Tips and Tricks

for FINDING

THE NEEDLE IN THE INTERNET HAY STACK

March 2003

by Abigail Schearer

Tips and Tricks for
FINDING THE NEEDLE IN THE INTERNET HAY STACK

http://www.keystonemac.com

***TECH/Cartoon**
Hand Out/Cartoon
Outline/Creating a SE

This program was designed to answer three questions re: searching the Internet.

1. What tools are the easiest to use with the best quality results?
2. Why do I get such weird, irrelevant responses to my questions?
3. How can I word my question to get better answers?

STARTING OUT A good way to begin your search is to create a search strategy in your head by asking yourself this question: What do I want to do? Your answer determines how you will conduct your search and which tools you will use: a Subject Directory, a Search Engine or a Meta Search Engine

1. "Do I want to just Browse, or
2. locate a specific piece of information
3. or retrieve everything I can on the subject?"

OVERVIEW

A. WEB DIRECTORY or SUBJECT DIRECTORIES: If you're **browsing and trying to determine what's available in your subject area,** start out by selecting a subject directory like Yahoo! This will give you a broad base for your search with more general information.

Unlike search engines, Subject Directories are created and maintained by real live people. They search the web pages, evaluate each one for subject-appropriate content and then add the best to the database. Therefore, Directories probably won't give you any where near as many references as a search engine will, but they are more likely to be more relevant.

B. SEARCH ENGINES are a better choice if you're **looking for a specific piece of information.** Go to a major search engine such as Google or Ask Jeeves, or to a specialized database such as PAJobs (with links to online job banks) or lowermybills (for comparing long distance phone rates)

A search engine has two parts for creating a database, a spider and an indexer. The spider is the program that fetches the documents, and the indexer reads the documents and creates an index or database based on the words or ideas contained in each document.

C. META SEARCH ENGINES are both Directories and Search Engines. A Meta Search Engine feeds a single query to several SE's and then combines and organizes all results in a single list. Enter your search keyword(s) into one of the mega-search engines, such as Ixquick, or ez2www just to see what's out there.

I. INTRO: THREE WAYS INTO THE INTERNET:

A. SEARCHING BY MEANS OF WEB or SUBJECT DIRECTORIES:

1. Subject Index: Remember, in the old library card catalogue you would find information by searching on either the author, the title, or the subject.
When you want to cover a broad range of information you would choose the *subject* option.

2. Subject-tree: Generally, when you open a Subject Directory, you read the list of major topics and decide which one best fits your search. Then under each of these *topics* is a list of *subtopics*. Each topic you select move you from the general to the more specific. Example: a query searching for Search Engines...
(Internet Demo: > looksmart.com > computing > computers & Internet > Internet & Web > Internet Resources for beginners)

3. Good Subject Directories to check out:

a. Yahoo.com is a actually a directory--a subject index. It attempts to catalogue and organize the entire Web search into a subject or topic. If you know exactly what subject you're searching for, and you can choose your subject from within a hierarchy of larger subjects, Yahoo is a good place to start.

b. StartBot.com: There is no bigger Search Engine's search engine directory. Large collection of search engines, meta search engines, worldwide search engines, yellowpages, and more.

c. Cosmix.com is a searchable directory of web directories and search engines. It has 3 search modes: The Insane Search (the most thorough search), The Web Search (a simplified meta search), and The Mother Load Search (a local directory search).

B. SEARCHING BY MEANS OF META SEARCH ENGINES

Good resource go to...**www.startbot.com/multi-meta-searchengines.html**

1. Dogpile.com: Dogpile meta search retrieves the greatest number of relevant results that match your query. Uses search keywords or

phrases. Can include Boolean operators, and " " for phrases, but these operators will be removed when search is sent where they are not accepted. The results can be inconsistent especially from the smaller sources. Useful if you customize Dogpile to have it search Google, AltaVista, and Infoseek first. Does not combine results.

>> Dogpile Searches AltaVista, Direct Hit, Dogpile open directory, Dogpile web catalog, FindWhat, Google, GoTo.com, InfoSeek, Kanoodle, LookSmart, Lycos, RealNames, Yahoo, but changes frequently.

2. MetaCrawler is designed to be simple and easy to use. MetaCrawler queries many of the Web's top search engines simultaneously, retrieving the best search results across the Internet and organizes them in a single list, ranking them by relevance.

3. Excite.com indexes some 50 million web pages. You can search them using the standard keywords, or use plain English phrases!

4. "Simple Art of Searching"
www.cs.buffalo.edu/faculty/miller/Courses/WWW-Sem/Sp97/Presentations/SearchEngines-Klemic/bigtable.html

5. Selected Meta-Indexes
http://reinert.creighton.edu/101/SrchEngext.htm

C. SEARCHING BY MEANS OF SEARCH ENGINES
path/berkley.edu/SE
www.lib.berkeley.edu/TeachingLib/Guides/Internet/ToolsTables.html

1. "How do Search Engines Work":
>>Links to detailed help,
>>size/type,
>>Phrase searching,
>>Boolean logic

2. Current Search Ratings
>> Biggest, Fastest: FAST **alltheweb**
Runner-up: **Google**
>> Coolest, Easiest, Most Fun: **Ask Jeeves**
>> Most Comprehensive Results: **Google**
>> Highest Overall Usability Rating: **Google**
Runner-up: **Yahoo**
>> Best Search Engine For Kids: **Ask Jeeves For Kids**
>> Most Relevant Results: **Google**
Runner-up: **AltaVista**
>> Most Likely to Find a Hit When Others Can't: **Northern Light**
Runner-up: **AltaVista** It's big, fast, and very popular

among dedicated web-surfers with lots of search-refining options.

3. "Things to Know About"
http://www.infotoday.com/searcher/oct01/price.html

Q & A

II. SEARCH ENGINES: Search Engines use two primary methods of text searching…keyword searching and concept-based searching**.**

A. KEY WORD-BASED SEARCH searches for word variations.
http://infopeople.org/search/chart.html

1. Databases: Search engines use software robots to survey the Web and build their databases. Web documents are retrieved and *indexed.* When you enter a query at a search engine web site, your input is checked against the search engine's keyword index. The best matches are then returned as hits.

2. Example of Indexing:: When the phrase is "within quotes" AltaVista prepares its index by searching every word on every page. A phrase without quotes will ignore the *stop words.* Example: "searching the web" contains two *stop words:* the & web. Consequently the search engine will only look for "searching". Aware of this, you can narrow your search with more relevant keywords, like "people search".

3. Stop words: Search engines ignore several hundred of the most common words in an effort to speed things up. They vary from engine to engine, but always contain words like the *a, an, the, and, or, to, in* & *is.* It doesn't matter whether they are embedded in a phrase or if they have a + before them, they will not be included in the search. Note: Google will let you know of any words it has excluded.

4. Confidence or Relevancy Rankings is how most of the search engines return results**.** SE's list hits according to how closely they think the results match the query. As far as the user is concerned, relevancy ranking is critical, and becomes more so as the sheer volume of information as the Web grows.

5. Term-frequency and the **positioning** of **keywords** is what most search engines use as a primary way of determining whether a document is relevant, reasoning that if the keywords appear early in the document, or in the headers, this increases the likelihood that the document is relevant.

6. Advanced query button on **AltaVista,** allows you to assign

relevance weights to your query terms before conducting a search. Although this takes some practice, it essentially allows you to have a stronger say in what results you will get back. *(dialogue box)*

7. Stemming is the ability for a search engine to search for variations of a word based on its *stem*. For example, when you enter the word *test*, the SE would search also for *tests, tested, tester, testers, testing,* and so on.

NOTE: **Google** does not support word stemming and does not provide a way to turn it on either. When using Google, if you want to search for variations of a word, you'll need to add those variations to your search query.

8. Keywords that mean the same: Search engines CANNOT return hits on Keywords that mean the same, but are not actually entered in your query. For example, query on heart disease would not return a document that used the word "cardiac" instead of "heart."

9. Keywords that are spelled the same way but mean something different often results in hits that are completely irrelevant to your query Example: *hard* cider, a *hard* stone, a *hard* exam, and the *hard* drive on your computer.

B. CONCEPT-BASED SEARCH.

1. Definition: Unlike keyword search systems, concept-based search try to determine what you mean, not just what you say**.** In the best circumstances, this system returns hits on documents that are "about" the subject/theme you're exploring, even if the words in the document don't precisely match the words in your query.

2. Clustering essentially means that words are examined in relation to other words found nearby.

>> *Excite* is currently the best-known general-purpose search engine site on the Web that relies on concept-based searching. It's software determines meaning by calculating the frequency with which certain important words appear. When several words or phrases, that are tagged to signal a particular concept, appear close to each other in a text, the search engine concludes that the piece is "about" a certain subject.

>> Example, the word heart, when used in the medical/health context, would be likely to appear with such words as coronary, artery, lung, stroke, cholesterol, pump, blood, attack, and arteriosclerosis. If the word heart appears in a document with others words such as flowers, candy, love, passion, and valentine, a very different context is

established, and the search engine returns hits on the subject of romance.

>> Problem: Concept-based indexing is a good idea, but it's far from perfect. The results are best when you enter a lot of words, all of which roughly refer to the concept about which you're seeking information.

III. SEARCH STRATEGY: Keys for constructing successful searches. (http://sc.edu/beaufort/library/lesson6.html)

www.noodletools.com/debbie/literacies/information/5locate/adviceengine.html

A. TIPS YOU CAN USE AT JUST ABOUT EVERY SEARCH SITE NOTE, There are a number of special characters that cause some search engines to consider variations of your keyword search terms, such as described in number 1, 2, & 3.

1. Spelling counts: If you spell a term incorrectly some sites might recommend a different spelling, based on the closest terms it can find. At the top of the "hit" parade it would say, "Do you mean....?" However, the spell-checker can't catch everything and it doesn't work at all in Advanced Search.

2. Wildcards or Truncation character, the asterisk (*) is used by many search sites to stand for parts of a word that you are not quite sure about. This is very useful if you aren't sure of the correct spelling of a term. In most instances, wildcards can match any character, or group of characters, from its particular position in the word to the end of that word. Enter the first four to six letters followed by " * ".

>> Example: renais* will find renaissance.

>> str*ed will find stretched, straightened and strapped;

>> sher* will find Sherry or Sheryl or Sherylyn;

>> auto* would return all three related words...auto, automobile, automotive.

3. Phrases in Quotation Marks (" ") forces the search engine to find a phrase or a specific grouping of words in a precise order. For example, most search engines can recognize the phrase "to be or not to be" as a famous quotation from Hamlet when bracketed by " "marks.

4. Nest them with Parentheses ( ) For example...Marx NOT (Brothers OR Moscow); ("Jesus Christ" NOT Humor) AND (Mary OR Magdalene).

>> BUT DON'T GO TOO FAR!! ((alphabet AND Soup) NOT (twinkies OR "KFC")) AND nutritious... is too confusing. Use "alphabet soup" AND nutritious... and if you get a lot of KFC hits, refine the results to exclude

them.
(NOTE: most top engines support this.)

5. Nouns as key words: Verbs and conjunctions are either ignored by the search engines or are too common to be useful. When searching a noun remember that most nouns are subsets of other nouns. Enter the smallest possible subset that describes what you want. Be specific. (Pets>dogs>collie)

6. Descriptive or specific words are used to narrow your search. Using multiple words helps the search engine get a handle on the concept you're looking for...the more descriptive or specific words the better. Buying a car? Don't just enter the keyword "car", enter the keyword "Toyota." Better still, enter the phrase "Toyota Dealerships" AND the name of the city where you live.

7. Reverse questions: Search engines look for pieces of text that match your query. Web pages are more likely to contain answers than questions Phrase your query how you would expect the answer to read - the difference appears slight, but it makes a huge difference. "IRS stands for" rather than "What does IRS stand for?"; "man first landed on the moon in" rather than "When did man first land on the moon?"; "sky is blue because" instead of "Why is the sky blue?" * Note: Ask Jeeves is the exception, and comes up with excellent answers to common, natural language questions.

8. Use rare words The more unusual or uncommon are the keywords you use, the more specific the results will be. Taking a moment to think of a valid yet uncommon word is a valuable technique. The word alcohol returns 912,620 hits (AltaVista) vodka fetched 120,740 but then it narrows down to 2754 hits when you enter Stolichnaya. NOTE: For a few engines the word order is important, so always enter the rare word first.

9. Capitalization: Some of the search engines discriminate between upper case and lower case; others store all words without reference to capitalization. This is essential information for searching on proper names of people, companies or products. Unfortunately, many words in English are used both as proper and common nouns--Bill, bill, Gates, gates, Oracle, oracle, Lotus, lotus, Digital, digital--the list is endless.

10. Help Files on the Search Engine Sites. All the search engines have different methods of refining queries. The best way to learn them is to read the help files on the search engine sites and practice!

B. BOOLEAN OPERATORS:
http://sc.edu/beaufort/library/lesson6.html
Many search engines allow you to use Boolean operators to refine your

search. These are the logical terms AND, OR, NOT, and the so-called proximal locators, NEAR and FOLLOWED BY. Capitalize all Boolean terms or they will be ignored.

1. AND or "**+**" requires the word to be present, like "heart" AND "attack." You might use this if you want to exclude common hits irrelevant to your query.

2. NOT or "**-**" excludes words, particularly important as the Web grows.

Example: If you want to find medical details about Alzheimer's Disease, try entering "Alzheimer's" AND "symptoms" AND "prognosis." If you want to find out about Alzheimer's care and community resources, query on "Alzheimer's" AND "support groups" AND "resources" AND NOT "symptoms."

3. OR allows either word to be present, i.e., bronchitis, acute OR chronic.

4. FOLLOWED BY means that one term must directly follow the other.

5. ADJ, (for adjacent *a-jay-sent)*, means that you to search on phrases by determining adjacency of keywords. This, essentially, serves the same function as FOLLOWED BY.

6. NEAR means that the terms you enter should be within a certain number of words of each other.

Example: The word heart, when used in the medical/health context, would be likely to appear with such words as coronary, artery, lung, stroke, cholesterol, pump, blood, attack and arteriosclerosis. If the word heart appears in a document with other words such as flowers, candy, love, passion and valentine, a very different context is established, and the search engine returns hits on the subject of romance. Results are best when you enter a lot of words.

Advanced Boolean Check pg. 3 & 4:
http://sc.edu/beaufort/library/lesson6.html

IV. LOOKING FOR ADDITIONAL SEARCH INFORMATION: Go to..
WEB SITES/AS recommended

Researched and Compiled by Abigail Schearer 1/03