

The background of the slide is a dark blue gradient with abstract, lighter blue geometric shapes. These shapes include parallel lines, zig-zag patterns, and rounded rectangular forms, creating a modern, technical aesthetic.

Searchability: Six Key Messages

by Dr. David Hawking

The Six Key Messages

1. Poor searchability on intranets and websites costs money and impedes achievement of business objectives.
2. Good searchability is needed at all levels.
3. Understand how modern search engines actually work - or your optimisation efforts will be misdirected.
4. Don't engage in dodgy optimisation or you may annoy people and be blacklisted by search engines.
5. Don't rely on subject and description metadata
6. Optimise your site to help searchers — good content, simple URLs, incoming links, useful anchor text.

Message 1

Poor searchability on intranets and websites leads to:

1. Wasted time & frustration for all
2. Increased reliance on less cost-effective delivery mechanisms
 - a. Phone calls
 - b. Email
 - c. Face-to-face
3. Lost employee productivity

Government	- Ineffective delivery of programs and services
Industry	- lost competitiveness
Education	- Reduced access to learning resources - Lost competitiveness with other institutions

Message 2

Searchability is needed at multiple levels

1. Whole-of-web
2. Whole-of-organisation
3. Local (site / agency / portfolio / portal / department) level

Government	The public often doesn't know which government level or agency is responsible for something that affects them. It's important to be findable in broad search environments.
Industry	It's important that your websites are findable in external search engines, not only when the name of your company (e.g. Sony) is used as the search term, but when the query matches your products or services (e.g. television, DVD player).
Education	Prospective students and staff may search for courses or research activities in a general search engine before becoming aware of your site.

Message 3

Understand how modern search engines work

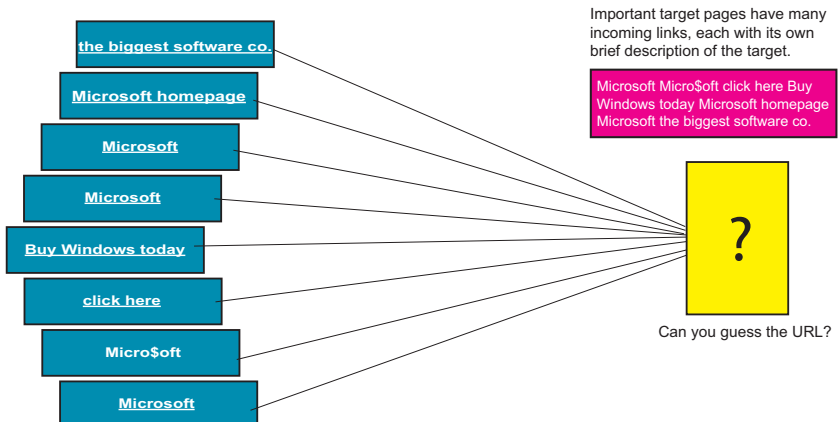
There are three distinct phases:

1. **Gathering content for indexing.** Web content is gathered by crawling, i.e. following links from a starting point, which may be a directory page like dmoz.org or a company home page.
 - If your site is not linked to by pages already in the crawl, it will not even be in the index!
 - URL submission is not reliable
2. **Indexing.** To give good query response, the search engine must build index structures which enable it to quickly identify which documents best match a word.
3. **Query Processing / Ranking**
 - Modern search engines consider tens or even hundreds of factors when deciding how to rank documents in response to a query. These may include query word counts, page length, counts of incoming links, URL characteristics, and presence of query words in incoming anchor text. Words in anchor text from other people's documents are more important than words in the document itself.
 - Of these ranking factors, referring anchor text and incoming links are key to ensuring important resources are ranked above millions of others which also match the query.

Incoming links are Invaluable

What is anchor text?

The text highlighted in your browser to indicate a link you can click on, is called anchor text. The figure shows anchor text in each of a set of pages (blue) linking to a target page (yellow). A search engine combines all of this anchor text (pink) and indexes it, as though it were part of the target.

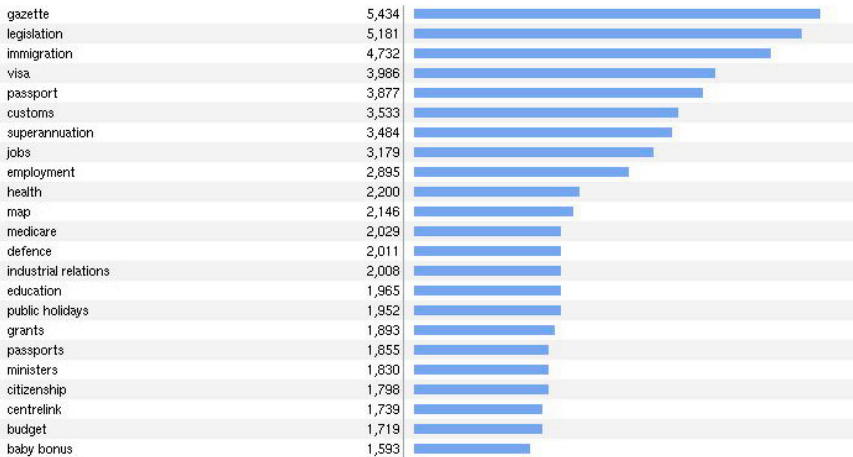


Web-style retrieval

- On the Web, ranking by “similarity” to the query isn’t good enough!
- The best results aren’t necessarily the ones which contain the most occurrences of the query words — those are probably spam!
- Some results such as homepages of key sites, or reliable directory pages are much more valuable to the average searcher than isolated paragraphs in individual documents.
- The Company X home page is a better result than the Company X privacy policy, but the privacy policy almost certainly uses “Company X” more often.
- Queries are short - result sets are very big.
- Most searchers use a **search-and-browse** approach — they type in a very short query and rely on the search engine to rank the homepages of sites they want at the top of the list. From there, they search and browse within the chosen site.

Search & Browse paradigm

The most popular queries on the Australian government search service (below) typically consist of a single word. Even though these queries may match hundreds of thousands of pages, this approach is effective because the search engine ranks the key sites (e.g. the home page of the Public Service Gazette, the Federal legislation website etc.) at the very top of the search results.



Message 4

Don't engage in dodgy optimisation

It's in your interest and in the interest of your customers and clients to optimise your site for visibility in external search engines, using suggestions like those in Message 6.

However, if you go too far you may only succeed in annoying people and having your site blacklisted. Examples of "dodgy" optimisations include:

- Paying a disreputable optimisation company to set up a forest of fake websites and IP addresses which link to yours
- Cloaking your site (i.e. delivering different content depending upon whether the request comes from a person or from a web crawler).
- Stuffing your pages with invisible keywords
- Arranging large numbers of links with misleading anchor text. For example, making a car maker's website rank highly for queries related to coffee or dog food!
- Inserting links to your page in all the open blogs you can find, regardless of relevance to the topic.

Message 5

Don't rely on subject and description metadata

Some people, particularly in libraries and government, believe that standardized metadata is essential to achieving effective search. While metadata such as author, title, journal, etc. is invaluable in the sort of searches conducted in a library environment, subject and description metadata often repeats content which is already indexed, and seldom adds value. External search engines can't afford to take much notice of subject and description metadata because:

- a. It can be a source of spam.
- b. It's often misleading, missing or inaccurate
- c. It doesn't improve search results.

[See JASIST 58(5), March 2007, pp. 613-628. *Does Topic Metadata Help with Web Search?* http://es.csiro.au/pubs/hawking_zobel_jasist.pdf]

Note, however, that modern search engines often display the description metadata from a document as the summary in search results.

Message 6

Optimise your site to help searchers

1. Make it easy for search engines to crawl your site.
 - a. Ensure all search-useful content is reachable via simple HTML links. (If it makes sense to have links in JavaScript, Flash, PDF, Word, etc., make sure they are all listed again in a simple HTML site map.)
 - b. Use a robots.txt file to avoid search engines wasting their time crawling content with no search value.
2. Prevent your web server generating error pages such as “Your browser doesn’t support JavaScript” (or frames etc.) when accessed by web crawlers. Instead, generate a text version of the page to ensure that your content is indexed
3. Encourage people to link to your pages (those with search value).
 - a. Publish stuff that people will want to link to. (If possible!)
 - b. Only publish each thing once, using a single consistent URL. (Every time you change a URL, all the external links to it break!)
 - c. Using simple, meaningful URLs. (Other web authors don’t like linking to long, complex URLs, particularly when they contain punctuation and parameters.)
 - d. Ask appropriate other sites to create links
4. Encourage linking sites to use anchor text which matches queries which are important to your business. For example, **ACME plastic cups** rather than just **ACME** or **click here**.
5. Make sure your web content complies with W3C standards. Use validation tools to check this.
6. Choose good titles for your pages - e.g. **ACME plastic cups - about us**
7. Make sure that the language you use in titles, content, and anchor text matches the language which people are likely to use as queries. Governments often use more sophisticated names for things than do the searching public. For example, “permit to acquire a longarm” rather than “gun licence”
8. Monitor your query and web server logs to find out what queries people do most commonly use.
9. Use thesaurus expansion and search short-cuts on your local site search but note that they are no help to external engines.

Published by:
Funnelback Pty Ltd
Internet & Enterprise Search
PO BOX 1441, Dickson, ACT, 2602
Ph: +61 2 6175 8500
www.funnelback.com