Tales from the Terminal Room
July/August 2005, Issue No. 64

Editor: Karen Blakeman
Published by: RBA Information Services
ISSN 1467-338X

http://www.rba.co.uk/tfttr/

Tales from the Terminal Room (TFTTR) is a monthly newsletter, with the exception of July and August which are published as a single issue. TFTTR includes reviews and comparisons of information sources; updates to the RBA Web site Business Sources and other useful resources; dealing with technical and access problems on the Net; and news of RBA's training courses and publications.

Tales from the Terminal Room can be delivered via email as plain text or as a PDF with active links. You can join the distribution list by going to http://www.rba.co.uk/tfttr/ and filling in the form. You will be sent an email asking you to confirm that you want to be added to the list. TFTTR is also available as an RSS feed. The URL for the feed is http://www.rba.co.uk/rss/tfttr.xml . Further information on RSS feeds can be found at http://www.rba.co.uk/rss/rss.htm and Wikipedia has a good article on the topic at http://en.wikipedia.org/wiki/RSS_%28file_format%29 .

In this issue:

Search Tools

Yahoo database overtakes Google

Yahoo now claims to index 20.8 billion web documents and images overtaking and nearly doubling Google's total of 11.3 items. The 20.8 billion breaks down into 19.2 billion 'documents' and 1.6 billion images. Google's 11.3 comprises 8.2 billion web pages, 2.1 billion images with the remainder coming from groups and discussions.

For many of us the size of the database, beyond a certain point, is irrelevant. In any case, you cannot view more than the first thousand or so of the results. What is more important is the quality of the results. 20.8 billion sounds impressive, but many of those could be duplicates that have multiple domain names pointing at them. A study comparing the size of the Yahoo and Google databases (http://vburton.ncsa.uiuc.edu/indexsize.html) casts doubt on Yahoo's claims.

Their main conclusions were:

> "Based on the data created from our sample searches, this study concludes that for a random set of words a user can expect, on average, to receive 65% more results using the Google search engine than the Yahoo! search engine. In fact, in the 10,034 test cases we ran, only in 16% of the cases (1606) did Yahoo! return more results. In 83.7% of the cases (8399) Google returned more results. In less than 1% of the cases both search engines returned the same number of results."

My own experience varies but for many 'ordinary' key word searches, that is searches that do not limit by filetype, domain etc Yahoo generally came out on top with more results. Only when I included file types and domains did this change. Below are some examples (G represents Google and Y Yahoo).

Test searches

peak oil	G: 5,050,000	Y: 12,700,000

Change that by enclosing the phrase in double quote marks:

"peak oil"	G: 666,000	Y: 1,950,000

Add in Hubbert:

"peak oil" Hubbert	G: 47,600	Y: 161,000

So that seems to confirm Yahoo's claim of a larger database. But my next search seemed to suggest otherwise:

vehicle emissions air quality	G: 4,210,000	Y: 3,090,000

Using double quote marks, though:

"vehicle emissions" "air quality"	G: 25,900	Y: 636,000

What happened here was that in the search without the double quote marks Google was automatically stemming my terms and searching on variations. As soon as I used double quote marks Google looked for exact matches for the phrases. Forcing Google to carry out an exact match search for individual words by preceding each term with a plus sign gave:

+vehicle +emissions +air +quality	G: 1,360,000	Y: 3,090,000

So, yes, Yahoo **does** seem to have a larger database. Still with me? My next step was to limit the search by filetype PDF, which gave:

G: 12,000 Y: 3,660

Google wins. Further limiting the search to UK government sites gave:

G: 5,910 Y: 1,810

Google wins again! I tried another search using a similar approach:

climate change emissions trading	G: 1,950,000	Y: 2,290,000
+climate +change +emissions +trading	G: 612,000	Y: 2, 290,000
"climate change" "emissions trading"	G: 372,000	Y: 1,630,000

Restricting my search to ppt as a filetype:

G: 485 Y:1,730

This time Yahoo wins when it comes to filetype limits. Limiting further to US government sites:

G: 11 Y: 11

So the same and, of these, each had 3 unique results.

Only one more to go - my usual gin/vodka search!

gin vodka sales UK	G: 34,800	Y: 61,600
Limit to filetype PDF	G: 587	Y: 695
Limit to filetype DOC	G: 30	Y: 19
(Google had 19 unique documents and Yahoo 8)		
Limit to filetype XLS	G: 2	Y: 2 (All unique documents).

Conclusions

From my own searches, both those shown above and day to day enquiries,Yahoo does indeed seem to have a larger database when it comes to 'basic' keyword searches. For some strategies Google comes up with more results, but this seems to be because it is automatically stemming the search terms. If an exact match is forced then Yahoo comes out on top in terms of numbers. The study mentioned earlier does not say if this factor was taken into consideration in their tests

Using advanced search techniques, such as limiting by filetype or domain, muddied the waters somewhat. Sometimes Google had more, sometimes Yahoo but each does often come up with unique results. This, combined with the fact that Yahoo and Google rank and sort search results in different ways, confirms that one really should use both for a more thorough search. Also, one should not forget that other search tools have unique search features that can bring up very different sets of results, for example Exalead (http://www.exalead.com/) .

Google adds RSS and Atom to news alerts

At long last Google has added Atom and RSS feed options to their news alerts service. Until now we have had to either receive them by email or use third party services such as ScrappyGoo. There were some teething problems during the first couple of days when some some RSS readers reported errors with the feeds, but it seems to be working fine now.

Google makes proximity searching fuzzier

For some time Google has had a proximity operator of sorts. You could use a single asterisk or multiple asterisks between words within a phrase to specify how far apart or close together you want the words to be. For example the search:

"phenol * ether" extraction

looks for documents where phenol and ether are separated by one word and extraction is anywhere in the page.

"phenol * * ether" extraction

looks for documents where phenol and ether are separated by two words, and so on.

That still works, but now you can use the asterisk between two words **without** the quotes marks and Google looks for the terms separated by one or more words.
For example:

phenol * ether extraction

I cannot find any information on the Google site that states the maximum number of words the asterisk can represent and I have not yet had time to site down and test it out. Any volunteers?

Google launches Desktop Search 2 plus Sidebar

http://desktop.google.com/

Google Desktop 2 has been launched with a new sidebar feature. I am not a great fan of Google Desktop for a variety of reasons - I actually use Yahoo Desktop - but decided to give this latest version a try.

The good news is that indexing of secure web pages such as bank statements and of password protected documents is switched off by default. You can even encrypt the cache that Google creates on your PC to protect it from prying eyes. The sidebar also looked promising with a news panel, options for RSS feeds, a scratchpad, share price monitor, weather, and a panel where you can have a sort of slide show of your favourite photos.

Now for the bad news. My enthusiasm started to wane as I discovered that the share prices and weather are US only, and that the news is from Google.com. I would have preferred news.google.co.uk but there seems to be no way of changing this. There is some UK content that you can specify, for example BBC News and you can train it by telling it not to show any more articles "like this". And if you have the advanced features switched on it is supposed to be able to work out the type of content you prefer to read. It didn't seem to work for me and, after three days, I gave up and removed the news panel as well as the Web Clips/RSS. (I find it much easier to use a proper RSS reader for feeds). I also found the Quick View of recently viewed files and documents irritating.

I expected the Sidebar email panel to be restricted to Outlook and Gmail but it did list new emails that appeared in my Thunderbird inbox. However I wanted it to alert me to just my Gmail messages so that I did not have to check my mail by logging in via my browser. The only way I could find to stop it picking up my Thunderbird messages was by creating filters. Far too time consuming and tedious, so that panel went as well. Which left me with the scratchpad (actually, quite useful), my photo slide show and lots of empty space.

I then filled up some of the space with an Adsense plugin to display how much my web pages are earning, and a to-do list. There is a list of plug-ins at
http://desktop.google.com/plugins/ .

As far as searching my hard drive, it still lags way behind Yahoo Desktop in terms of accuracy, number of documents found, and indexing procedure. The main problem though was that I found it to be a serious

resource hog in terms of CPU usage. It was mainly related to when it decided to index my email or leap into action to index new and modified documents. Even when I told it to stop it took a while before it obeyed and then about 10-15 minutes later off it would go again. I can't see any way of controlling when it indexes documents, unlike Yahoo Desktop and many of the other desktop tools.

I have now uninstalled it :-(A pity, because I rather like the sidebar. It has real potential, and I am sure that there will be plenty more useful panel plug-ins along soon.

Trovando - search different

http://www.trovando.it/

I picked this one up from Phil Bradley's blog. Think of it as Turboscout with lots more search engines and lots more types of search. The search tools are organised under tabs such as Web, Images, Reference, News, Blogs etc. Click on a tab, enter your terms and click on each tool in turn to run your search. I particularly like the URL tab which, amongst other things, finds pages that link to your known URL (backlinks), runs a Whois on the domain name, and finds archived copies of the page. And there is a custom tab where you can build your own collection of search tools.

Gigablast Toolbar

http://www.gigablast.com/

Gigablast is the latest search engine to launch a toolbar for your browser. At first glance it seems rather Spartan when compared with other toolbars such as Yahoo's and Google's, but it does have two unique and very useful features: search the sites linked from the current page and search the sites in your bookmarks.

Unfortunately it is only available for IE at present. There is a plugin for Firefox users at http://mycroft.mozdev.org/download.html?name=gigablast&submitform=Find+search+plugins or http://tinyurl.com/dyopr , but this only adds Gigablast as an option to the built in search box.

Dogpile Adds MSN

http://www.dogpile.com/

Meta search tool Dogpile had added MSN to its collection of search engine. It now searches Google, Yahoo, Ask Jeeves and MSN at the same time. (You can compare three search tools at a time using Dogpile's new Search Comparison (http://comparesearchengines.dogpile.com/) and see how much overlap there is between them in the first 20 results.)

When you run a web search, as well as combining and removing the duplicates from the results Dogpile automatically shows you the Top 12 from Google and Yahoo side by side on the right hand side of the screen. Unique results are highlighted. You can close these boxes or add columns for MSN and Ask Jeeves by clicking on the relevant icons. In addition to web meta search Dogpile has an images search (Yahoo and Ditto), audio (Yahoo and Singingfish), video (Yahoo and Singingfish) and a News option (Yahoo, Topix News, Fox News, ABC News). Interestingly, Google and MSN are omitted from the images and news meta search. The Yellow and White Pages search covers the US only.

Information Resources

Alacra Store

http://www.alacrastore.com/

As one who is a great supporter of pay as you go services, I was delighted to see that Alacra have launched a pay-per-view version of their priced service. The Alacra Store enables "business professionals", or anyone else for that matter, to find and purchase premium business information with a credit or debit card. The types of content accessible include company fundamentals and financials, credit research, economic data, market and investment research, and news.

Still in beta, it covers just 30 of the 100 or more databases available via the subscription service and many highly desirable databases, such as EIU, are not included. I have been reassured by Alacra, though, that the EIU files will be available: they are on the long list of other sources that they are working on

integrating for the Store.

Looking at the company information my first thoughts were that any half competent researcher should be able to find much of the Alacra content free on the net, at least for publicly traded companies. Having had a closer look there is some great additional stuff which one would have to go to one or more priced services to locate, and for many of those you would have to take out substantial subscriptions. Also, it takes time to pull information together from the free web and evaluate it, and we don't always have that luxury. Many's the time in the past I have been asked to produce a company snapshot - 2 -3 pages max - for my MD who was about to step into a taxi for a meeting with his opposite number at that company. The Alacra snapshots are perfect for that.

The market research, like the company information, comes from a variety of sources and pricing depends on the publisher, report series and length of report. For some you can purchase individual pages or sections, but many of the documents I retrieved with my searches were only available as the full report. This might be the policy of the publishers of the documents that I just happened to retrieve with my searches. I did not cross check with the publishers' web sites or other market research aggregators to confirm whether or not this was the case.

Alacra has some good market research sources and - Oh Joy! - they have Tablebase, one of my favourite databases. It is the quickest way I know of tracking down rankings and market shares in a particular industry/country. I have tried accessing it via Dialog's Open Access and Skyminder but have always had mixed and sometimes rather odd results. Dialog, in particular, I find clunky and slow. In contrast, the Alacra Store interface is much 'smoother' and faster, and I seem to get far more sensible results.

When it come to news, though, Alacra will not be my first port of call. The articles from Business and Industry are a bit pricey ($ 9.95 regardless of the length of the article) compared with the per article pricing of LexisNexis ($3), a comment made by several other researchers.

Alacra admits the cost of the news articles is relatively high. They are focusing more on the premium content (i.e. company profiles, market research, investment research, etc.) that is not available anywhere for free. The reason they have the archival news stories, expensive though they are, is so that they can offer a "one-stop shop" for business information. Having said that B&I will not be my primary news source, I would add it to my list after the cheaper services if I required as comprehensive a search as possible, just in case B&I had industry publications not covered elsewhere.

The purchasing process is straightforward once you have registered, and the usual cards are supported. You may get a shock, though, when you see the final price. The prices quoted for each document in the Alacra Store do not include VAT; this is added on at the check out. That doesn't bother me as I am VAT registered and I can claim it back, and I can understand why it has been set up this way. However, I do think it would be helpful if the fact that prices do not include VAT or local sales tax were mentioned before the final purchase screen. It does make one wonder if there are other hidden costs; there aren't in this case but that is not the point. VAT on $9.95 is not much in terms of £s and pence, but it is a significant wad of dosh on top of, say, $250.

Overall, I like this service and will use it. It is still in beta and I found Alacra to be very responsive to comments and suggestions. Definitely recommended.

Automotive Industry Information Sources

http://www.rba.co.uk/sources/automotive.shtml

I have just added a page listing some key resources in the automotive industry to the section on Industry Sectors on my web site. Once again, many thanks to Paul Pedley who supplied me a with a list of web sites to help me build up this section.

Aroq industry specific market research

http://www.aroq.com/

This is a gateway to Aroq's four industry specific sites: just-drinks.com, just-food.com, just-auto.com and just-style.com (fashion). All four pull together news and market research in their particular sector. The market research comes from a wide range of suppliers and can be purchased online. The news is free and each of the sites have blogs commenting on events in their sector. RSS feeds of the news headlines and the blogs are available.

Thomas Register becomes ThomasNet

Thomas Register has changed its name to ThomasNet and has a new URL - http://www.thomsanet.com/ . ThomasNet is a database of over 650,000 US industrial suppliers that can be searched by product, company name or brand name. If you register (free of charge) you can save searches, company details, subscribe to newsletters and alerts, and customize news feeds. You can receive newsletters by email or select a category from the Industrial News Room and receive the latest news as an RSS feed or add it to your My Yahoo.

For more International coverage there is the Thomas Global Register (http://www.tgrnet.com/). This is a directory of 700,000 manufacturers and distributors from 28 countries, classified by 11,000 products and services categories. Search by product/service, company name or browse the categories. Both ThomasNet and the Thomas Global Register are free of charge.

If you are only interested in European suppliers then try the Thomas Global Register Europe (http://www.tremnet.com/), which covers 210,000 industrial suppliers from 21 European countries. Search by company name or product/service. The information is free but registration is required.

Our Property launches alerts

http://www.ourproperty.co.uk/

Our Property repackages Land Registry data and data from the Registrars of Scotland. It enables you to search for properties that have been sold since 2000 by street, town and postcode and displays the price that the property or properties were sold for. They now offer free sales alerts. You can monitor new sales at up to 10 different locations around the UK. Enter the postcodes that you are interested in, and if a new sale is registered in the database within 500 metres of that postcode you receive an email alert.

Very useful if you are buying or selling a house and want to compare prices, or if you are just plain nosey and want to know how much your neighbour paid for their property!

Searching Questions

Are the English versions of foreign language pages identical in content to the original?

Question:

This may seem a rather strange question, but we went to Prague Zoo while we were on holiday and were very impressed by their two large hippos Slavek and Maruska. While we were out there, a Czech friend showed us some photos of them on the Prague Zoo web site (http://www.zoopraha.cz/). Back home I can't seem to find them. In fact the site seems to have far less information than I remember. I am wondering if the information on the English version of the site is different form the original Czech version?

Answer:

Not a strange question at all. Not many people realise that the non-English language web sites sometimes only have translations of basic information. The Prague Zoo web site is a good example. So how to find the photos without knowing any Czech. As it is not obvious where the search option is on the Czech pages, but at least we know the names of the animals. I went to Google Advanced Search, typed Slavek Maruska in the top search box and then zoopraha.cz in the domain box. The entry at the top of my results list had 9 photos of the hippos.

It is important to be aware that translated pages may not always be exact translations of the original language. For example, the official Norwegian companies register at http://www.brreg.no/ gives only key points about the service in English. To search the database, you have to use the Norwegian

pages. In these circumstances, I usually end up clicking on every link until I find a page containing something that resembles a database search box.

An extreme example is Wikipedia (http://www.wikipedia.org/). You may have noticed that the home page has links to several language 'versions'. The articles on these sites are not always translations at all They are often totally different articles written by completely different people. This is not necessarily a disadvantage as they may present different information and have a different emphasis. If your linguistic skills are up to it, they can be used to complement each other. I find this particularly when looking for biographies, for example the English and French articles on Charles de Gaulle.

These things are sent to try us!

Google's My Search History - did you know you had one?

Sorry, but this is another dig at Google. The potential privacy/security threat posed by Google's Desktop cache, and which I mentioned in the previous issue of TFTTR, has been fixed to a large degree. The latest Google Desktop (see earlier in this issue of TFTTR) has the caching of secure web pages and password protected documents switched off by default. Also, you can now encrypt the plain text cache that Google creates on your PC. This time it is My Search History that is under the spotlight.

My Search History, to be found under under Google Labs (http://labs.google.com/), keeps track of all the searches that you do on Google and the results that you click on. They are ordered in reverse chronological order with the most recent first.

One advantage is that you can easily go back to previous searches and double check which pages you used for data and information. Another is that it is also tied into Personalized Search, which orders your search results based on "what's most relevant to you." Personalized Search attempts to learn what interests you by analysing your searches and the pages that you visit.

Personalized Search may or may not work. I did not notice a significant difference but I may not have used it for long enough. The My Search History is more worrying. Of course you do not search for or visit "dodgy" web sites (but what is dodgy?). But do you really want to keep a record of every piece of confidential research? Of course you do not have My Search History switched on, do you. Actually, you might, if you have a Google account for news alerts or email. After you have finished with your news alerts or your Gmail do you log out before carrying out any searches? If you don't, then go to http://labs.google.com/, click on My Search History and log in.

If, to your horror, the gory details of your chequered past is revealed all is not lost. You can remove individual searches or all of them. If you do not want My Search History to record your research click on the Pause button to halt the eavesdropping.

What seriously annoys me about this is that My Search History is, or certainly used to be, switched on by default.

Gizmo of the Month

Customise Google Firefox Extension

https://addons.mozilla.org/extensions/

CustomizeGoogle is a Firefox/Mozilla extension that allows you to add or remove elements from Google results pages.

Options include:
- use Google Suggest when you search
- add links to other search tools from the Google results page - tools vary according to type whether you are using web, image, news search etc,
- add position counters to the search results
- add links to the WayBack Machine (webpage history)
- filter out specific web sites or pages - you can either do this manually or use the filter link next to each entry in the results list
- anonymize your Google userid

Meetings and Workshops

Workshop: Untangling your web: effective web site management
Organiser: Manchester Business School
Presenter: Karen Blakeman
Venue: MBS, Manchester
Date: Wednesday, 14th September 2005
Course fee: £215 + VAT - BIS/BINN members. Others £250 + VAT.
URL: http://www.mbs.ac.uk/bis-training
Course Outline: Whether you already have a web site or are in the process of setting one up, keeping it in pristine condition and ensuring that your organisation benefits to the full can drive you to distraction. Once it is operational it is all to easy too easy to forget about it; but out of date pages and inaccurate information are bad news! This course looks at how to painlessly manage the process of keeping your web site up to date, how to identify key areas that can drive your business forward, and how to encourage visitors to come back for more.

Workshop: Business Information on the Internet: Free vs. Fee
Organiser: TFPL
Presenter: Karen Blakeman
Venue: TFPL, London
Date: Thursday, 6th October 2005
Course fee: £295.00 + VAT.
URL: http://www.tfpl.com/
Course Outline: To pay or not to pay for business information is a question that faces many users of the Net. This one day course compares what is available for free with pay-as-you-go and subscription services and will examine quality of content, coverage, functionality and price. There will be practical sessions throughout the day and delegates will receive comprehensive notes.

Workshop: Searching the Internet: Google and Beyond
Organiser: UKeiG
Presenter: Karen Blakeman
Venue: CILIP, 7 Ridgmount Street, London WC1E 7AE
Date: Friday, 14th October 2005, 9.30 - 16.30

This course is now fully booked but it is being re-run. Please contact cabaker@ukeig.org.uk for further information and dates.

Contact Information

Karen Blakeman, RBA Information Services
UK Tel: 0118 947 2256
Int. Tel: +44 118 947 2256

UK Fax: 020 8020 0253
Int. Fax: +44 20 8020 0253

Address: 88 Star Road, Caversham, Berks RG4 5BE, UK

Archives

TFTTR archives:
http://www.rba.co.uk/tfttr/archives/

Subscribe and Unsubscribe

To subscribe to the newsletter fill in the online registration form at http://www.rba.co.uk/tfttr/

To unsubscribe, use the registration form at http://www.rba.co.uk/tfttr/ and check the unsubscribe radio button.

If you wish to change your subscription from the PDF to the Plain Text version, send an email to tfttr@rba.co.uk with "Plain Text Version Please" in the subject line.

Privacy Statement

Subscribers' details are used only to enable distribution of the newsletter **Tales from the Terminal Room**. The subscriber list is not used for any other purpose, nor will it be disclosed by RBA Information Services or made available in any form to any other individual, organisation or company.