UNIVERSITY OF FLORIDA

EXTENSION

Institute of Food and Agricultural Sciences

Editor: Robert H. Stamps

Christine Taylor Waddill, Dean, Cooperative Extension

Cut Foliage Grower

Volume 15, Number 4 **October–December, 2000**

Searching for Information on the Internet

Robert H. Stamps[1]

Please note: In the fast moving world of the Internet, companies, web sites, addresses, categories, features, etc. change and come and go rapidly. Therefore, some of the information contained herein may be obsolete by the time this is published.

The amount of information available on the Internet is immense and growing rapidly. It is estimated that there are 2 billion pages (2) on the WWW (see Table 1 - glossary) alone. Despite, or perhaps because of, this plethora of knowledge, it can be frustrating trying to find information about a specific topic. Of course, the quality of the information found can vary from worthless to extremely useful — but that is a topic for another article.

Luckily there are tools that can make the job of finding the information you need easier. These are directories and search engines. In addition, there are metasearch tools that can connect to several directories and/or search engines at a time, purportedly saving you time and effort and returning a greater number of informative hits than available from a single directory or search engine (more on this later).

Directories (Table 2) are classification systems created by editors. They are a hierarchal set of specific categories into which selected Web sites are placed. Yahoo, which has evolved into a Web portal (see glossary), started out and still contains a directory. Besides proprietary directories, many sites use a Web categorization scheme from the Open Directory Project (ODP). The stated goal of The Open Directory Project is "to produce the most comprehensive directory of the web, by relying on a vast army of volunteer editors". As the time this was written, over 32,000 editors have combed over 2.2 million sites and created over 325,000 categories of information. The ODP started out as an open-source competitor to Yahoo and is now owned by Netscape/America Online (3). The ODP Web categorization is being used by many web sites including All the Web, AltaVista, Deja, Go, Google, HotBot, Lycos and Northern Light. If you have time on your hands or an interest in a particular area not already covered by an editor, you can sign up to be a volunteer. An example hierarchy from the ODP is given below:

- Home
  - Gardens
    - Plants
      - Perennials
        - Australian Acacias

[1]Professor of Environmental Horticulture and Extension Cut Foliage Specialist, University of Florida, Institute of Food and Agricultural Sciences, Mid-Florida Research and Education Center, 2725 Binion Road, Apopka, FL 32703-8504. Phone - 407/884-2034, Fax - 407/814-6186, E-mail - rhs@gnv.ifas.ufl.edu.

Table 1. Glossary of terms/acronyms associated with the Internet.		
Term	**Meaning**	**Comments**
algorithm	A set of programming instructions (code) written to solve a problem or do a particular task.	Programers may develop differing ways of doing similar tasks. The results of the tasks (e.g., web searches) may, therefore, vary.
Boolean operators	Commands that connect (AND), exclude (AND NOT, NOT), associate (NEAR, OR), categorize (parentheses)	Boolean operators can be used to develop very specific and powerful search criteria.
crawler	Search software that follows hypertext links from server to server and indexes information according to search criteria (algorithm).	Indexing speeds and quality of the information collected vary from crawler to crawler. Results may change overtime.
hit	Web site that the search engine selects as appropriate based on your search criteria and the cataloging and searching algorithms used by the programs.	Some "hits" may be partially or totally irrelevant. Selection of appropriate search terms and criteria and use of sophisticated search engines can result in fewer misses.
extinct	Term used to describe URLs (see below) that are links to web pages that no longer exist. The information in these obsolete links is sometimes saved by search engines as cached pages (see right).	As time goes on, more and more of these outdated links develop. Search engines should either remove these listings or, at the least, cache the information originally present, so as to save searchers time and effort. With the latter option, users should realize that the cached information may be dated or incorrect.
metasearch tools	Software that takes a query, formats it appropriately for multiple directories and/or search engines, sends the queries out, and reports back the results from each site that was contacted.	Provides a quick way to search many directories and search engines at one time; however, better results can often be provided by making queries directly at each site.
portal	Informational hubs with a collection of links that connect visitors to information sources, such as on-line yellow pages, auction and shopping sites, news services, etc. Portals consist of topical tree listings (directories) of sites combined with at least one search engine.	Portals are typically positioned as starting points for users logging onto the WWW. Examples include AOL (America On Line), Excite, Yahoo.
search engine	Software for finding information on the Internet. Program that take search criteria and use them to find web pages deemed relevant.	Some can only accept simple search criteria whereas other search engines can accept natural language or complex criteria.
spider	Same as crawler	Same as crawler
surfer(s)	Someone that explores the WWW by jumping from website to website.	
usenet newsgroups	Discussion groups about a topic that is reflected in their titles, such as corel.wpoffice.wordperfect9 or fl.gardening.	Group titles can be obscure and explicit language is used in some newsgroups. Comments about a particular topic can be grouped together in "threads" listing questions and subsequent responses.
URL	**U**niversal **R**esource **L**ocator	Address for a web page. Example: mrec.ifas.ufl.edu/cutfol/cutpage.htm

WWW	World Wide Web	This is a subset of the Internet that has a graphical and acoustical user interface.

Table 2. Some World Wide Web directories.

Name	Address (URL, see glossary)	Comments
About	www.about.com	A search for "acacias" under Home/Garden resulted in 10 directory "hits" of which only a few were useful. Numerous web site "hits" were simultaneously found by the WWW search engine at this site.
Looksmart	www.looksmart.com	A search for "acacias" under Lifestyle, Gardening resulted in 12 useful "hits". An option is available to have a web search done on "acacia" with a single mouse click (an Inktomi Corporation search engine is used).
Open Directory	www.opendirectory.org	A search for "acacias under Home/Gardens turned up one "hit". Searching the whole directory yielded 11 "hits", with two being useful.
Yahoo	www.yahoo.com	A search for "acacias" under Recreation/Home and Garden was fruitless, and a search of the whole Yahoo site yielded only one useful hit out of 59 reported back by Yahoo.

Search engines use programs called crawlers and spiders to search the WWW where they index all the text they encounter into searchable forms. The search engines use different methods of perusing and categorizing that information. Also, each search engine has categorized a different volume of information and web sites. In fact, search engines have categorized only a fraction of the WWW, and each has crawled differing portions of the total content. In addition, this process is ongoing so there is always a lag period between the time the information is entered on the web and the time when it is discovered by the search engines. On top of that, websites come and go so many links become "extinct" (obsolete) — search engines should clean up these defunct connections; however, many remain.

Despite the hype, search engines that base their hit rankings on where previous web users have gone do not do well when searching for uncommon topics such as those concerning potential or current cut foliage crops such as our search engine test phrase — "leatherleaf fern".

Additionally, the format in which the information is placed on the WWW influences if, and how, it becomes available to search engines. For example, at this time, most search engines cannot recognize and index documents in Adobe Acrobat portable document format (*.pdf). This means that they cannot, for example, index the text of the *Cut Foliage Grower* and *Cut Foliage Research Notes* that are located on our website (mrec.ifas.ufl.edu). The Acrobat format is widely acknowledged as a standard for publishing formatted information to the web so that it will print out as the author desired, regardless of the output device. Fortunately for those searching the web, Adobe is providing a service (searchpdf.abobe.com) that finds these publications (including the cut foliage ones mentioned above). Adobe has found and summarized over a million *.pdf files located on the WWW. Adobe summarizes them in a format that search engines can detect and creates links to the original *.pdf files. Besides information in Acrobat format, there is much other information on the Internet that cannot be accessed by search engines. Examples include library catalogs, certain reference works and many government documents.

Search engines also differ in the ways that they search for information when you request it. The methods (algorithms) used vary in complexity, configuration and effectiveness. This and the differences in the sites categorized make for unique results for different sites. Because of the variation, metasearch tools that send queries to numerous search engines have the *potential* to find the greatest number of informative sites (see Metasearch section below and comments in Table 3).

Only selected search engine sites are listed in Table 3, ones that are especially superficial and geared mainly for selling products are excluded. Also, sites that depend on gimmicks to attract users and that have very poorly performing search engines are not listed. Multiple sites 3 may use the same search engine, but results may be different from one site compared with another using the same search engine. Of the search engines listed in Table 3, only a few (AltaVista/Raging Bull, Google, Northern Lights, Yahoo) seem suited for use doing specific, obscure inquiries.

Metasearch tools format queries in forms appropriate for each of the directories and search engines that they contact. If your query is too sophisticated for a particular site it may not forward a request to that site. Despite the hype, many searches using these tools provided far fewer hits than one obtains from directly querying search engines directly. In addition, none of the metasearch tools allowed for specifying detailed search criteria.

Formatting a search request

How you request information from a search engine (or metasearch tool) can be as important as the site from which you make your search request. The following are a few suggestions that should make your searches more fruitful.

1. ***Search for phrases*** – One of the easiest ways to increase the number of relevant hits is by searching for phrases. This usually results in fewer but far more pertinent web pages when using multiple words. For example at Excite, the phrase "leatherleaf fern" (in quotes) returned 163 hits while the two words not in quotes returned 19,505 — all sites that contained the words leatherleaf or fern, as well as the combination. The phrase search, of course, provided a far higher percentage of relevant hits. Most, but not all, search engines accept search phrases (see Table 2). Phrases are selected using Quotation marks (" ") and/or by selecting phrase from drop down menus.
2. ***Use other available search criteria*** – Check to see if search engine has any advanced search options such as these listed below.
   a. ***Word groupings –*** As indicated above, some web sites have drop down menus from which to select search attributes. For example, All the Web lists *any of the words*, *all the words* and *exact phrase*. The number of hits for the two words "leatherleaf" and "fern" returned almost 450,000 "hits", 878 "hits" and 319 "hits", respectively, for those three search criteria. Some search engines allow you to specify words that occur NEAR one another in a web page.
   b. ***Capitalization –*** Some search engines are sophisticated enough to be able to handle capitalization and some can even deal with mid-word capitalization (eBay, pH).
   c. ***Word inclusion/exclusion –*** Some web sites allow searchers to include and exclude specific words from searches. For example, searching for "leatherleaf fern" but excluding sites that include the word roses (AND NOT roses) resulted in 621 fewer hits than a search that did not exclude roses on AltaVista. On some sites, + and – signs are used before words to include and exclude them, respectively. Other sites have boxes where words to be included and excluded are entered. Still others use Boolean operators like AND NOT or NOT to exclude words.
   d. **Wildcards/truncation symbols –** Some search engines are capable of handling wildcards. For example, if you want to find sites listing "leather fern" or "leather ferns" or "leatherleaf fern" or "leatherleaf ferns" you could type in "leather* fern*". However, be

aware that the search might also find links to leatherWOOD and other ferns. The use of wildcards can help find plurals, as just illustrated, and different spellings for the same word (e.g., colo*r = color or colour).

e. ***Boolean operators –*** The more powerful search engines allow the use of Boolean operators (AND, OR, NEAR, NOT, parentheses) when defining search criteria. Check to see which operators a particular search engines supports. Boolean operators allow you to make very specific search requests. For example, "leatherleaf fern" AND "leather fern" NOT (roses OR carnations OR chrysanthemums). NEAR is used when you want to find words in fairly close proximity, say within 10 words, of each other.

f. ***Domain filtering –*** To further refine searches and eliminate many non-pertinent sites or focus only on specific types of sites, use domain filtering where available. For example, excluding commercial sites (*.com) from the leatherleaf fern phrase search using Google dropped the number of hits from 474 to 207. Alternatively, searching for only educational sites (*.edu) reduced the number of hits to 103. Domain filtering is usually an advanced or power search option where you choose include or exclude and then type in the domain extensions (.xxx).

g. ***Language filtering –*** Some sites allow searchers to limit the selection of hits to specific languages.

h. ***Date filtering –*** A useful feature, especially if you do repeated searches over time using the same criteria, is date filtering. This allows searches to be limited to date ranges in which the searcher is interested.

i. ***Display of URLs –*** It is very useful to have web site URLs displayed in the results. This can give an easy indication of on what type of site the information is located. This is a type of domain filtering. However, not all search engines offer this convenient feature.

Some of the more sophisticated search engines will even allow the searcher to specify where in the web page (which fields) specific information is to be looked for. For example, you might want to search for sites that have ifas.ufl.edu in their URL or "cut foliage" in the title.

Additional search refinements include such things as limiting the number of web pages identified per web site and offering language translation services. Reporting options may include web page sizes, the last date the web page was modified and limiting results to one page per site.

There are differences between the directories and search engines but, in general, they are all getting better and more powerful. Not all search engines are equally effective and even the same search engine used at different portals can return differing results (Table 3). These tools, when used efficiently can be a good way to find information on the Internet. Once you find information, it is up to you to determine it quality and accuracy. Happy surfing.

References

1. Charski, M. 2000. Google's ad program stresses simplicity. Inter@ctive Week 7(33):14.
2. Sirapyan, N., C. Metz, S. Pike and C. Gero. 2000. In search of … PC Magazine 19(21):186–190, 193, 195–196, 198.
3. Willmott. 2000. 10 insightful search engines. Computer Shopper 20(11):174–175, 180–182.

Table 3. Some search engines for finding information on the WWW.

Search Engine Name (number of "hits" for search phrase "leatherleaf fern"	**Address (URL)**	**Phrase search**	**Word filters**	**Boolean operators**	**Domain filtering**	**Language filtering**	**Limit search by date**	**Limit pages per site**	**Display URL in results**	**Translations available**	**Comments**
About (40)	www.about.com	X							X		Best suited for general queries, not specific and/or complex searches. Has its own human organized directory but also includes a WWW search engine (powered by Inktomi) and links to other search engines, metasearch tools and directories.
All the web (319)	www.alltheweb.com	X	X		X	X			X		Reportedly has 575 million URLs in its database and is increasing that number rapidly. Provided a large number of hits for specific searches.
Alta Vista (712)	www.av.com (www.altavista.com)	X		X	X	X	X	X	X	X	Fast and powerful, perhaps the most used search engine on the WWW (1). Does well on specific searches but may produce a few more extinct web links because it started crawling the web earlier than some of the newer search engines.
Ask Jeeves (60)	www.ask.com										Results of the natural language queries not very relevant; sites are rated by popularity which is of little use for cut foliage queries. Some "hits" did not contain search words at all and some were no longer in service (extinct). A popular site for consumers.
Deja News (2)	www.deja.com	X			X[Z]	X	X		X[Y]		Different from most other search engines because it searches Usenet newgroups postings, not web pages. Can be useful for finding specific information but only if it has been the topic of a newsgroup discussion.

Search Engine Name (number of "hits" for search phrase "leatherleaf fern"	**Address (URL)**	**Phrase search**	**Word filters**	**Boolean operators**	**Domain filtering**	**Language filtering**	**Limit search by date**	**Limit pages per site**	**Display URL in results**	**Translations available**	**Comments**
DirectHit (144)	www.directhit.com		X		X				X		"Hits" ranked by confidence that they are relevant to the query based on earlier surfers. For specific and obscure searches, this method is of little or no value. Search results for our specific search disappointing. A subsidiary of Ask Jeeves, Inc.
Excite (163)	www.excite.com	X	X		X	X			X		Web database reportedly has about 250 million URLs and works well for specific queries. Another part of the Excite Network, Blue Mountain electronic greeting cards (www.bluemountain.com), is an entertaining e-card site.
FindWhat (40)	www.findwhat.com	X	X	X[X]							Not powerful and results very limited. Uses proprietary Pay for Position™ technology, basically selling top search rankings to the highest bidders. This site powered by Inktomi Corporation's search engine.
Go (104)	www.go.com (www.infoseek.com)	X	X						X		GO.com is part of the Walt Disney Internet Group. Web search powered by Infoseek. Mediocre results with many relevant pages missing and some irrelevant ones included.
Google (474)	www.google.com	X	X		X	X			X		Claims database includes over 1.3 billion web pages. Produced a large number of relevant hits.
GoTo (27)	www.goto.com	X	X								Another site powered by Inktomi that produced few relevant hits, some not at all relevant, and others extinct. Advertisers pay for their placement in the search results at this very commercially oriented site.

Search Engine Name (number of "hits" for search phrase "leatherleaf fern"	**Address (URL)**	**Phrase search**	**Word filters**	**Boolean operators**	**Domain filtering**	**Language filtering**	**Limit search by date**	**Limit pages per site**	**Display URL in results**	**Translations available**	**Comments**
HotBot (154)	www.hotbot.com	X	X	X	X	X	X	X			Provides many ways to specify what to search for and where to search. Returned a moderate number of relevant hits but many significant ones were missing. Part of the Terra Lycos Network.
Lycos (316) FAST	www.lycos.com	X	X		X	X			X		Although Lycos provides a number of search options, they cannot be used together. For example, a search by domain cannot be combined with one for language or host. Lycos is a strategic investor in FAST and, therefore uses FAST as its search engine.
Magellan (163)	magellan.mckinley.com	X	X	X					X		Magellan reportedly has over 50 million Web sites in its database, can search for ideas and concepts, and sorts hits by relevance. Although the database is limited, there were a good number of relevant hits for the search phrase — including PDF files (through Search Adobe PDF Online). Magellan is a subsidiary of Excite, Inc.
Northern Light (452)	www.northernlight.com	X	X	X	X	X	X		X		Many relevant hits, gives summaries of some published articles and will provide complete articles for a price. Its Special Collection is a unique combination of data representing over 6,700 journals, books, magazines, databases and newswires not easily found on the WWW. Most of this information is unavailable on the Internet. Also, organizes results into custom research folders.

Search Engine Name (number of "hits" for search phrase "leatherleaf fern"	**Address (URL)**	**Phrase search**	**Word filters**	**Boolean operators**	**Domain filtering**	**Language filtering**	**Limit search by date**	**Limit pages per site**	**Display URL in results**	**Translations available**	**Comments**
Oingo (10*)	www.oingo.com					x[W]					Although Oingo indexes information based on meaning rather than text, initially found few relevant hits. Oingo's Web Hits Filter allows you to control whether or not you see the AltaVista web hits, you can them transfer to AltaVista. To save time, try AltaVista first.
Raging Search (288)	www.raging.com	X		X	X	X	X		X	X	AltaVista search engine stripped of the surrounding portal trappings. Extensive search criteria and results customization available. Interestingly, results from phrase search different than AltaVista.
WebCrawler (163)	www.webcrawler.com	X	X	X					X		Supports "natural language searching. Returned a decent number of relevant "hits" but many were duplicates or triplicates. Users are directed to try Excite for more comprehensive results.
Yahoo (316)	www.yahoo.com	X[V]					X[U]				Popular but somewhat superficial online directory, best for finding popular WWW sites. Search engine powered by Google. Can connect to Google usenet search engine from here

[Z] Forum filtering since this tool searches Internet newgroups rather than WWW pages.
[Y] Displays newsgroup name.
[X] Limited to AND, OR and parentheses ().
[W] Limited to English, Spanish or both.
[V] Applies only to Yahoo Directory Searches.
[U] Only for specific periods from the present back, i.e., one day, three months, four years.

Table 4. Some metasearch engines available on the World Wide Web.			
Name	**Address (URL, see glossary)**	**Query formats**	**Comments**
C4	www.cyber411.com	Phrase using double quotation marks (“”), Boolean (AND, OR, NOT)	Searches AltaVista, Excite, HotBot, InfoSeek, Lycos, Magellan, NBCi, WebCrawler, and Yahoo. Results are sorted by relevance.
Dogpile	www.dogpile.com	Little control other than phrase search (“”)	Searches three of 12 search engines at a time (AltaVista, Direct Hit, Dogpile Web Catalog, FindWhat, Google, GoTo.com, InfoSeek, Kanoodle, LookSmart, Lycos, RealNames and Yahoo) plus the Open Directory. Results variable; for example, 172 hits reported from AltaVista but none from Yahoo. However, the same search phrase made directly on Yahoo (which uses Google’s search engine) resulted in hundreds of hits. Curiously, all Dogpile attempts to query Google timed out.
GoGettem (formerly SuperSeek)	www.gogettem.com	phrase search	User selectable queries of up to eighteen search engines at a time, from AltaVista to Yahoo. A nice feature is that a browser window is opened for each search engine selected putting the user in direct contact with each search engine. The number of “hits” reported is therefore the same as would be found doing the search directly at each search engine site.
Ixquick	www.ixquick.com	+, -, “”, (), CapS, *	Sends queries to AOL, AltaVista, EuroSeek, Excite, FastSearch, GoTo, HotBot, InfoSeek, LookSmart, MSN, NBCi, WebCrawler, Yahoo, xRefer. Unfortunately for those doing an obscure search, say for leatherleaf fern, reports ~300 hits but only links to a few “top ten” pages.
Mamma	www.mamma.com	Limited Boolean operators option (PHRASE or AND or OR) available under power search.	Searches Askjeeves, FindWhat, Go To, InfoSeek, Lycos, MSN, NBCi, and Yahoo.
Metacrawler	www.metacrawler.com	Any, all or phrase searchs.	AltaVista, DirectHit, Infoseek, Lycos. However, Lycos timed out on. Time out period can be increased. Shows URLs.
Search (SavvySearch)	www.search.com	+, -, “”, Boolean (AND, OR, NOT)	Default queries are to AltaVista, DirectHit, FindWhat, GoTo, Lycos and NBCi. Can query a total of 19 search engines from About to Yahoo. Can click on links to search engines that return results.