WEB SEARCHING BASICS: Q & A

HOW DO SEARCH ENGINES WORK?

Search engines use automated software programs known as spiders or robots to survey the Web and build their databases. Web documents are retrieved by these programs and analyzed. Data collected from each web page are then added to the search engine index. When you enter a query at a search engine site, your input is checked against the search engine's index of all the web pages it has analyzed. The best URLs are then returned to you as hits, ranked in order with the best results at the top.

WHAT ARE KEYWORDS AND HOW ARE THEY USED?

A keyword search is the most common form of text search on the Web. Most search engines do their text query and retrieval using keywords.

What is a keyword, exactly? It can simply be any word on a webpage. Unless the author of the Web document specifies the keywords for her document (this is possible by using **meta tags**), it's up to the search engine to determine them. Essentially, this means that search engines pull out and index words that appear to be significant. Since engines are software programs, not rational human beings, they work according to rules established by their creators for what words are *usually* important in a broad range of documents. The title of a page, for example, usually gives useful information about the subject of the page. Words that are mentioned towards the beginning of a document are given more weight by most search engines. The same goes for words that are repeated several times throughout the document. Some search engines index every word on every page. Others index only part of the document.

WHAT IS CONTROLLED VOCABULARY AND HOW IS IT USED?

Controlled vocabulary is standard words or phrases, often in a hierarchical thesaurus format, that can help in the searching process. Often controlled vocabulary is subject terms assigned by a librarian or other information professional in a formal process. Controlled vocabulary is useful to find documents that are about similar topics, but which use different keywords in the full text (because of different authors or disciplines). By searching for the standard vocabulary, the searcher can find all the documents, even if different keywords are used in the actual document text. Oftentimes, searchers can determine standard or controlled vocabulary by looking for a thesaurus in the database help files, or by looking at the subject terms assigned in the bibliographic record for a document, and then revising a search to find those words or phrases.

WHAT IS A META TAG?

Some search engines are now indexing Web documents by the meta tags in the documents' HTML (at the beginning of the document in the "head" tag). Meta Tags do not affect the appearance of a web site and are not visible when you look at a Web page, but they provide information regarding the content of the site. What this means is that the Web page author can have some influence over which keywords are used to index the document, and even in the description of the document that appears when it comes up as a search engine hit.

Adapted from: The Spider's Apprentice – "How to Use Web Search Engines"
Retrieved July 21, 2005 from: http://www.monash.com/spidap2.html