within Web sites). According to an August 2000 study by BrightPlanet, an Internet content company, the World Wide Web is 400 to 550 times bigger than previously estimated.[9] According to this study, the Web consists of hundreds of billions of documents hidden in searchable databases unretrievable by conventional search engines—what it refers to as the "deep Web." The deep Web contains 7,500 terabytes of information, compared to 19 terabytes of information on the surface Web. A *single* terabyte of storage could hold each of the following: 300 million pages of text, 100,000 medical x-rays, or 250 movies.[10]

Search engines rely on technology that generally identifies "static" pages, rather than the "dynamic" information stored in databases. Deep Web content resides in searchable databases, the results from which can only be discovered by a direct query. Without the directed query, the database does not publish the result. Thus, while the content is there, it is skipped over by traditional search engines which cannot probe beneath the surface. Examples of Web sites with "dynamic" databases are: THOMAS (legislative information), PubMed and Medline (medical information), SEC corporate filings, Yellow Pages, classifieds, shopping/auction sites, library catalogs, etc. BrightPlanet has developed a software called "LexiBot" which searches not only pages indexed by traditional search engines, but delves into Internet databases as well.

## Standards for Determining Information Quality

Almost anyone with an Internet connection can "publish" on the Web. Some criteria to consider when judging an Internet site's quality are:

*Content.* Is the site a provider of original content or merely a pointer site to other sources? What is the purpose of the site? Is it stated? Sites containing durable, timely, fresh, attributable information are more useful.

*Comprehensiveness.* What is the scope of the information? How deep and broad is the information coverage? If the site links to other resources, the links should be up-to-date and to appropriate resources.

*Balance.* Is the content accurate? (You may have to check other Internet or print resources.) Is it objective? If there are biases in the information, they should be noted at the site. The organization's motivation for placing the information on the Web should be clear (is it an advertisement? does it support a particular viewpoint?). Generally, an organization's Web page will provide information it wants to release and nothing more.

---

[9] The Deep Web: Surfacing Hidden Value. BrightPlanet, July 2000. [http://www.completeplanet.com/Tutorials/DeepWeb/index.asp].

[10] The Life Cycle of Government Information: Challenges of Electronic Innovation. 1995 FLICC Forum on Federal Information Policies, Library of Congress. March 24, 1995. [http://lcweb.loc.gov/flicc/forum95.html]

*Currency.* Is the site kept up-to-date? If it points to other sites, what percentage of the links work when clicked on? Dates of updates should be stated and correspond to the information listed in the resource.

*Authority.* Does the resource have a reputable organization or expert behind it? Who is the author? What is the author's authority? Does the author or institution have credibility in the field? Can the author be contacted for clarification or to be informed of new information? There is nothing intrinsically deficient about amateur, club, or fan sites; in fact, they may deliver more passion and enthusiasm than professional sites. The researcher must remember, however, that many amateur sites have no standards for accuracy, no fact checkers, and no peer review board.

## Where to Start

The first thing to decide is what type of resource is needed. One possibility is to obtain information from the World Wide Web; another would be to explore information posted to special interest e-mail lists or Usenet newsgroups. Some search engines concentrate on the Web, others focus on Usenet, and others, such as AltaVista and InfoSeek, let you search both. Many search engines scan for gopher and FTP sites as well.[11]

If you are looking for general information on a subject, start with subject guides, which are compiled and categorized by human indexers (discussed below). These are organized hierarchically, so you can move from broad topics to narrower ones. Once you find the correct terminology for your subject, you can use search engines to locate additional information. A rule of thumb for a comprehensive search would be to check three subject indexes and three search engines.

You will retrieve more information from a search engine than a subject index, because software robots[12] visit many more sites than human indexers. However, human indexers add structure and organization to their indexes.

A good source for choosing the best directory or search engine for your purpose is the Nueva School's "Library Help: Choose the Best Search for Your Purpose."[13] For example, if you "need a few good hits fast," the site recommends Google, because

---

[11] For additional information on finding the best search tool for your needs, see: *How to Choose the Search Tools You Need* from the University of California at Berkeley Library (updated June 1999) at:
[http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/ToolsTables.html].
See also *Internet Searching Tools*, from Southern Oregon University,
[http://www.sou.edu/library/cybrary/search.htm].
It is a well-organized selection of search engines, subject directories, and resources on how to search the Internet.

[12] See Glossary for definition.

[13] Choose the Best Search for Your Purpose. Latest revision: August 22, 2000. [http://NuevaSchool.org/~debbie/library/research/adviceengine.html]

it "returns important relevant hits quickly." If you have a general "broad academic subject" to explore, the site recommends Northern Lights, the Librarians' Index to the Internet, or Infomine.

## Subject Guides

Subject guides typically present an organized hierarchy of categories for information browsing by subject. Under each category or subcategory, links to appropriate Web pages are listed. Some sites (for example, the Argus Clearinghouse) include subject guides that function as bibliographies for Internet resources and are authored by specialists.

The lack of a controlled vocabulary within and among different subject trees increases the difficulty of browsing them effectively. Some subject guides allow keyword searching, which is useful. Examples of well-organized and comprehensive subject guides include:

- About.com [http://a-zlist.miningco.com/]
- Argus Clearinghouse [http://www.clearinghouse.net/]
- Galaxy (formerly EINet Galaxy) [http://www.einet.net]
- Google [http://directory.google.com/]
- Internet Public Library [http://www.ipl.org/ref/]
- Librarians' Index to the Internet [http://lii.org/]
- Open Directory Project [http://www.dmoz.org/]
- Snap [http://www.snap.com/]
- WebGEMS [http://www.fpsol.com/gems/webgems.html]
- World Wide Web Virtual Library [http://vlib.org]
- Yahoo [http://www.yahoo.com/]

## Specialized Subject Directories and Search Engines

Specialized search engines or indexes focus on collecting relevant sites for a particular subject. Some examples are:

- Academic Publications: All Academic [http://www.allacademic.com]
- Education: SearchEdu [http://www.searchedu.com/]
- Engineering: Edinburgh Engineering Virtual Library [http://www.eevl.ac.uk/searchengines.html]
- Government: FirstGov [http://www.firstgov.gov]
- Government: Govbot [http://ciir2.cs.umass.edu/Govbot/]
- Government: Google Uncle Sam [http://www.google.com/unclesam]

- Government: SearchGov [http://www.searchgov.com/]

- Health: Achoo [http://www.achoo.com]
- Health: HealthAtoZ [http://healthatoz.com]
- Health: HealthFinder [http://www.healthfinder.gov]
- Health: MedicalWorld [http://www.mwsearch.com/]

- Legal: FindLaw [http://www.findlaw.com]
- Legal: LawCrawler [http://lawcrawler.findlaw.com]
- Legal: LegalEngine [http://www.legalengine.com/]

- Politics: iPolitics [http://www.ipolitics.com/community/default.asp]

- Science: BioLinks [http://www.biolinks.com]
- Science: Life Sciences - BioCrawler [http://www.biocrawler.com]
- Science: SciSeek [http://www.sciseek.com]

- Specialized Directories - Cyward [http://www.cyward.com/speciali.htm]

- Travel: bopLOP [http://www.bopLOP.com/]

## Search Engines: Spiders, Crawlers, Robots

Search engines are automated software robots which typically begin at a known page and follow links from it to others, downloading pages and indexing them as they go.[14] At its most basic level, a search engine maintains a list, for every word, of all known Web pages containing that word. The collection of lists is known as an index. Search engines vary according to the size of the index, the frequency of updating the index, the search options, the speed of returning a result, the relevancy of the results, and the overall ease of use. In reality, no two search engines work the same way.[15]

To decide on which search engine to use, it helps to understand which parts of a Web page the search engines index. All search engines do not use the same syntax.

---

[14] For more information comparing the features of different search engines, spiders, robots, and crawlers, see Web Search Engines: Features and Commands, Online Magazine, May/June 1999, p. 24-28.

See also *Comparison of Search Engine User Interface Capabilities*, from the Curtin University of Technology (last modified July 5, 1999) at: [http://www.curtin.edu.au/curtin/library/staffpages/gwpersonal/senginestudy/compare.htm], or Search Engine Features for Searchers, from Search Engine Watch (May 24, 1999), at: [http://searchenginewatch.com/facts/ataglance.html].

See also *A Higher Signal-to-Noise Ratio: Effective Use of Web Search Engines* (updated March 13, 1998), from the Wisconsin Educational Technology Conference, Green Bay, WI, at: [http://www.dpi.state.wi.us/dpi/dlcl/lbstat/search2.html].

[15] *Search Engine Watch*, from Mecklermedia, produces "Search Engine Facts and Fun," which gives information on how search engines work. Check the "Under the Hood of Search Engines" links at: [http://searchenginewatch.com/facts/index.html].

For example, some search engines index every word of a Web page, while others index the title, heading, and the most significant 200 words.

Search engines will also check to see if the keywords appear near the top of a Web page, such as in the headline or in the first few paragraphs of text. They assume that any page relevant to the topic will mention those words at the start. Many search engines ignore words of three or fewer letters, or will not search numbers or a date.

These differences contribute to the different results returned by different search engines for the same query. Search engines are not in any way comprehensive maps of the Internet. The World Wide Web is simply too vast for even the most advanced search engine to cover exhaustively.

Frequency is the other major factor in how search engines determine relevancy. A search engine will analyze how often keywords appear in relation to other words in a Web page. Those with a higher frequency are often deemed more relevant than other Web pages.

Many Web users do not realize that the results of their searches may be skewed by a new industry that has emerged to advise Web page owners about how to improve their site's rankings in search engines.[16] None of the major search engines or directories accepts payment to increase a ranking, although some, like Yahoo and LookSmart, offer an express service for a fee, meaning that sites are reviewed for listing in a few days, rather than weeks or months. Some search engines offer "keyword buying," which means that a company's advertising banner appears when a searcher types a certain word. For example, if a searcher typed "vacuum cleaner" on AltaVista, an advertisement for a particular vacuum company might appear, at least until another company pays for that term.

Commercial sites are increasingly likely to be ranked higher than purely informational ones because they are the most likely to invest their resources in trying to manipulate search engines. For example, some site owners create what are called bridge or doorway pages, which are written for the sole purpose of getting high rankings on search engines. A site may have dozens of those pages, each focusing on different keywords, and each aimed at a particular search engine's ranking formula. Once you reach one of those bridge pages, you are immediately forwarded to the site's real home page.

Parallel or meta-search engines (Debriefing, Dogpile, MetaCrawler, etc.) scan several search engines sequentially and eliminate duplicates, though not always reliably. Meta-search engines are good for uncomplicated searches of very general concepts or very narrow searches of unique words or concepts, because you cannot use advanced search techniques with them.

---

[16] Berkman, Robert. Internet Searching Is Not Always What It Seems. *The Chronicle of Higher Education*, July 28, 2000. [http://chronicle.com/weekly/v46/i47/47b00901.htm].

Examples of some useful search engines are:[17]

- AltaVista [http://www.altavista.com/]
- CNET Search.com [http://www.search.com]
- Excite [http://www.excite.com/]
- Google [http://www.google.com]
- Hotbot [http://www.hotbot.com/]
- InfoSeek [http://www.infoseek.com/]
- Lycos [http://www.lycos.com/]
- Metacrawler [http://metacrawler.com/]
- Northern Light [http://www.nlsearch.com/]

There is no "best" search engine, and one search engine is not necessarily better than another at finding different types of documents (for example, government reports, corporate press releases, or movie reviews). Search engines look for keywords, not concepts, so to find information on a particular topic, you need to create a precise search. That is why it is important to learn the advanced search syntax for a few different search engines in order to refine and narrow a query when the number of items retrieved is too large.

## New Generation Search Engines

A new generation of search engines is emerging, armed with next-generation technology. Several search engines have begun considering factors such as the number of links made to a page (Google), or the number of times a page is accessed from a results list (Direct Hit). Direct Hit measures which sites are most frequently selected from a search results list—a sort of "popularity" engine. The system observes which pages are selected from search results and how long visitors spend reading the pages. These approaches attempt to locate authoritative sources on the Web and use the information to compile relevance rankings.

Google, a search engine originally developed at Stanford University's Computer Science Department, measures link importance based on the concept that Web page authors generally create links only to other pages they think are important. Using link analysis, Google's technology gives a numerical rank to Web pages based on the number of times those pages are linked from an authoritative site, a virtual peer-review process for Web pages. One problem with this concept is that if searchers are trying to find obscure information that is not likely to have been linked from other Web sites, Google probably will not find it.

A new type of search engines are those which use natural language concepts in retrieving results. Oingo searches what it calls the realm of "semantic space," bringing up categories and documents that are close in meaning to the concepts the

---

[17] Some sites with compilations of multiple search engines are:
All-in-One Search Page [http://www.allonesearch.com/], Scout Toolkit: Searching the Internet [http://scout.cs.wisc.edu/toolkit/searching/index.html], and WebCrawler: Database of Web Robots, Overview [http://info.webcrawler.com/mak/projects/robots/active/html/index.html].

searcher is interested in.[18] Oingo provides a sophisticated filtering mechanism that allows successively greater degrees of control over search results by specifying the exact meaning of query words and eliminating irrelevant alternate definitions.

Researchers at Lucent Technologies Bell Labs have invented a technique that would allow a quantum computer to almost instantaneously search massive databases and return very precise results.[19] The technique depends on having a quantum computer, a machine that is largely theoretical, but is slowly starting to become a reality in research labs.

## Search Engine Features

- Most search engines allow for phrase searching, usually by enclosing the phrase in quotation marks, for example, "aurora borealis."

- Most are case-insensitive, so you can enter a keyword in lower case, and the search engine will find both upper and lower case matches. Other search engines allow an exact match, which means you can retrieve words that are capitalized, such as "AIDS," or all lower case, such as "e.e. cummings."

- Most can search for word variations. Some search engines support the asterisk (*) symbol (known as a wildcard) to find word variations. For example, if you enter "sing*," you will retrieve pages on singers, singing, and Sing Sing.

- Most allow for advanced searching. All of the top sites use Boolean search operators to help limit the set if a large number of results is retrieved. The most important of these is "AND." When you use "AND" in a search—for example, "travel AND Antarctica"—the search engine will find Web pages where both those words appear. Another useful Boolean operator is "NOT" (or "AND NOT" in AltaVista). For example, if the search is for "beetle NOT volkswagen," the search engine will find information on the insect and not the automobile. Some search engines allow you to use the Boolean operator "NEAR." For example, "vaccine NEAR HIV." In this case, both words will be in the document and within a few words of each other.

## Search Tips

- Read the help pages of the search engines you use regularly. These explain how to search, what is and is not covered by the database,

---

[18] Sherman, Chris. The Future Revisited: What's New with Web Search. *Online*, May 2000. [http://www.onlineinc.com/onlinemag/OL2000/sherman5.html]

[19] Kahney, Leander. Quantum Leap in Searching. *Wired News*. May 25, 2000. [http://www.wired.com/news/print/0,1294,36574,00.html].

and special syntax or retrieval rules. Take advantage of advanced searching features, such as narrowing the results by document title, date, or domain (i.e., .gov, .edu, .com, etc.)

- To increase the chance of precision searching, try to use unique or uncommon words or acronyms, especially when using a parallel search engine such as Metacrawler or SavvySearch. If there is a synonym or less common word, this will reduce the number of items retrieved. Also remember to vary the spelling to account for differences in British or other spelling (for example, colour or labour.)

- If you want to eliminate commercial sites from your search results, you can add "not" or the minus sign to exclude terms like "order" and "buy" which appear on many commercial sites (i.e., "not order" or "- order.)

- Think of which organizations are interested in the subject and visit those Web sites to see if they provide position papers or link to material on it. For example, if you wanted to find information on handgun control issues, check the Web pages for the National Rifle Association and the Center to Prevent Handgun Violence.

- If you do not find anything useful with one search engine, try another. There is surprisingly little overlap when using the same query in more than one Web search engine.

## Some Common Problems

- *The search engine did not find a Web page you know is available.* No search engine—**none of them**—indexes everything on the Web. If the page is new, it is possible the Web robot has not found it yet. The search phrase or term is checked against an index of documents that the robot has scanned on a previous indexing run. While some robots search the Web continuously, others go out only once a week or once a month.[20] Some dynamic sites,[21] by their very nature, are impossible to index correctly. News sites such as Cable News Network (CNN) or the *New York Times* are updated daily. Hotbot [http://www.hotbot.com] allows you to search for items within the last week, but no search engine can consistently find very recent material, for example, information posted within the previous couple of days.

---

[20] *Search Engine EKGs*, from Search Engine Watch, compares database update times for six major search engines: AltaVista, Lycos, Excite, InfoSeek, Northern Light, and Inktomi: [http://www.searchenginewatch.com/reports/ekgs/index.html].

[21] Dynamic Web sites use programming that allows the developer to create Web pages more animated and responsive to user interaction than previous versions of HTML.

- *The Web robot found the document but was not permitted to access it.* If the page you want is on a server protected by a firewall,[22] access will be denied. Most search engines skip sites that demand a password or registration for entrance, even those, like the *New York Times*, which offer passwords free of charge. Additionally, some Web servers install software specifically to prohibit Web robots from entering. Some search engines cannot index sites with frames,[23] Adobe Acrobat PDF formatted files, CGI output (data provided by users by filling out an online form ), or image maps. Many search engines cannot index Intranets (internal sites which do not link to the Internet) and non-Web resources (i.e., files on gopher, FTP, or telnet servers). Some search tools index only HTML[24] files on Web servers.

- *The Web robot could not access the document, at least for the moment.* This problem is related to the vagaries of Internet traffic and connectivity. The Internet is most congested during the afternoon hours. If you see a message such as "no DNS entry found," this is an indication that the host server is busy or unavailable. Frequently, an immediate attempt to reconnect will be successful.

- Many search engines put a limit on how many Web pages from any individual domain will be indexed, so they do not index free Web hosting services such as GeoCities and its reported 34 million home pages. Web authors who want their sites to be found should register them with individual search engines for inclusion in the search engine's index. Dynamically delivered pages represent another barrier to spiders. The hallmark of a dynamic Web page is a "?" in the URL. Most search engines will not read past the "?," resulting in an error and preventing pages from being indexed.

Information can vanish for other reasons. Webmasters move pages or entire sites without notifying search engines. Pages are deleted when customers' accounts are terminated. The challenge of keeping search engine indexes up-to-date is formidable.

## Usenet News Groups and E-mail Discussion Lists

Usenet is a discussion system distributed worldwide. It consists of a set of "newsgroups" with names that are classified hierarchically by subject. There are approximately 15,000 newsgroups organized according to their specific areas of

---

[22] See Glossary for definition.

[23] Frames is the use of multiple, independently controllable sections on a Web page. A typical use of frames is to have one frame containing a selection menu and another frame that contains the space where the selected (linked to) files will appear.

[24] See Glossary for definition.

concentration. The groups are organized in a tree structure which has seven major categories: *Alt* (anything-goes discussions), *Biz* (discussions of business products and services), *Comp* (of interest to computer professionals and hobbyists), *K12* (education discussions), *Humanities* (literature, fine arts, and other humanities), *Rec* (oriented towards hobbies and recreational activities, *Sci* (research or applications in the general sciences), *Regional* (discussions about a country or U.S. state), *Soc* (discusses issues of different world cultures), *Talk* (debate-oriented, general topics), *News* (concerned with the newsgroups network, maintenance, and software), and *Misc* (groups not easily classified into the other headings, or which incorporate themes from multiple categories). For example, fans of musical composer Stephen Sondheim could read articles posted to the *alt.music.sondheim* or the *rec.arts.theatre.musicals* newsgroups.

"Articles" or "messages" are "posted" to these newsgroups by people on computers with the appropriate software; these articles are then broadcast to other interconnected computer systems via a wide variety of networks. Some newsgroups are "moderated"; in these newsgroups, the articles are first sent to a moderator for approval before appearing in the newsgroup.[25]

Human expertise is very accessible on the Web. A researcher can find information from other people via Usenet newsgroups, listservs, or an e-mail link on a Web page.

Before posting to a Usenet group, read its Frequently Asked Questions (FAQ) guide. Chances are good that your question will be answered there. The FAQ is often compiled by the experts who moderate a particular newsgroup. Two good sources of Usenet FAQs are the *FAQ Archive* at: [http://www.cis.ohio-state.edu/hypertext/faq/usenet/FAQ-List.html] and *The FAQ Finder* at: [http://faqfinder.cs.uchicago.edu:8001/].

Another good practice is to read a few discussion threads before posting a question to a newsgroup. You will get a feeling for the group's style and attitudes and will reduce the chance of getting "flamed"[26] for posting an inappropriate query.

When sending a message to a Usenet group, the question may be sent out globally. People who take the time to answer are likely to feel strongly about the issue or have information that you need. Such direct personal communication is one of the Usenet's strengths. Some of its weaknesses, however, are that some Usenet groups are unmoderated, and that there is no way to verify that a poster is who he/she claims to be, or whether the statements are true or not.

If you see that a particular person frequently posts to a certain Usenet group or seems to be well-informed on a particular subject, you can search for the poster's name in *Deja* [http://www.deja.com/] to see what else he/she has written on that (or any other) topic.

---

[25] For more information on Usenet, see "What is Usenet" at the *FAQ Archive* at: [http://www.cis.ohio-state.edu/hypertext/faq/usenet/usenet/what-is/part1/faq.html].

[26] See Glossary for definition.

An e-mail discussion list server[27] is a computerized mailing list in which a group of people is sent messages pertaining to a particular topic. The messages can be articles, comments, or whatever is appropriate to that topic. There are more than 70,000 electronic mailing lists covering every imaginable topic.

E-mail lists have been used for more than a decade to distribute information efficiently to research and academic communities. Scholarly lists/newsgroups are still more common than scholarly Web sites. To find listservs on various topics, check the *Publicly Accessible Mailing Lists* at: [http://www.neosoft.com/internet/paml/] or *Liszt* at: [http://www.liszt.com].

## Gopher versus Web

The probability of finding something current, valuable, important, and unique on a gopher[28] diminishes as the Web becomes more popular and gophers less so. Gophers are becoming less well-maintained. However, gophers cannot be ignored because a lot of static (but still useful) information is conveyed via gopher. Most search engines also index gophers. A catalog of many of the best gopher sites by category is *Gopher Jewels* at: [http://galaxy.einet.net/GJ/].

## Miscellaneous Sources

Additional information on Internet searching is available at the Library of Congress Home Page. See "Internet Search Tools" at: [http://lcweb.loc.gov/global/search.html].

### Internet News

The sites listed below provide annotated evaluations of new Internet resources within a few days of their availability. A user can also subscribe to them via e-mail. Most of the sites archive their previous issues, so it is not usually necessary to keep copies of postings.

- CNet Digital Dispatch [http://www.cnet.com/]
- Edupage [http://www.educom.edu/]
- Net Happenings [http://scout.cs.wisc.edu/scout/net-hap]
- Netsurfer Digest [http://www.netsurf.com/nsd/]
- Scout Report [http://scout.cs.wisc.edu/index.html]

---

[27] A list server (mailing list server) is a program that handles subscription requests for a mailing list and distributes new messages, newsletters, or other postings from the list's members to the entire list of subscribers as they occur or are scheduled.

[28] See Glossary for definition.

# Glossary of Selected Internet Terms

*Bookmark*—Using a World Wide Web browser, a bookmark is a saved link to a Web site. Like bookmarks for paper books, Web bookmarks are markers that permit you to quickly return to a Web page. Netscape and some other browsers use the term "bookmark," while Microsoft's Internet Explorer uses the term "favorite."

*Firewall*—A dedicated gateway machine with special security precautions on it, used to protect the resources of a private network from outside users. The firewall protects a cluster of more loosely administered machines hidden behind it from individuals attempting to gain unauthorized access.

*Flame*—An electronic mail or Usenet news message intended to insult, provoke, or rebuke; the act of sending such a message.

*FTP*—The file transfer protocol (FTP) command allows an Internet-connected computer to contact another computer, log-on anonymously, retrieve texts, graphics, audio, or computer program files, and transfer desired files back to itself.

*Gopher*—The gopher software program, developed at the University of Minnesota, organizes information into a series of menus. Using gopher is like browsing a table of contents: a user clicks through a set of "nested" menus to zero in on a specific subject.

*HTML*—Hypertext Markup Language is the set of markup symbols or codes inserted in a file intended for display on a World Wide Web browser page. The markup tells the Web browser how to display a Web page's words and images for the user.

*Robot*—A program that automatically explores the World Wide Web by retrieving a document and retrieving some or all the documents that are referenced in it. This is in contrast to Web subject guides that are maintained by humans and do not automatically follow links other than graphic images and redirections (pointers to new URLs).

*Search engine*—A remotely accessible program that lets you do keyword searches for information on the Internet. There are several types of search engines; the search may cover titles of documents, URLs, headers, or full text.

*URL*—Uniform Resource Locator is the unique Internet address which begins with "http://." This address is used to specify a WWW server and home page. For example, the House of Representatives URL is: [http://www.house.gov] and the Senate URL is: [http://www.senate.gov].