Specialist Tools for Tackling the Hidden Web

Karen Blakeman
RBA Information Services, Reading, Berkshire, UK
karen.blakemam@rba.co.uk

INFORUM 2006: 12th Conference on Professional Information Resources
Prague May 23rd – 25th, 2006

Abstract: *When we fail to find information through our favourite search tool we often blame the so-called "hidden web". Some industry commentators imbue the hidden web with an aura of mystery as though esoteric, magical incantations are needed to reveal its secrets. The reality is more prosaic. The size of the search engine databases is part of the problem. Yahoo claims to have over 20 billion pages and the rest range from 5-12 billion pages. Valuable information can so easily be "lost" because of the huge volume of data. To make matters worse, information is now presented in an increasing variety of formats: ebooks, photographs, videos, podcasts of interviews and conference presentations, blogs, RSS feeds, TV and radio programmes. The possibilities seem endless. Tackling this level of information overload requires lateral thinking on the part of the searcher when building a search strategy. Google is a fantastic search tool but not foolproof. It is not comprehensive and other services may be more appropriate. This presentation looks at the specialist tools and techniques that are available, and how they can help us uncover those essential but elusive resources.*

What is the hidden web?

The term “hidden web” was originally used to describe pages and sites that search engines failed to index, either because of the way in which the pages were designed or because the information was password protected or held in a database. This missing information is sometimes referred to as the invisible web, the deep web or the dark web; but is that really the problem? Pages may be indexed by the major search tools but relevant documents could be buried at number 2,300,279 in your results list and most search engines rarely display more than 1000. In fact, most of us rarely go beyond the first 20-30 hits in our results.

We certainly do have a problem and it is the result of the ever-expanding web and the increasing range of types of resources that are now available. It is bad enough having to deal with the enormous number of web pages out there – Yahoo claims to have over 20 billion - but we also have news, pictures, video, audio and podcasts, books, institutional repositories, wikis, blogs, RSS feeds, specialist databases, formatted files (PDF, XLS, PPT) etc. to deal with. Finding a needle in a haystack is child's play compared with trying to locate key documents that are poorly ranked for our search terms. Perhaps the “hidden web” should really be called the “massive web”?

There is also the disappearing web: documents and pages that have vanished from a web site. Sometimes they are still there but the site navigation or search engine does not find them any more.

Tackling the massive web – general tricks and techniques

Your search has found 2,726,780 results. How can you improve the relevance of your search and bring key documents further up the list?

1. Imagine what you would like to appear in your ideal document and include those word in your strategy. For example, If you want to know the market share of various brands of beer in a country use the terms 'beer market share' together with the country name and the names of the beers you would like to see in the document.

2. Ask or partially answer your question. For example:

   “how fast can a hippopotamus run”

   “a hippopotamus can run at”

3. Repeat the most important term or terms in your search strategy. For example the following give different results:

   beer market share belgium czech
   beer market share belgium czech czech
   beer market share belgium czech czech czech
   beer market share belgium belgium czech czech czech

   Phil Bradley has called this the Google Sinker, but it can be used with many of the other search engines.

4. Use synonyms or variations of your search terms. In Google use the tilde (~) before the word. For example ~banking will find pages containing the words banking, finance, financial, commerce.

5. Use a search engine that suggests alternative strategies. For example Ask, Exalead, AlltheWeb, Livesearch.

Use Advanced Search options

Just making better use of the mainstream search engines and the advanced search options such as file formats can bring better quality information nearer the top of your list. For example, search for PDF and DOC (Word) formats for official documents and industry/market information, XLS (spreadsheet) for statistics and data, PPT (Powerpoint) for presentations by experts, RSS/XML for current news,and opinions.

Limiting your search by site or domain can narrow down your search by type of organisation, for example site:gov will restrict you search to US government pages. You can even search a single web site, for example site:europarl.eu.int limits your search to just the European Parliament web site. This is invaluable when trying to track down documents on a large site with poor navigation or internal search engines, and especially information that seems to have disappeared from the site.

Use a different search engine

Google is not the only search engine, nor is it the most comprehensive. It can also be the most out of date. Different search engines cover different sites and pages and, more importantly, sort the results for a given search strategy in different ways. Thumbshots Ranking is a tool that enables you to compare the first hundred results of two search engines at a time and analyse the overlap, that is the number of links that are common to both. The degree of overlap between, for example, Yahoo and Google varies depending on the search but can be as little as 4-7% or as much as 70%

Yahoo Search search.yahoo.com

Re-launched in January 2004, a surprising number of people are still unaware of this excellent alternative to Google. It claims to have over 20 billion pages in its index and supports full Boolean search options. It has most of the Google advanced search features and indexes the first 500 K of a document as opposed to Google's 100 K (important when searching for PDF or PPT file formats). The image search is far superior to Google's in terms of relevance and there is an

RSS/XML filetype in the Advanced Search if you are looking for blogs or RSS feeds.

Ask (formerly Ask Jeeves) www.ask.com

Ask has had a major overhaul and is a worthy competitor in the search engine arena. It uses the Teoma database and has a very useful "Zoom" that suggests terms and strategies – listed to the right of your results - for narrowing down and broadening your search.

Exalead www.exalead.com

A relative newcomer to general web search, Exalead has resurrected some of the search functions that were dropped by the the major search engines long ago. There is a NEAR command that by default looks for terms within 16 words of one another but you can specify the number by using NEAR/n where n is the number of words separating your term. You can use the asterisk (*) as a wild card at the end of a word and after a full stop in the middle of a word to stand in for one or more letters. You must start and end the term with a forward slash. For example, /psych.*ist/ will find psychologist, psychiatrist, psychoanalyst etc.

AlltheWeb Livesearch www.alltheweb.com

Owned by Yahoo, this new search tool suggests alternative search strategies as soon as you start typing and changes the results display as you enter more words.

MSN Search search.msn.com

Relaunched in Autumn 2004, this offering from Microsoft offers most of the advanced search features one would expect from a major search engine. Results tend to be more "consumer" orientated. For example a search on gin and vodka sales tends to pick up more online stores in the top 20 than market research reports. This is not necessarily a problem as you may want your results to be biased in this way.

Rather than search each tool one by one, **Dogpile** (www.dogpile.com) combines the results from Google, Yahoo, Ask and MSN and has an option for viewing them side by side with the unique results highlighted.

Think "type of information"

Perhaps the first question we should ask ourselves before starting a search is "should I really be using a general, all purpose search engine?" Like most people I tend to go to Google as my first port of call. Google often does come up with relevant pages but we could do better by thinking about the **type** of information for which we looking:

- need the latest information and discussions on climate change and global warming: think news sources, blogs and RSS feeds
- looking for statistics: go for official government or international sites and spreadsheet formats
- trying to track down information an a small or medium sized company: the national official company registers are your key starting points
- desperate for a quick overview of a subject: try reference sources such as Wikipedia or Answers.com
- and when it comes to quality, peer reviewed articles there are numerous evaluated subject

portals and databases, although many of these will be priced.

For peer reviewed articles many end-users head for Google Scholar. There are problems with some of the search options in Google Scholar, particularly the author search, and it does not include Reed Elsevier who publish a substantial number of journals in the scientific, technical and biomedical area. Reed Elsevier have Scirus, their own web based interface to their journals, and which has far better search and display options than Google Scholar. Microsoft have recently launched Windows Academic Live: this is a direct competitor to Google and is far superior in search and results management but covers a limited subject area at present. All three include free resources but many of the articles are available only on a pay per view basis.

For some searches, it is far better to go straight to the specialist priced databases. They are often quicker in returning relevant results than the general search engines, more focussed, and the provenance and quality of the information is known. Well established services such as STN, GEM, Dialog, Datastar, Proquest, Factiva, LexisNexis and Alacra are familiar to most information professionals but there are many niche databases that are less well known. Tracking these down can be difficult and this is where evaluated subject listings and portals come into their own.

Evaluated subject listings and portals

These are lists of sites covering a particular subject area or type of information, and evaluated and assessed by people knowledgeable in that subject. They often include links to databases not easily found by the general search engines. For example:

- BUBL www.bubl.ac.uk – a good starting point for any subject and type of information. Resources are organised by country, subject, alphabetically, or you can carry out a keyword search. Some searchers looking for commercial or business information are deterred by the academic address but BUBL covers all types of web sites.

- Virtual libraries such as Aerade (aerospace industry), EEVL (engineering), Biome (biomedical), Sosig (social sciences) ad many, many more. For a fuller list go to Pinakes at www.hw.ac.uk/libwww/irn/pinakes/pinakes.html .

- For industry statistics and market information Alacrawiki Spotlights at www.alacrawiki.com is a superb starting point. For each industry Alacra editors have pulled together and commented on web sites, mostly free, that provide quality information. There is a US bias in some areas but more than enough European and International sites to make this a key portal for business information

The right search engine for the right type of information

Consider using the specialist areas of the general search engines. Try out the tabs and links that take you to news, image search, video, audio, or blog searches. Or go to a search tool that covers only one type or format of information.

For blogs and RSS use Google Blogsearch, the RSS/XML file format in Yahoo Advanced Search, Technorati or Blogpulse.

For images click on the images link on the search engine home page but also take a look at the image specific services, in particular those that offer images that are in the public domain:

- Flickr - www.flickr.com – consists of photo collections made available by individuals for sharing within closed groups or the world in general. It can be difficult to search unless you know the tags or keywords that have been assigned to the images by the photographer, so go to www.flickr.com/photos/search/ and use the search box for titles, tags and descriptions. Each photo will have a copyright and licensing statement associated with it indicating what you can or cannot do with it. As a short cut, you can search for photos that have been assigned a specific Creative Commons license at flickr.com/creativecommons. Details of the different licenses are on that page as well.

- Morguefile - www.morguefile.com. This is not a collection of photos of bodies and corpses! Morgue file is a journalist term for a news archive. This site contains free high resolution digital stock photography for corporate or public use.

- Wikimedia Commons - http://commons.wikimedia.org/ . All the images, and other content, on this web site are free to use. Everyone is allowed to copy, use and modify any files as long as the source and the authors are credited, and as long as you release your copies/improvements under the same freedom to others.

Looking for TV and radio interviews, corporate promotions, presentations on video, or monitoring competitor's advertisements? Yahoo Video and Audio are excellent, if slightly US biased. Blinkx.tv searches many of the standard news sources such as radio, the BBC and CNN, but also has links to advertising databases. Visit4info.com covers European advertising on TV, the cinema and in the press. You can search by company, brand/product name, keyword, media and date and view the advertisements. The service is free for the adverts broadcast over the last three months; earlier adverts can be viewed for an annual fee of GBP 45.

And so the list goes on.....

Identifying specialist search tools

It is impossible to remember all of the search tools that one should or could be using. Luckily there are several services that can help you with this. Two of my personal favourites are GoshMe and Trovando.it.

GoshMe www.goshme.com

GoshMe's home page has a number of major headings such as science, environment and society, audio, health, law, reference, sports. Select the most appropriate subject or subjects and type in your search. GoshMe looks at the results from different search engines – it claims to cover approximately 1,000 - and then displays what it thinks are the best tools in which to run the search. You can then choose the tools you want to use.

Trovando www.trovando.it

This is an interface to hundreds of search tools covering a wide range of formats and information types. You type in your strategy just once, click on the type of information (for example web, image, blog, video/audio or reference) and then choose your service (for example Reference includes Wikipedia, Google Scholar, FindArticles, Scirus, Encarta). This is a very quick way of running your search sequentially in a whole range of tools. I find that it provides an excellent reminder of the resources I should be considering.

Keeping up to date

Keep up with what Google is doing on labs.google.com and googlepress.blogspot.com

I report on new resources and search features in my own blog at www.rba.co.uk/rss/blog.htm and in my free newsletter Tales from the Terminal Room (www.rba.co.uk/tfttr/).

For my own current awareness I use a mix of newsletters and RSS feeds, but the top two for me are:

- Internet Resources Newsletter www.hw.ac.uk/libWWW/irn/irn.html
  A monthly newsletter available on the Web and via email highlighting key sources of information and new search tools.

- Phil Bradley's Blog http://philbradley.typepad.com/
  My first port of call for the more unusual search tools and new search engine features.

Summary

I hope that this paper has demystified the hidden web and given you some indication on how to tackle it. For myself, I find it helps to remember 5 main points:

1. Use the search engine advanced search features to bring relevant documents further up your results lists

2. Google is not the only search tool – try a different one

3. Use the media-type areas of the search engine such as images, video, blog or try a specialist tool

4. Try an evaluated subject listing

5. Think “type of information” - that will guide you to the most appropriate resource

URLs

Mainstream alternatives to Google

Yahoo – search.yahoo.com
Ask (Jeeves) – www.ask.com
Exalead – www.exalead.com
Alltheweb Livesearch – livesearch.alltheweb.com
MSN – search.msn.com

Comparing search engines and identifying specialist tools

Thumbshots Ranking – ranking.thumbshots.com
Dogpile – www.dogpile.com
GoshME – www.goshme.com
Trovando – www.trovando.it

Evaluated subject listings

BUBL – www.bubl.ac.uk
Pinakes – www.hw.ac.uk/libWWW/irn/pinakes/pinakes.html
Alacrawiki – www.alacrawiki.com

Reference sources, databases and peer reviewed articles

Google Books – books.google.com
Official National Statistics World-wide www.library.auckland.ac.nz/subjects/stats/offstats/
Nationmaster – www.nationmaster.com
Official Company Registries – www.rba.co.uk/sources/registers.htm
Scirus – www.scirus.com
Google Scholar – scholar.google.com
Windows Academic Live – academic.live.com

Blogs

Google Blogsearch – www.google.com/blogsearch/
Technorati – www.technorati.com
Blogpulse – www.blogpulse.com

Images

MorgueFile – www.morguefile.com
FreeFoto – www.freefoto.com
Flickr Creative Commons – www.flickr.com/creeativecommons
Wikimedia Common – commons.wikimedia.org

Video

Yahoo Video – search.yahoo.com and click on the Video optio
Blinkx TV – www.blinkx.tv
Visit4Info – www.visit4info.com