Choosing a search facilty

Helen Varley Sargan

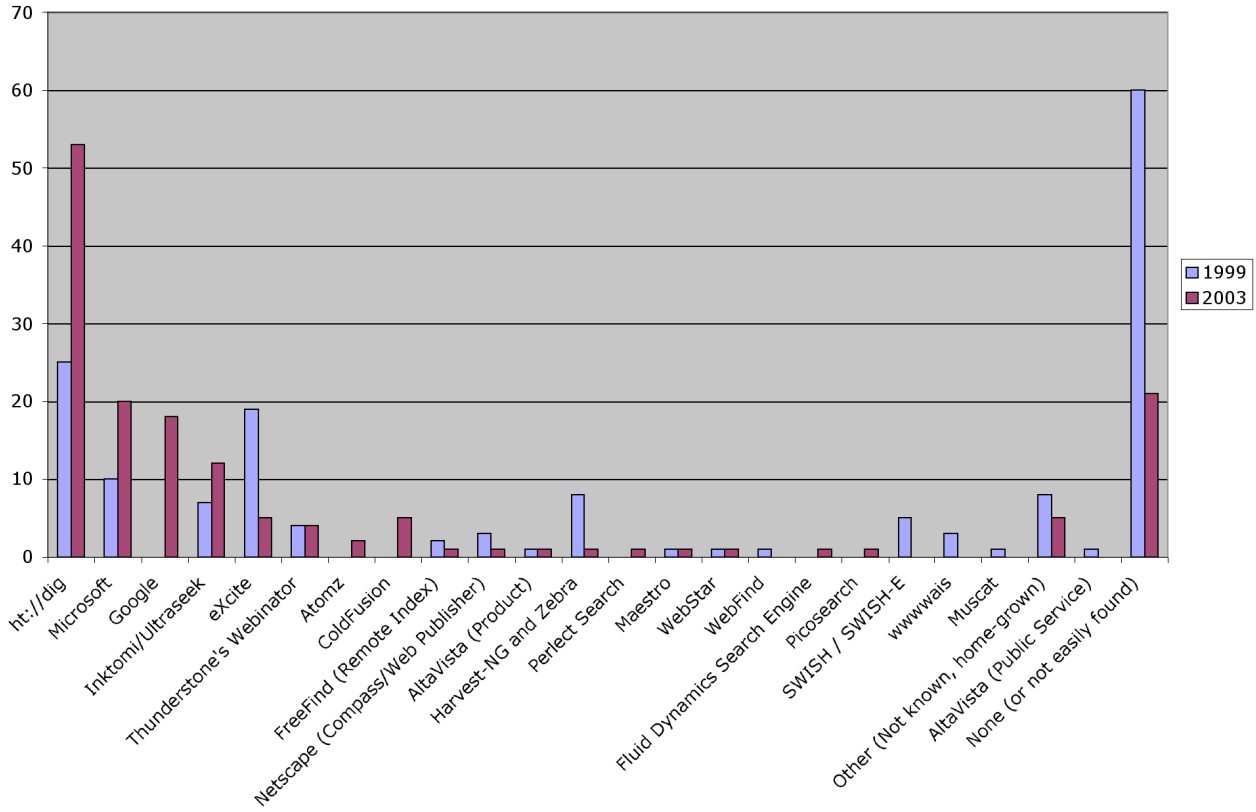
University of Cambridge Computing Service

Introduction

How can you provide the type of search facility your institution wants, needs and can afford? This session will give an overview of the options and issues.

Search facilities in use in UK HE (1999-2003)

I first gave a session about search engines in 1999. Comparing this data with some for February 2003 gives the following results:



What does this tell us?

Over the fours years we see the following changes:

- Many more sites are now using a search engine
- Free solutions are popular (ht://dig, Microsoft, Google and Atomz) although there has been an increase in paid-for software as well (Inktomi/Ultraseek, and Coldfusion – it comes as part of paid-for software)
- There is evidence that once a solution has been found (particularly paid-for) sites will stick with it (for instance Webinator, Atlavista product, Inktomi/Ultraseek). Hardly suprising.

Issues with categories of search software

Search software falls into one of these categories:

- Free products
- free (open source) software (ht://dig, Harvest/Zebra) or
- free-of-charge externally hosted indexers (Google, Atomz, FreeFind)
- Products that are part of other software
- Commercial products

Issues with free (open-source) software

- Free or open source software can be developed intermittently and incompletely and it is difficult to know when this might happen.
- There will be a requirement for manpower to research how appropriate the software is for your circumstances, tailor or develop the solution to your requirements and to run and trouble shoot it.
- It may be difficult to tell whether your software will ever be able to do all you want.
- There may be little support, although online communities often provide a wealth of help.

Issues with free-of-charge externally hosted indexes

Free-of-charge externally hosted indexes have been the big growth area since 1999. They started as a solution for small sites with little expertise but have grown (particularly with the introduction of Google University search) into a workhorse solution for many larger sites. The major attractions are no cost and no manpower implications, and with Google, also, the brand name and experience users already have with it. The drawbacks are as follows:

- You have little to no control: The index is hosted externally, often in the US. If network connections are compromised it will not work.
- You will not have any say over which web servers within your domain are indexed
- You will not have any say over how often the index is updated. Google University search updates in weekly blocks but only once per month.
- Documents that are only visible internally will not be indexed

All these free-of-charge externally hosted indexers are spiders that may be controlled from your web server robots.txt file, should you want to restrict parts of the content from being indexed.

- Google information for Webmasters - http://www.google.com/webmasters/

Issues with 'built-in' software

Of the products in the table, Microsoft, Coldfusion, WebStar, are all associated either with the web server or the content management software that is running on it. My concern over using these products would be:

- the coverage of the results
- whether you are limited to indexing one server

Problems can arise when indexing from built-in or add-on indexing software, such as Microsoft Site Server or WebStar indexing software, which may be vendor specific. This is because of the differences in the ways web servers respond when the indexing robot approaches them (the API of the web server). The APIs of the above servers are subtly different from, say, Apache, and the indexing software may have been written with their particular API in mind, so may baulk at unexpected server responses. This problem is more likely to arise when you are indexing a large number of servers (and encounter more different types of server software during the process).

Issues with commercial software

Commercial products such as Inktomi/Ultraseek and Altavista (product):

- works 'out of the box' so ongoing work is reduced
- may be tailored to produce exactly what you need (depending on what that is), and will generally offer more features but some of these may be opaque
- limited (or possibly no) scope for development of features, since application is a 'black box'
- (usually) ongoing product development and support
- cost can be very high and recurring (annual maintenance)

In the US and Canada, Google sell the Google Search Appliance (http://www.google.com/appliance/), which you lease as an internal service and control yourself, and a customised service is also available, but both are very expensive.

Rolling your own

There is a temptation to hire a developer and create just exactly what you need, however life is never that easy. A few problems:

- creating an exhaustive list of features you want is almost impossible
- you never know if the final product you want **is** attainable until it's been attained
- the project could very possibly be open-ended in terms of time and effort (and money)
- Forth bridge scenario for updating and maintaining
- you have to use something else while development is ongoing

What do you need?

It is essential that you start off the exercise with a clear idea of what you are looking for. Things to think about may include the following:

- What manpower and/or money is available for the project and will this be seen as a one-off cost?
- Do I want to/ am I able to run this on the web server, on a separate machine, or have someone else host it?
- What platform do I want to use (is there the expertise or facilities for using a different platform)?
- How many servers do I want to index (ballpark figure of number of pages to be indexed useful here too)?
- Is the data to be indexed subject to frequent change, if so in part or as a whole?
- What type of files do I want indexed (just HTML, or including PDF, Office files, etc.)
- What type of search facilities do I want to offer (keyword, phrase, natural language, constrained searches)?
- Do you want to offer separate searches for external and internal users?
- Can you ever solve the problems with just a search engine? – see http://vivisimo.com/solutions/universities.html for 'clustering engine'

Controlling access

Intentionally

It is a useful lesson that if you don't want people to read files, then they shouldn't be on the web server. Adding a new indexing facility should remind people to 'spring clean' their files and remove all the information that is no longer pertinent.

There will be some directories that you do not want your indexer to look at and index. When using a spider or robot based indexer, controls over indexing are through a number of means and will be observed by Internet indexers such as AllTheWeb, Google, and HotBot, as well as your local indexer. Obviously, if you can kill all your indexing requirements with one stone it will save you work in the long run.

These controls are:

- the robots.txt file
- robots metadata tag giving noindex and nofollow information (and combinations) in individual files

All 'proper' search engines will observe a robots.txt file and do what it says, and observe the robots metadata tag (see http://www.searchenginewatch.com/webmasters/article.php/2167891).

At another level, access to branches of a web server can be limited by the server software. Combining access control with use of metadata can give information to those within the access domain and some limited information to those outside.

robots.txt

Robots.txt sits at the root level of your web server and give information about what should not be indexed. An example might look like this:

```
# A comment line just to show what one looks like; it is ignored.
User-agent: *
Disallow: /bin/
Disallow: /cgi/
Disallow: /includes/
Disallow: /tmp/
Disallow: /~
Disallow: /stats/

Disallow: /local.html
```

Another example, including a reference to a named search engine robot, might look like this:

```
User-agent: Ultraseek (webmaster@ucs.cam.ac.uk)  # local search engine
Disallow: /bin/
Disallow: /cgi/
Disallow: /includes/
Disallow: /tmp/
Disallow: /~
Disallow: /stats/
Disallow: /local.html

# tell all others to go away
User-agent: *
Disallow: /
```

Robots meta tag

If the information providers can neither update the robots.txt file nor request changes to it, they can use robots META tag to specify within an HTML page whether indexing robots may index the contents of the document and/or follow links from it to other documents. This is of limited use, since it can only be used in HTML documents, but does not require changes to any robots.txt file. If there is also a robots.txt file, the exclusions there are processed first.

All META tags must be placed within the <HEAD> section of the HTML. The name attribute must be "robots", and the content attribute contains a comma-separated list of directives to control indexing, chosen from

- INDEX or NOINDEX – allow or exclude indexing of the containing HTML page.
- FOLLOW or NOFOLLOW – allow or exclude following links from the containing HTML page.
- ALL – allow all indexing (same as INDEX,FOLLOW)
- NONE - no indexing allowed (same as NOINDEX,NOFOLLOW)

The values of the name and content attributes are case-insensitive. Repeated or contradictory values should be avoided. The defaults are INDEX,FOLLOW, i.e. all indexing is allowed. Note that INDEX and/or FOLLOW cannot override exclusions specified in a robots.txt file, since an excluded document would not be fetched and the tag would not be seen. Also, the NOFOLLOW exclusion applies only to access through links on the page containing the tag - the target documents may still be indexed if the search engine finds links to them elsewhere.

Ignoring the "shorthand" ALL and NONE variants, the following examples show all the possible combinations:

<meta name="robots" content="index,follow">

<meta name="robots" content="noindex,follow">

<meta name="robots" content="index,nofollow">

<meta name="robots" content="noindex,nofollow">

Controlling access unintentionally

Depending on how your pages are generated, you may be excluding indexers. Problems will arise with indexing of:

- Framed sites (see http://www.searchengines.com/frames.html)
- Graphics (including Flash and Shockwave) – use alt text
- Pages requiring passwords, registration or cookies (local search engines can be given passwords)
- XML
- Java applets
- Acrobat files (except Google)
- Dynamic pages with queries in the URL (except except Google, Altavista, FAST)
- Multimedia files (but see later)

Adding metadata

You can add information to your html files that will be indexed in addition to the content of the rest of the page. The standard model is for description and keywords - the description being used as the summary instead of the start of the file, and the keywords being picked up as search terms. You cannot depend upon the description and keywords metadata tags being used - many of the major search engines now ignore keywords but use description (except Google) - but if they work for your local search facility and make it more valuable, they must be worth pursuing.

For instance;

```
<HEAD>
<TITLE>Stamp Collecting World</TITLE>
<META name="description" content="Everything you wanted to know about stamps, from prices to history.">
<META name="keywords" content="stamps, stamp collecting, stamp history, prices, stamps for sale">
</HEAD>
```

Dublin core is another standard for metadata, which in its simplest form consists of 15 terms (see http://webreference.com/xml/column24/index.html). No major web indexers use Dublin Core but specialist search engines do.

Title

Although title is a standard tag, it is the most important piece of information for indexing purposes. There may be a dilution effect (keyword in 6 words ranked higher than keyword in 10 words), so keep the title short and succinct. Remember that this is what goes into bookmark lists, so it also needs to adequately reflect the content of the file.

Metadata and PDFs

When you generate a pdf, information may be inserted into the description, keywords and title slots from the original file, unless you tell the processor otherwise. You can change these values in the pdf by using the full version of Acrobat (Document properties>Summary), but increasingly users generate their own pdfs and do know know that these values are there, or how to influence them before the pdf is generated.

PDFs from Word

By now there are many versions of Word about in the world and this advice does not apply to all of them. In addition to versions of Word being a difficulty, the process is handled differently by various means of producing an pdf file. In the main, the most information is passed from Word if a postscript file is distilled, rather than a pdf file written

directly. Word may take the information from the following fields in the 'Properties' information, and use it as metadata:

Original File Properties	Mapped Meta Tags
Title	title
Subject	subject
Author	author
Keywords	keywords
Comments	doccomment
Last Saved by	lastsavedby
Revision Number	revisionnumber
Category	category
Abstract	description

PDFs from scans

PDFs created from scans, either as a graphic, where no metadata will be present, or OCR, where the metadata will probably be faulty, need careful checking before being made available on the web.

Changing metadata

Edit the metadata using Acrobat and reindex the document (if you are able to) as soon as possible.

Further info

- http://www.searchtools.com/info/pdf.html

Indexing other pages and filetypes

To index dynamically generated pages in a foolproof way, the server or the page generating software should be used to rewrite the URL to a static form. For more information, see excellent article at http://www.searchtools.com/robots/goodurls.html (with refs).

Indexing graphics and Flash will hinge on supplying information in the alt tag (which search engines index) and/or using the longdesc tag or a D-link, which will be followed by some search engines - if you follow accessibility guidelines then the information should be available for search engines. For more info see:

- Macromedia Flash accessibility info http://www.macromedia.com/macromedia/accessibility/features/flash/hints.html
- How search engines see Flash files http://www.searchguild.com/seflash.html

There is a multiplicity of other filetypes that may have metadata that can be indexed by some search engines, amongst these are images, audio (including mp3) and video. To find out more see http://www.searchtools.com/info/multimedia-search.html.

See reference on What search engines can find - http://www.teacherlibrarian.com/pages/infotech30_5.html

Indexing OPACs

Increasingly, OPACs are seen as a repository of information that may be searched via a web search instead of (or as well as) via its own search interface. If the OPAC is run out of a database using PHP and MySQL, for example, the records can be rewritten with 'real' URLs and be indexed. An example of this is at the Fitzwilliam Museum (for example record see http://www.fitzmuseum.cam.ac.uk/opacdirect/2811.htm). We do not index these locally as

there are 90 000 records and we do not have the spare capacity in our indexing licence, but they will be indexed by Google, when it gets round to it (see http://www.fitzmuseum.cam.ac.uk/robots.txt).

Further info

Making dynamic pages searchable

- http://www.searchtools.com/robots/goodurls.html (with refs)
- http://www.traffick.com/article.asp?aID=106
- PHP: http://www.phpbuilder.com/columns/tim19990117.php3

More general information about search engines

- http://www.searchengines.com/
- http://www.searchenginewatch.com/resources/article.php/2156591
- Tutorial from Words in a row http://www.wordsinarow.com/seo.html
- http://www.searchtools.com/
- Review of free site search tools: http://www.microdocs-news.info/newsGoogle/2003/05/29.html#a656
- Google Dance syndrome: http://www.searchenginewatch.com/sereport/article.php/2216081
- http://www.customultraseek.org/otherengines.html
- http://www.researchbuzz.com/
- http://www.searchengineworld.com/
- Search engine forum: http://www.ihelpyouservices.com/forums/forumdisplay.php?s=f684b3962e8a0ef8d2555f1689ac67da&forumid=31
- Search tools chart - http://www.infopeople.org/search/chart.html

Appendix: Search engine details

Free search facilities

Some of the products listed below will index either a single server or a group of servers. In many cases it is difficult to find out exact capabilities without installing the product. See http://www.searchtools.com/tools/tools-opensource.html for a longer list.

Search engine	Version	Platforms	Memory and disk space	Searchable document formats	Notes
Alkaline http:// alkaline.vestris.com/	1.9 March04	Linux, (intel/Alpha) FreeBSD SGI Irix Solaris BSDI, BSD/OS Win NT		HTML, ASCII, filters for PDF, User defined filters may be made	Free to non-commercial sites
Glimpse & Webglimpse http://webglimpse.net/	4.18.0 and 2.11.0 June 2004	UNIX of various sorts, with more coming		HTML, ACSII	New effort not connected with original developers Free for non-commercial use. Hosted service also available at low cost. Webglimpse is the spider, Glimpse the indexing software.
**ht://DIG http://www.htdig.org/	3.2.0 June04, but 3.1.6 Feb 2002 ack. To run faster than releases since	Sun Solaris 2.X SPARC (using gcc/g++) Sun SunOS 4.1.4 SPARC (using gcc/gcc 2.7.0) HP/UX A.10.X (using gcc/g++) IRIX 5.3 and 6.X(SGI C++ compiler. Unknown version) Most Linux and BSD – inc Mac OSX(using gcc/g++)	disk space – approx 12KB for each document for wordlist database, 7.5KB without wordlist.	HTML ASCII	You will need a Unix machine, a C compiler and a C++ compiler. With libstdc++ instlled and Berkeley 'make' Will index multiple servers understanding http 1.0 protocol. Developer site at http://dev.htdig.org/
Isearch http://www.etymon.com/tr.html	v 1.47 available 2002	Unix machines from Linux PCs to Crays		wide range of document types, with facilities to add new types	No longer under development
Nutch http://www.nutch.org/docs/en/	In development July2004	Java (Tomcat)	Disk space - around a total of 10kb per web page		Java crawler, and an indexer and search engine based on the Lucene open source search code library See http://www.searchtools.com/tools/nutch.html
SWISH-E (SWISH Enhanced)	2.4.2 March2004	Many unix based OSs Windows (all 32-bit varieties)	disk space – approx 1-5% of size of HTML data	Large number of file formats and filters also	Discussion lists and newsgroups, and on the website

http://swish-e.org/			HTML data	available	
Thunderstone Webinator http://www.thunderstone.com/texis/site/pages/webinator.html	5.0 2004	Lots of Unix flavours Linux (Intel) Windows NT x86		ASCI HTML Other formats in the commercial versions	Technical support via a mailing list. The free version is limited to 10,000 pages per index
Zebra http://indexdata.dk/zebra/	1.3.15 Jan2004	Many Unix flavours and Windows		Variety of filetypes and large databases	Mailing list but can buy optional support

Commercial products (not supposed complete)

All of these products will cost real money but many will negotiate a price, so do not be put off from asking about prices or immediately write off using a commercial product. The money spent may well be saved by staff having no development work to do and having access to ready technical support. Many of these products have a limited-time trial version for you to assess before you commit yourself to buying, but you may have to pre-register with them to get access to trial software. Information on web sites varies enormously, but check the basic facts there before you go any further with assessment.

Commercial products are marketed primarily to companies, not to academic institutions, and information about them reflects this. It may not be readily apparent how or if the software will work in your particular environment until you investigate, particularly if you are seeking to index a group of independent servers that are not on an intranet, or are wishing to produce indexes of subgroups of information.

Some of these products will support metadata but the information is not readily available so no information about metadata has been recorded. Support of Dublin core metadata is almost non-existent.

Search engine	**Version/Price**	**Platforms**	**Searchable document formats**	**Notes**
ALISE http://www.pspinc.com/ntsg/htm/pro-alise.htm	2.0 US$1999	Visual Basic		Up to 500 000 docs
Fast WebSearch (now incorporates AltaVista Enterprise Search) http://www.fast.no/		Solaris Intel: NT, Linux, BSD (FreeBSD), Solaris Alpha: Digital Unix, NT.		Comes with hardware option for generating and searching extremely large indexes Demos available
Hummingbird SearchServer http://www.pcdocs.com/products/km/ss_overview.html		Windows, some Unix flavours	Over 200	Supports Korean and Asian languages and Java
InText http://intext.com/	2.0	Windows NT UNIX	Wide range	

Mondosearch http://www.mondosearch.com/	5.1	Windows NT (.NET)		Multilingual. Indexes many different types of content. Has content management add-on. Free trial version available
Oracle http://www.oracle.com/				Indexer for sites generated by an Oracle database
Smartlogik Discover (was Muscat) http://www.aprsmartlogik.com/products/discover/			Many	Almost no technical information
Thunderstone Webinator http://www.thunderstone.com/texis/site/pages/webinator.html	5.0	Lots of Unix flavours Linux (Intel) Windows NT x86	Many formats in commercial versions	Technical support via a mailing list. The free version is limited to 10,000 pages per index
** Ultraseek [Verity] http://www.verity.com/products/ultraseek/	5.2.2 2004 Discount for academic use,	Sun Solaris 2.5 and above Linux Windows NT 4.0 and above	Many	Full feature trial download for month available. Numerous awards
Web Express http://www.convera.com/Products/products_we.asp		Many	Over 200	

Synopsis

Search engine software that is available free of charge is generally either a cut-down version of a commercial product that is limited to producing a small index (**Lycos Site Spider**), or a product that might require quite advanced expertise to set it up correctly and keep it running smoothly (there are, of course exceptions to this). Maintenance of products is a problem area - for a server manager to install and configure a search engine only to find its development is discontinued or it is turned into a commercial product is a blow. Many free products are for Unix platforms since this is where such expertise and enthusiasm for free software lies.

The Perl-based search engines suffer from the disadvantage that the whole index needs to be loaded before a search can be done, and these products might have a limited life when more engines written in Java are available. Java-based search engines have the problem that users have to be running Java-enabled browsers to use them, and many users prefer to disable Java because of security problems. Several other Perl and Java based search engines are available, other than those listed here -see list at http://www.searchtools.com/tools/tools.html but likelihood of '10-day wonders' is high.

While it requires some technical expertise **ht://dig** does accomplish the job with no direct cost, although the index files may grow excessively large and the day-to-day running may require large amounts of time from a technically able staff member.

The only way to find **commercial products** that really are suitable for your needs is to pay close attention to making your 'shopping list', investigate available information about the products capabilities, then talk to the local contact. We were seeking a product that was well supported, had a good interface, ran under Unix, was essentially self-managing, and could index a large number of diverse web servers. The product that appeared most suitable was **Ultraseek**. We were able to download a trial version (restricted to one months' use) and use it to confirm its suitability before buying a licence for the product. I would suggest that if you cannot use the product on a trial basis first, you shouldn't buy it. Be aware of how much time and effort you are investing into your search engine – should it start to decline or another product appear and transcend it, you will not want to feel incapable of moving on.