**Search Engines and Directories** 

Presented by: Chris Moore (chrism@ccohs.ca) Canadian Centre for Occupational Health and Safety (CCOHS)

April 2001

## The Challenges

"Picture a vast, echoing library, with row upon row of books, containing all the world's knowledge. Citizens wander in from time to time and add a book wherever they please. The volumes are in no particular order; there's no card catalog and no Dewey decimal system. There are, however, a number of independent librarians who scurry about perusing and indexing the books. Each librarian's index is fiercely separate from the rest, and each index references only a fraction of the library's contents. A bad dream? No, it's a description of the Internet as we know it today. " ("How to search the Internet...better" by Neil J. Rubenking, PC Magazine, 2000)

Two of the main challenges that people face when searching for information on the web are: knowing where to start; and knowing how to avoid being overwhelmed with junk! In this session, I hope to give you a bit of help with both of these challenges.

## The Tools

There are two main types of WWW tools that are used to locate health and safety information - directories and search engines.

#### 1. Directories

WWW directories or resource lists provide links to other resources on the Internet - usually WWW servers, and sometimes to other types of resources as well (e.g. Gopher servers, FTP servers, mailing lists, newsgroups.) Hypertext links to WWW servers usually point to an organization's site. Entries in directories may be arranged hierarchically by subject, they may be alphabetical, they may have keyword search capabilities, or they may have all three. There are both general purpose directories, and health and safety-specific directories.

#### **1.1 General purpose directories**

The most popular general purpose directory is **Yahoo!** (Yet Another Hierarchical Officious Oracle,) accessible at <u>http://www.yahoo.com</u>. Yahoo! indexes thousands of WWW sites on every conceivable subject, including sites containing health and safety information. Yahoo! also provides a keyword search capability. Health and safety information is accessible using a number of different routes:

- Health: Public Health and Safety
- Health: Workplace
- Regional: Countries: Canada: Health: Public Health and Safety
- Regional: Countries: Canada: Health: Workplace
- Government: Agencies: Executive Branch

The example below shows part of the list of sites accessible under Regional: Countries: Canada: Health: Workplace: Organizations.

File Edit					> Works	slace > 0:	ganizatio	ns - Micro	soft Inte	unet Explor	6
Tao Too	t ⊻jew F	Bhoapes	Toop H	elp							
÷.	. ⇒.	. 🕲 .	<u>_</u>	<u>a</u>	0		3	- B-	۵.	Edit	
Back	Formand	Stop	Refrech	Home		Favorites	History	Mail	Print	Edit	
fdress 🧟	] Μφ.//dit.ye	hoo.com/F	Regional/Co	ountries/Car	iada/Healt	h/w/orkplax	se/Organiz	alions/			
	egional > C	ountries ?	Canada	> <u>Health</u> >	Workpla	<u>ce</u> >					
)rganiza	ations										
			Sear	ch @ al	l of Yahoo	i C justi	his categ	ory			More Yahool
side Ya	haal										BOILE LABOR
side ta	inoo:										News: Health
- <u>X</u>	ahoot Ca	nada -	offering	a unique	ly Cana	dian per	spectiv	θ.			Administration
				-	-	-					News: Health Care
											News: Public Health
• Ce	nadian (	entre f	for Occu	pationa	Healti	h and Se	afety ((	COH8)			W. Canada Grantinan
<ul> <li>Ca</li> </ul>	nadian In	jured W	orkers A	liance -	exists t	o enforc	e and s	trengthe	n the v		YI Canada Greetings free virtual cards
	al and priperiences							ging info	rmatior	1 and	
								oms, wor	kplace	risk facto	rs, advice, and more.
• Ca	inadian Re	gistered	d Board	of Occup	ational	Hygienis	ts				
	<u>dustrial A</u> hieve safi					IAPA) -	not-for	-profit or	ganizat	tion helpin	ig Ontario businesses
						on (PPHS	sa) - no	t-for-pn	ofit safi	e workpla	ce association, assisting
• Pu		íp, pape	er, and r	elated in	dustries	in the e	liminati	on of wo	rkplace	illnesses	and injuries.
On			Anticina	tion Rect	anition	Evaluat	ion Con	trol of He	salth Ri	sks Inc. =	focusion on
• <u>54</u>	REC: Scie										locasing on
• SA											incusing on
• <u>\$4</u>	REC: Scie										focusing on

Yahoo does provides keyword search capabilities, but bear in mind that directories of this type are not comprehensive. If you want to do a keyword search for information on a particular subject, using a search engine is usually a better approach.

There are a number of directories similar to Yahoo! These include: Magellan (<u>http://magellan.excite.com</u>), LookSmart (<u>http://www.looksmart.com</u>), Lycos (<u>http://www.lycos.com</u>) and Snap (<u>http://www.snap.com</u>).

### 1.2 Health and Safety Directories

Several health and safety-specific directories are accessible via WWW. The most comprehensive directories are:

- Health and Safety Internet Directory Maintained by CCOHS, this directory provides links to health and safety-related WWW servers, e-mail-based mailing lists and Usenet Newsgroups. Site listings are annotated. Sites are accessible using alphabetical and subject lists, or by keyword searching -<u>http://www.ccohs.ca/resources/hshome.html</u>
- OSHWeb This directory is maintained by the Finnish Institute of Occupational Safety Engineering. It is a subject-oriented listing with keyword search capabilities. WWW sites and Usenet Newsgroups are included -<u>http://oshweb.me.tut.fi</u>
- Safety Related Internet Resources This directory is maintained by Christie Communications, an Alberta Canada safety training company. It includes WWW sites, mailing lists, Usenet Newsgroups, etc. – <u>http://www.christie.ab.ca/safelist/</u>
- Directory of Sites in Occupational and Environmental Health Maintained by Dr. Raymond Agius at the University of Edinburgh, this is a somewhat "Euro-Centric" directory of occupational and environmental health resources. <u>http://www.agius.com/hew/links/</u>

These directories have a variety of uses, including: browsing to see what organizations have resources available; finding an organization s site; and finding resources in general subject areas.

As an example, if you wanted to find WWW sites providing chemical-related information, you could get a list from CCOHS' directory at <u>http://www.ccohs.ca/resources/hshome.html</u>.

### 2. Search Engines

There are hundreds if not thousands of WWW sites that allow you to search various subsets of the pages accessible on the world's WWW servers. They all work approximately the same way. You are presented with a fill-in form for your query. When you submit your word, phrase or search string to the search engine, you will be presented with a list of WWW pages that match your query to a greater or lesser degree. Most search engines provide an extract from the web page referenced. Most present the search results based on their own formulae for relevance ranking. What is relevance ranking? As an example, if you searched for indoor air quality, any documents containing all three words would be ranked as more relevant than those containing one or two of the words. Documents containing all three words in the title, those containing the words near the beginning of the document and those where the words appear most frequently throughout the document would be ranked higher than documents where the words appear somewhere in the body of the document. The specific ranking rules vary from search engine to search engine, but the general approach is the same.

Some search engines offer "advanced" search features, which may include boolean logic (AND, OR, NOT,) quoted phrase searching (e.g. searching for the phrase "indoor air quality", rather than just the words indoor, air and quality) and other features for fine-tuning your search.

Some of the better search engines include:

- AltaVista <u>http://www.altavista.com</u>
- Excite <u>http://www.excite.com</u>
- Fast Search <u>http://www.alltheweb.com</u>
- Google <u>http://www.google.com</u>
- HotBot <u>http://www.hotbot.com</u>
- Northern Light <u>http://www.northernlight.com</u>

It is very difficult to say which search engine(s) will be best for you. The instructor's picks are Google, AltaVista, Fast Search and Northern Light.

The image below shows the result of an AltaVista search for biomedical waste information. Using the search string +"**biomedical waste**"-domain:com - this means that AltaVista should return documents containing the phrase "biomedical waste" from non-commercial sites.



#### 2.1 Features of popular search engines

The table below compares the search features of some of the most popular search engines.

Ranking refers to the ability of the user to change the "weighting" of specific words, making them more or less important.

Other features refer to the ability to limit the search some other way. As an example, one search feature would be the ability to limit a search by domain name.

	AltaVista	Excite	search	Google	HotBot	Northern Light
phrase	yes	yes	Yes	yes	yes	yes
boolean	yes	yes	no	no	yes	yes
ranking	yes	yes	yes	yes	yes	yes
date range	yes	no	no	no	yes	no
wildcards	yes	no	no	no	no	yes
case sensitive	yes	no	no	no	no	no
other features	yes	yes	yes	yes	yes	yes

More detailed explanations for each of the search engines follow.

#### AltaVista (http://www.altavista.com)

- phrase searching
- boolean searching
- include/exclude terms using "+" or "-"
- date range
- other constraints you can put on your search: title, anchor, link, image, host, domain, etc.
- case sensitive searching possible
- Canadian-only version (http://www.altavista.ca)

#### Excite (<u>http://www.excite.com</u>)

- phrase searching
- Boolean searching
- Include/exclude terms using "+" or "-"
- Retrieve information in a specific language
- Include information from a specific top-level domain (e.g. .ca)

#### Fast Search (http://www.alltheweb.com)

- phrase searching
- include/exclude terms using "+" or "-"
- include/exclude information from a specific top-level domain (e.g. .ca) or by organizational domain (e.g. gov.on.ca)
- retrieve information in a specific language
- limit search to page text, title, link name, URL

#### Google (http://www.google.com)

- phrase searching
- automatic "AND"
- include/exclude terms using "+" or "-"
- include/exclude information from a specific organizational domain (e.g. gov.on.ca)
- limit search to page title, URL
- retrieve information in a specific language
- relevance ranking largely based on popularity

#### HotBot (<u>http://www.hotbot.com</u>)

- phrase searching
- boolean searching
- include/exclude terms using "+" or "-"
- limit search to page title
- limit search by date range
- include information from a specific top-level domain (e.g. .ca) or by organizational domain (e.g. gov.on.ca)
- limit search to pages that include images, javascript, PDF files, etc.
- case sensitive searching

#### Northern Light (<u>http://www.northernlight.com</u>)

- search the web and/or a large full-text document collection (includes major OSH journals)
- phrase searching
- Boolean searching
- include/exclude terms using "+" or "-"
- truncation
- limit search to title, publication name, URL
- limit search by subject or category
- custom search folders

More detailed descriptions of search engines' features can be found in "Search Engine Features for Searchers" at <a href="http://www.searchenginewatch.com/facts/ataglance.html">http://www.searchenginewatch.com/facts/ataglance.html</a>

# 2.2 Health and Safety Web Pages Indexed by Popular Search Engines

This table shows the number of web pages retrieved by each search engine for each of the terms. The searches were performed during April 2001.

	Alta Vista	Excite	Fast Search	Google	HotBot	Northern Light
industrial hygiene	56480	5845	57441	143000	36000	44886
safety	18581675	4428959	10353156	16500000	8095800	6872845
occupational health	274985	18125	280829	285000	134700	194719
occupational medicine			42315	70600	25100	31904
toxicology	210350	21375	305832	509000	210400	209260
indoor air quality	61797	7675	69933	125000	46500	51900
confined space	34599	2585	48036	69600	21800	49858
ergonomics	216400	96548	232656	359000	221200	282397
exposure assessment	14541	1060	16119	44700	9400	13090
non ionizing radiation	4820	3475	6119	9090	4200	4750
protective clothing	152331	3115	106562	160000	41800	74501
respiratory protection			52917	85600	14600	41151
risk assessment	441838	17065	292224	646000	173100	230386
CCOHS	18479	1535	7356	50500	6100	11316
Total Pages	20144777	4610852	11861495	19057090	9040700	8112963

In summary, AltaVista and Google seem to have the largest number of health and safety-related pages indexed. Number of pages is not necessarily a clear indicator of relevance, but the instructor's subjective opinion is that AltaVista and Google also have the highest degrees of relevance of the search engines listed above.

### 2.3 Meta Search Engines

Although the search engines described above are extremely useful, none is comprehensive. To do a more comprehensive search, you may want to use a "meta search engine" such as MetaCrawler (<u>http://www.metacrawler.com</u>.) or ProFusion (<u>http://www.profusion.com</u>). Meta search engines perform searches on a number of search engines and directories, then combine the results, eliminating duplicates in the process. This is probably the most comprehensive way of searching by subject on the World Wide Web. Since MetaCrawler and ProFusion must reach the "lowest common denominator" of the search engines they query, they do not have as many features as the likes of AltaVista. In the author's opinion, meta search engines are only necessary when you have not been successful searching on one of the individual search engines.

#### 2.3.1 Profusion Alerts

ProFusion has an additional search service worth mentioning – Profusion Alerts. ProFusion Alert Technology gives you the ability to track Web pages retrieved by your query for updated information. Once those Web pages change, ProFusion will notify you automatically by email and send you a complete report of changes. Plus, you can create an alert that searches the Web for you - even when you're away from your computer!



#### 2.4 Workshop – Search Engines and Directories

When answering the questions below, write down the steps you used to find the answers. This will help you remember how to answer similar questions in the future.

- The US government put out a short publication a couple of years ago on back belts – something about whether they prevent injury or not. It might have been NIOSH or OSHA, but I'm not sure. Can you find this document using a search engine? (Hint: Search AltaVista, Google or HotBot and limit your search to the .gov top level domain).
- 2. I need to make a presentation on confined space entry. Are there any PowerPoint presentations available on the web that I could use as starting points?
- 3. Can you find a list of health and safety-related web sites from universities and colleges? (Hint: Look at CCOHS' Health & Safety Internet Directory)
- 4. I'm writing a research paper on the toxic effects of solvent exposures, and I want to reference a document I found on the web in my bibliography. How do I describe web pages in a bibliography? (Hint: Search for something like "citing electronic resources" in the Yahoo directory.)
- 5. I need a picture of a respirator. Can you find one for me? (Hint: Try AltaVista. On the home page, choose "Images, Audio and Video", then do your search.)
- 6. I already have a lot of information about indoor air quality, but I want to see if there is anything new in the last two months. (Hint: Try AltaVista's or HotBot's advanced search. Both allow date range searching.)