search insider

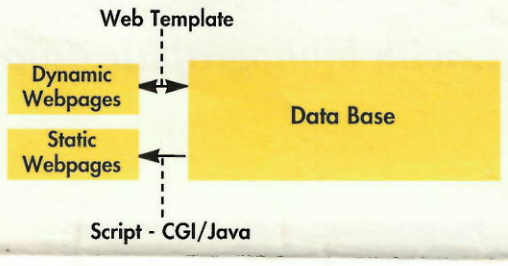
In the first of two articles on Internet searching, David Green, Head of Publishing at Informed Business Services, looks at the advantages and disadvantages of Web directories and search engines.

In that constellation of computers known as the Internet, transmitted data is split into small 'packets', an exponentially more efficient utilisation of bandwidth. This single fact has forever changed the economics of electronic publishing.

Collapsed costs have resulted in an explosion of information. Great! Couldn't easily get that type of data before. Just one problem – the Internet lacks the structured indexing of the more 'traditional' online hosts and 'free information that takes too long to find and format is expensive information' (*Information Today*, Feb 98). True ... but according to a recent survey by *Business Week*, Internet users spend 50 per cent of their online time conducting research. However there are a variety of weapons available to the information searcher's armoury and in this article, I'll be focusing on Web directories and search engines.

It is important first to distinguish between two types of Web page: static and dynamic. Static pages are those which have been created manually by a Web designer, posted onto a Web server and are available to anyone or anything that visits the Web site of which it is a part. Any changes to that page must be performed manually by the designer. Don't be confused by flashing or rotating images.

Dynamic Web pages are those created by a computer using a piece of programming known as a script (often CGI or Java). This script acts as an intermediary between the user requesting or submitting information on a Web page (the 'front end') and a database (the 'back end') which supplies or processes the information. The script slots the results into a blank page template, and presents the visitor with a unique, customised Web page – in other words, one which is dynamically generated. The diagram below illustrates this process:



Question: *Which is the odd one out of Altavista, Excite, Yahoo! or Infoseek?*
Answer: *Yahoo! is a Web directory and not a search engine like the others.*

But doesn't Yahoo! have a search facility? Well yes it does – Yahoo! uses search technology to enable the visitor to search within its directory listings in much the same way that you can use search engine technology to search within an intranet.

So what's the difference between a Web directory and a search engine? A Web directory:

- is a predefined list of Web sites,
- is compiled by (a) human editor(s),
- is categorised according to subject/topic, and
- provides useful related links.

Because Web directories (sometimes also called indexes) are compiled by humans, a qualitative decision concerning the content on each listed Web site has already been made. Therefore, with Web directories, you know you've got a head start in identifying 'the best of the Web' for the topic that you're interested in. However, precisely because they are compiled by humans, who are more expensive and slower than computers, Web directories can only ever hope to cover a very tiny percentage of what's available – the world's largest directory, Yahoo!, only covers about a million Web sites.

Web directories are broadly categorised into two camps: 'commercial' and 'non commercial'. Commercial directories are those operated for profit. Consequently they tend to be larger, broader in scope and have no (or few) dead links. Examples include:

Worldwide	Magellan	www.magellan.com
	Mining Co	www.miningco.com
	Yahoo!	www.yahoo.com
UK	UK Index	www.ukindex.co.uk
	UK Directory	www.ukdirectory.co.uk

It is important always to try more than one directory when searching for information as there can be major differences in their listings. For example, Yahoo! refuses to list any Web site featured by upstart competitor The Mining Co.

Non-commercial directories usually lack the snazzy design or breadth of coverage of their commercial cousins. However, they are often topic specific and have been compiled by someone who is an expert in or has a passion for that topic, so if there is a useful Web site for that topic, it is very likely to be listed.

Some directories of interest to information professionals would include:

- Sheila Webber's excellent 'Business Information sources on the Internet' www.dis.strath.ac.uk/business
- Informed Business Services' directory of over 700 business information links www.informed-ibs.com/training/bnet.html
- Business Researcher's Interests www.brint.com/interest.html

Nobody ever searched the Web

When using a search engine, you do not 'search the Web'. What you are searching is a database of indexed Web sites. Users enter their search term to a form on a Web page, the script interrogates the databases according to your requirements and the results are presented in a dynamically generated Web page.

These search engine databases are primarily built up by 'spiders'. Dispatched on an automatic and frequent basis by the search engines, spiders are programmes that search the Web for new Web pages, index words on those pages and match the indexed word with the URL site of the page on which it appears. AltaVista's spider, 'Scooter', visits and indexes six million Web pages a day.

And they're popular: according to Michael Tchong of Iconocast (www.iconocast.com) the top five search engines process more than 75 million search requests per day, and research conducted by CommerceNet/Nielsen Media Research in April 1997 showed that 71 per cent of frequent Web users cited search engines as the most popular means of locating useful information.

Despite this popularity, there are several problems with using search engines.

Size matters

The April 1998 issue of *Science* reported on research into Internet search engines conducted at the NEC Research Institute in Princeton, US. The survey, reported that of 320 million publically indexable pages on the Web, search engines only covered a small percentage:

RANK	SEARCH ENGINE	URL	% Web Coverage
1	Hotbot	www.hotbot.com	34%
2	Altavista	www.altavista.com	28%
3	Northern Light	www.nlsearch.com	20%
4	Excite	www.excite.com	14%
5	Infoseek	www.infoseek.com	10%
6	Lycos	www.lycos.com	3%

Search engines can't keep up now – yet the Web is projected to grow by over 1000 per cent over the next two years.

The Invisible Web

Increasingly information is accessed via the Web, not on it. Web sites are being integrated with databases of information – information which does not exist in Web page format until you, the searcher, extract it from that database by entering your query to the Web 'front end'. In other words, while the information is presented to you in a dynamically created Web page, it never existed in Web page format until that moment. Search engines can only index HTML Web pages. If the information is in a database, then search engines won't find this information because they can't search 'the invisible Web'.

Chinese Walls

Search engines can only search publicly available Web sites. Database integration is one trend in Web publishing, 'subscriber only areas' are another. Much 'value-added' (read more in-depth) content on the Web is now hidden behind 'search walls' in special password protected sections of a Web site which are not accessible to search engines or non-subscribers.

Portalisation

In late 1996 search engines began to encourage visitors to hang around their sites for longer, to transform them to destination sites themselves. Why? To sell more ad inventory. The search engines started to mimic AOL, CompuServe and others by offering free email and content. This has had two effects: a realignment of control and influence of the information distribution chain, and a growing disillusionment with the core services provided by the main search engines.

This disillusionment, however, has in turn fostered a new market for other exciting search engine technology applications. These are helping to circumvent and overcome some of these search engine problems.

One already well entrenched search solution is 'meta searching'. In a nutshell, meta search sites enable the user to search across several search engines and Web directories simultaneously. The results are often de-duplicated and then re-ranked for relevancy. Examples include Dogpile (www.dogpile.com), which searches 14 different search engines and directories but doesn't eliminate duplicates; Mamma (www.mamma.com) which only searches seven engines but deduplicates and re-orders according to its own relevance-ranking algorithm; and Metacrawler (www.metacrawler.com) which is one of the earliest and most highly rated search engines.

Another interesting development has been 'natural language searching' (available only in English to date, as far as I am aware). Most search engines only find exact matches of words, with little consideration of semantics (it ignores related words), or use of thesauruses (some search engines offer a 'more like this' option with results), and with no consideration of syntax (such as where words appear in a clause). Natural language search engines aim to overcome these problems. One of my favourites, for relevance of results, is the Electric Monk (www.electricmonk.com). Named after a character in a Douglas Adams novel, this search engine operates by performing a syntactical analysis of the search term using artificial intelligence natural language algorithms. These convert the user's search term to a complex Boolean query, which is then executed on the AltaVista database.

The main search engines have now come full circle. The combination of focusing on 'portalisation', together with competition from new search engines with added functionality, has forced them to refocus on providing meaningful search results to searchers' enquiries. Both HotBot and AltaVista have incorporated Web directories, AltaVista added natural language searching in October, and HotBot launched 'Direct Hit' technology in August. In fact, we're at the beginning of a very positive time for anyone who uses the Web as a research tool.

In part 2:

- Search engine metatool plug-ins (eg Discovery for Altavista, Express for Infoseek)
- Searching the invisible Web — search utilities/intelligent agents and XML
- The future of search engines:
  - Google (results dependent on other sites mentions)
  - Direct Hit (results dependent on other visitors' choices)
  - Clever (results dependent on other pages' links)