

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

---

## Yahoo Hacks

---

While Google hacks—tips, tricks, techniques, and scripts that make Google more powerful and useful—are plentiful and fairly well documented, the same cannot be said (yet) for Yahoo Hacks, despite the fact that O'Reilly published a *Yahoo Hacks* book in late 2005. Part of the reason for this was the absence of Yahoo APIs, a problem Yahoo recognized and rectified with its Developer site.

Yahoo Developer Network <http://developer.yahoo.net/>

Yahoo Developer Network Blog <http://developer.yahoo.net/blog/>

While many of the hacks, mostly employing some form of API, are geared toward maps, Yahoo launched a webpage devoted exclusively to Yahoo and “mixed” API applications.

Yahoo Search Application Gallery <http://developer.yahoo.net/search/applications.html>

I recommend you pay special attention to the following applications that use Yahoo APIs, although you may find others even more useful to you:

Link Harvester <http://www.linkhounds.com/link-harvester/>

This is a very powerful—but very slow—tool for examining links to a domain or a specific url. The example below shows the links to [www.mfa.gov.cn]. Link Harvester does the following:

- quickly finds almost every single site linking into a domain or page.
- scrapes past the 1,000 search result limit by making domain filtering a snap.
- grabs number of pages indexed.
- grabs links to any page.
- grabs total inbound links, home page links, and deep link ratio.
- tool is fast and free. which is great considering all it does.
- grabs C block IP address information.
- tool provides links to Wayback Machine and Whois Source (now Domain Tools) next to each domain.
- free & open source

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//FOR OFFICIAL USE ONLY

- uses the Yahoo API so it complies with their TOS [terms of service].<sup>51</sup>

URL (ex. www.site.com):  Link Type:  Query  
 Depth:  Start Over

http://api.search.yahoo.com/WebSearchService/V1/webSearch?query=linkdomain:www.mfa.gov.cn

Enter sitenames you want to filter:

Showing 201 unique domains from the first 250 results of 273 total results

Links To Domain: 610 Pages Indexed: 121  
 Links To Homepage: 255 Deep Link Percentage: 58%

7 Unique Educational Domains (.edu) with 7 Unique C Block Addresses

<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">job.svau.edu.cn</a> (2) 210.47.174.208	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">job.hztc.edu.cn</a> (2) 221.12.26.151
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">cs.whu.edu.cn</a> (2) 202.114.121.41	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">career.ruc.edu.cn</a> (2) 202.112.117.116
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.htsz.edu.cn</a> (2) 218.17.227.219	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">job.hliu.edu.cn</a> (2) 210.46.96.35
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.sdnqy.edu.cn</a> (2) 211.64.116.10	

12 Unique Government Domains (.gov, .mil) with 10 Unique C Block Addresses

<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.mfa.gov.cn</a> (4) 211.99.196.166	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.gov.cn</a> (2) 202.123.110.3
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">embassy.tajikistan.fmprc.gov.cn</a> (2) 211.99.196.218	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">wqzc.ywlu.gov.cn</a> (2) 61.153.32.13
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.scpta.gov.cn</a> (4) 61.157.75.21	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.changchun.gov.cn</a> (2) 221.8.13.136
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.zjgthz.gov.cn</a> (2) 218.4.101.3	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">portal.praefectura.sp.gov.br</a> (2) 200.230.190.68
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">qwy2006.mop.gov.cn</a> (2) 202.106.181.242	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">wcm.fmprc.gov.cn</a> (2) 211.99.196.166
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">bsq.sh.uov.cn</a> (2) 218.242.255.118	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.zjcx.gov.cn</a> (2) 218.75.53.69

182 Unique Commerical Domains (.com, .net, etc) with 126 Unique C Block Addresses

<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.atimes.com</a> (2) 204.14.134.23	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.quoshi.net</a> (2) 203.194.128.198
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.chinaliss.org</a> (4) 210.51.190.236	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">bubblepucker.li.net</a> (4) 211.100.24.5
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.titaaq.com</a> (2) 203.88.198.16	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.freerepublic.com</a> (2) 209.157.64.201
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.comefromchina.com</a> (2) 67.15.83.143	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">spaces.msn.com</a> (2) 65.54.153.254
<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">www.popyard.org</a> (2) 72.4.161.148	<input type="checkbox"/> [W] [A] [G] [T] [H] [D] [Y] <a href="#">zh.wikipedia.org</a> (4) 207.142.131.213

For each unique domain, Link Harvester provides [W]=Whois Source data for domain; [A]=Internet Archive data for domain; [G]=Google cache of actual webpage; [T]=Google's text only cache of actual webpage; [H]=Google's text only cache of domain; [D]=Whois Source's information about the domain from the Open Directory; [Y]=Yahoo's Directory Listing of Whois Source data about the domain.

Hub Finder

<http://www.linkhounds.com/hub-finder/>

"Hub Finder looks for sites which have co-occurring links to related authoritative websites on a particular topic." Basically, Hub Finder locates authoritative websites on a particular subject, as in the example below, for *java*. In this case, the top sites (most authoritative resources) for *java* are shown. Hub Finder also permits users to download the data in CSV (Comma Separated Value) format that can be easily merged into a spreadsheet or database.

<sup>51</sup> Link Harvester, *Linkhounds*, <<http://www.linkhounds.com/link-harvester/>> (14 November 2006).

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

Subject:  Results:    Min Match:  Depth:

Google Key:

Include this site (optional):  Link Type:  Application:  Query:

Enter up to 10 sites:

Querying the following 5 sites

- 1: http://www.java.com
- 2: java.sun.com
- 3: www.java.com
- 4: rdw1.yahoo.com
- 5: javaboulique.internet.com

Showing 45 sites with at least 2 matching backlinks from 178 search results

1	2	3	4	5	Site Name			
X	X	X	<a href="#">W</a>	<a href="#">A</a>	<a href="#">H</a>	<a href="#">D</a>	<a href="#">Y</a>	<a href="#">blogs.sun.com</a> (209.249.116.203)
X	X	X	<a href="#">W</a>	<a href="#">A</a>	<a href="#">H</a>	<a href="#">D</a>	<a href="#">Y</a>	<a href="#">java.sun.com</a> (209.249.116.141)
X	X	X	<a href="#">W</a>	<a href="#">A</a>	<a href="#">H</a>	<a href="#">D</a>	<a href="#">Y</a>	<a href="#">www.jcp.org</a> (192.18.97.62)
X	X	X	<a href="#">W</a>	<a href="#">A</a>	<a href="#">H</a>	<a href="#">D</a>	<a href="#">Y</a>	<a href="#">www.microsoft.com</a> (207.46.18.30)
X	X	X	<a href="#">W</a>	<a href="#">A</a>	<a href="#">H</a>	<a href="#">D</a>	<a href="#">Y</a>	<a href="#">www.sun.com</a> (209.249.116.195)
X	X	X	<a href="#">W</a>	<a href="#">A</a>	<a href="#">H</a>	<a href="#">D</a>	<a href="#">Y</a>	<a href="#">www.talkcity.com</a> (66.37.219.37)
X	X	X	<a href="#">W</a>	<a href="#">A</a>	<a href="#">H</a>	<a href="#">D</a>	<a href="#">Y</a>	<a href="#">atlantis.bigfishgames.com</a> (63.251.168.82)
X	X	X	<a href="#">W</a>	<a href="#">A</a>	<a href="#">H</a>	<a href="#">D</a>	<a href="#">Y</a>	<a href="#">camelot.stratics.com</a> (64.158.108.35)
X	X	X	<a href="#">W</a>	<a href="#">A</a>	<a href="#">H</a>	<a href="#">D</a>	<a href="#">Y</a>	<a href="#">dessert.net</a> (69.60.119.225)
X	X	X	<a href="#">W</a>	<a href="#">A</a>	<a href="#">H</a>	<a href="#">D</a>	<a href="#">Y</a>	<a href="#">...</a>

The following Yahoo Hacks generally mirror certain Google hacks, with the exception of the **originurlextension:** syntax, which is unique to Yahoo and very powerful.

- Disabling Word Stemming. Yahoo does not give users the option to turn off word stemming, which can frustrate users trying to perform precise searches. To run a precise search, enclose the term in double-quotes, e.g., ["drink"] will not find *drinks* (except in sponsored results).
- Searching by Filetype. Despite the fact Yahoo mysteriously disabled its *filetype* syntax, you can use **originurlextension:** to search by file type, but this syntax is imperfect.

Examples of how to use the **originurlextension:** command:

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

- [originurlextension:pdf "white paper"] finds pages indexed by Yahoo that are in PDF format and contain the phrase ["white paper"] anywhere in the text, title, or url.

To search by specific type of file, use the syntax *originurlextension:* plus one of these or **any file extension**, such as *cgi*, *log*, *zip*, etc. Because this workaround is not a true filetype search, you can search on any file extension.

- htm or html—standard webpage
- pdf—Adobe Acrobat
- xls—MS Excel
- ppt—MS PowerPoint
- doc—MS Word
- txt—text
- xml, rdf, rss—RSS or XML feeds<sup>52</sup>

Searchroller. Searchroller uses a JavaScript to let you create a neat little search query bookmarklet<sup>53</sup> for your future use. The bookmarklet comprises a set of domains you like to search on routinely but don't want to type in each time. For example, perhaps you'd like to search simultaneously on a whole group of news sites. Tara Calishain's script lets you input the urls for the news' sites once, then save them to your Favorites or Bookmarks. Each time you click on the bookmarklet, a screen will appear asking you to enter a query term or terms, then the bookmarklet will automatically go to Yahoo and run that query against all the urls you have previously selected. It's a great timesaver when you consider this is a typical Searchroller bookmarklet query, although it could be much longer:

[iraq (site:cnn.com OR site:msnbc.com OR site:usatoday.com) OR  
[site:nytimes.com OR site:washingtonpost.com OR site:bbc.co.uk )]

Searchroller

[http://www.researchbuzz.org/2004/10/new\\_yahoo\\_hack\\_searchroller\\_fo.shtml](http://www.researchbuzz.org/2004/10/new_yahoo_hack_searchroller_fo.shtml)

Artificial Proximity Search. Since Yahoo's APIs are so new and as yet not fully exploited, clever folks like Tara Calishain have come up with ways to force Yahoo to perform new types of searches. The proximity search lets you input one search term and look for it from 1 to 5 "spaces" (really, words) from a second search term. For example, I can search for *henry* within two words of *thoreau* and find many instances

---

<sup>52</sup> In order to read RSS or XML feeds, you need a reader or aggregator to parse this type of data.

<sup>53</sup> A bookmarklet is a tiny JavaScript application contained in a bookmark that can be saved and used the same way you use normal bookmarks. Bookmarklets do not require users to download and install software. For more on bookmarklets, visit <http://www.bookmarklets.com/>.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

of *Henry David Thoreau*. This tool is very good for finding names with the last name listed first, e.g., *Thoreau, Henry David*.

---

### YNAPS -- Yahoo Non-API Proximity Search

Try using an artificial NEAR search for Yahoo:

Find Word One:

Within  spaces of

Word Two:

Any additional words?

### Yahoo Proximity Search

[http://www.researchbuzz.org/2004/10/ynaps\\_yahoo\\_nonapi\\_proximity\\_s.shtml](http://www.researchbuzz.org/2004/10/ynaps_yahoo_nonapi_proximity_s.shtml)

Boilerplate Words or Phrases Yield Gold. Used in combination with keywords, standardized words or phrases can produce very useful results from Yahoo as well as Google. Whether it's "company proprietary," "not for distribution," or a copyright disclaimer, these are the kinds of identifying query terms that searchers need to look for.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

---

## Windows Live Search

---

MSN Search is no more. As of mid-September 2006, Windows Live Search was out of beta and officially supplanted MSN Search. It came as a surprise to no one that the new Live Search has the familiar clean, uncluttered look popularized by Google. Live marks a clear change in Microsoft's overall direction from a multipurpose portal to a search service: "Live.com is now first and foremost a search destination," according to Christopher Payne, Microsoft's corporate vice president.<sup>54</sup>

The question on everyone's mind is whether or not Live Search is any better than MSN Search or Google or Yahoo or any number of other search engines. Thus far, Live is not noticeably superior to MSN Search, but it is a one of the top three largest and most powerful US-based search engines.

The new Windows Live Search:

- uses its own database for web search.
- indexes at least 5 billion pages.
- offers cached links with the date Microsoft estimates the page was last updated (usually the date the Microsoft spider last crawled the page); sometimes a date will appear next to the cached link on the results' page if that page has recently been updated.
- has a "Near Me" search option that only works in the US; it uses your IP address to determine your location; users can override this location by changing it on the Options page. Note that you cannot leave the default location empty. ***If you do not enter a location, Live Search will default to what it reads as your IP address's geolocation.***
- offers web, news, image, local, Q&A, academic, feeds, video, products, and new "build your own" searches.
- offers preference control via "options."

The "Search Builder" query customization tool has been replaced by the "Advanced" option; as with "Search Builder" the Advanced option opens a little window beneath the search form.

---

<sup>54</sup> Chris Sherman, "Microsoft Upgrades Live Search Offerings," SearchEngineWatch, 12 September 2006, <<http://searchenginewatch.com/showPage.html?page=3623401>> (5 October 2006).

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

## **Customizing Live Search Settings (“Options”)**

Live Search currently offers these user-defined options (preferences):

- Display: display the site in a specific language (most major languages with some notable exceptions, e.g., Arabic, Thai).
- Number of Results: choose to display 10, 15, 30, or 50 results at a time.
- Group results from the same site: Show the first 1, 2, or 3 results.
- Open Links in New Browser Window: yes or no.
- SafeSearch Filter: choose among Strict, Moderate, Off.
- Location: set a default location; Microsoft detects your physical location from your IP address, but you may enter a new geographical location in its place. Remember: you cannot leave the default location empty. If you do not enter a location, Live Search will default to what it reads as your IP address's geolocation.
- Search Language: search in any language or search in one or more of 38 languages including Arabic, Japanese, Chinese, Korean, and Hebrew.

## **The Live Search Results Page**

The clean look continues on the results' page. Once you have entered your search term(s) and clicked the Live Search button, Live will present you with a list of results. Depending on the search you are running, you will see some or all of the following for a web search:

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//FOR OFFICIAL USE ONLY

Live Search

Web Images News Local Q&A More Feeds Academic Video

cardinal Page 1 of 5,443,180 results • [Options](#) • [Advanced](#) **A**

**Cardinal Health** **B**  
Offers drug development services in the pharmaceutical and biotechnology industry. Contains information on company, investor relations and news  
[www.cardinal.com](http://www.cardinal.com) • [Cached page](#) **C**

**Careers | Cardinal Health** **C**  
Cardinal Health is the leading provider of products, services and technologies supporting the health ... At Cardinal Health, everything we do has one ultimate purpose: to help our customers fulfill ...  
[www.cardinal.com/careers/index.asp](http://www.cardinal.com/careers/index.asp) • [Cached page](#)  
[Show more results from www.cardinal.com!](#) **D**

**Cardinal Brands** **D**  
Manufacturer of business forms, accordion files, loose-leaf binders and craft product organizers.  
[www.cardinalbrands.com](http://www.cardinalbrands.com) • [Cached page](#)

**Why Cardinal Health? | Careers | Cardinal Health**  
Cardinal Health is the leading provider of products, services and technologies supporting the health ... Working at Cardinal Health, Cardinal Health has quietly transformed itself from the fastest ...  
[spdc.cardinal.com/careers/why/index.asp](http://spdc.cardinal.com/careers/why/index.asp) • [Cached page](#)

**Related searches:** **E**  
St Louis Cardinals  
Cardinal Health  
Arizona Cardinals  
New Cardinals  
Cardinal Bird  
Louisville Cardinals  
Cardinal Fitness  
Cardinal Pictures

**SPONSORED SITES** **F**  
**Cardinal!**  
Looking for Cardinal? Find exactly what you want today.  
[www.ebay.com](http://www.ebay.com)  
**Arizona Cardinals Shop**  
Great Arizona Cardinal Gifts For Men, Women, & Kids!  
[www.fogtballfanatics.com](http://www.fogtballfanatics.com)

- **A** Type of results (web, image, etc.): the number of resulting pages and estimated total number of results.
- **B** The title of the webpage found, an excerpt from the webpage with the search terms bolded, the url of the webpage.
- **C** Cached page: links to a copy of the page as saved by the Live Search engine; Live Search shows the last date the page was examined by its spider; search terms are not highlighted on the cached page. **Important Note**: the cached copy of Microsoft file types are safe to view.
- **D** Additional results from the same site; clicking on “Show more results from...” will bring up the pages from that site that match the keyword(s).
- **E** Related Searches offer options either for similar terms or terms with multiple meanings, e.g., “cardinal.”
- **F** Sponsored Sites are paid results.



~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

## Live Basic Search

<http://www.live.com/>

Live Search has changed little about its basic search internals in terms of how it handles queries and the number of basic search options.

Live Search assumes as its default that multiple search terms are joined by the **AND** operator, so that a search on the keywords [windows explorer] will find all the webpages that contain both search terms.

Live Search recognizes **double quotes as enclosing a phrase**.

Live Search **will not return any results** if there is no webpage containing all the search terms. Try this query to see what I mean:

[rollerskate handshake specktioneer]

Unlike Google, Live Search **does not limit the number of search terms** to 10 keywords. Live will try to match all the keywords you enter.

Live Search is **not case sensitive**.

Live Search **does not offer any word stemming or truncation**, i.e., searching for variations of search terms. A search for [child] will not find [children].

Live Search **automatically clusters search results**. If you want to see more pages from a specific site, simply select the link following the url of the result.

Live permits the use of nested **boolean** queries in simple search. The operators must be **capitalized**. Live Search will run nested boolean queries (those using parentheses), such as:

[cardinals AND ("st louis" OR arizona) NOT (bird OR catholic)]

Live Search will **ignore stop words**, i.e., commonplace words, **if the query contains non-stop words**; the query [to be or not to be] will only search for the term "not." However, you can search on any single letter or number by itself, e.g., [1]. You can also force Live Search to look for stop words either by enclosing the query in double quotes ["to be or not to be"] or by placing a plus sign in front of the stop word, e.g., [+1 number] or [+to +be +or +not].

Otherwise, it is unnecessary to use the plus sign (+) with any terms because by default Live Search searches for all keywords. However, many times searchers need to exclude certain terms that are commonly associated with a keyword but irrelevant to their search. That's where the minus sign (-) comes in. Using the **minus sign** in front of a keyword ensures that Live Search excludes that term from the search. For example, the results for the search ["pearl harbor" -movie] are very different from the

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

results for ["pearl harbor"]. You may use the boolean operator NOT instead of the minus sign.

Live Search interprets the ampersand [&] as a space, so these searches are virtually identical: [at&t], [at & t], ["at t"]. Also, while Live Search will not actually search on a plus sign, the search engine will search for [c++], although it does not recognize [c+].

### Live Search Advanced Search

Thus far, while Live Search added more advanced search options, it still falls behind Yahoo and Google in the number and type of advanced search options it offers. Nonetheless, Live Search has several advanced search features that are accessible by clicking on the "Advanced" link, which opens a small window that used to be labeled "Search Builder" in MSN Search and is still called that on the Help pages. The advanced search options may also be employed directly by using the correct syntax in the query box. Live Search's web search help is accessible from a link on the Live Search home page, but I prefer the old MSN Search, which is still available and at this point still accurate.

Windows Live Search Help

<http://search.msn.com/docs/help.aspx>

Live Search now offers as many **languages** in which users may search as Yahoo and Google. Using either the language preference settings or the advanced search window, users can select from nearly 40 languages in which to search and see results. There are three ways to specify a search language:

1. in the Advanced search window, select Language, then pull down and click on a specific language.
2. type your search terms into the search box, and then add language: followed immediately by the two-character language code. For example, to search only for sites in French: [language:fr keyword]
3. a more permanent change would be to go to the Options page and change your primary search language.

Live Search does not distinguish words using **diacritical marks** such as accents or umlauts. Live Search finds terms matching those with and without the diacritic. The term [façade] finds façade and facade, and vice versa.

Live Search offers several special search terms to restrict searches and make them more effective.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

- **site/domain:** restricts results to a specific website or domain, including a specific top-level domain. Can be used with or without keywords.

Advanced Search > Site/Domain returns results from specific domains (com, gov, dell.com, a country digraph, etc.)

Examples of how to use the site: command:

[site:amazon.com] finds www.amazon.com, auction.amazon.com, www.amazon.com/dvd/. However, it will not find www.amazon.com.br.

[books -site:amazon.com] finds pages containing the keyword "books" that are not at any amazon.com website.

[site:ir] finds all the pages from the Iranian (.ir) top-level domain indexed by Live Search.

- **country/region:** on the Advanced menu; it is identical to the site/domain search for a country digraph. However, if you do not know a country's top-level domain, you can use the Country/Region pull-down menu to select the country, and Live Search will automatically enter the correct country digraph for you.
- **language:** restricts results to pages in a specific language. Users must specify a language using the two-letter code or use Advanced Search. Can be used with or without additional keywords.

Advanced Search > Language uses pull-down menu to select languages.

Examples of how to use the language: command:

[language:ro] restricts results to sites written in Romanian.

[language:es domain:mx méxico] restricts results to sites written in Spanish in the Mexican top-level domain that contain the term "mexico."

- **url:** unlike Google's url query, Live's url query checks to see if the domain or web address is in the Live Search index. This query is not really intended to be used with other search terms.

Examples of how to use the url: command:

[url:nasa.gov] or [url:education.jpl.nasa.gov] will check to see if a site is indexed by Live Search.

- **inurl:** restricts results to pages that contain search terms within the url of a site. Multiple terms can be used, but all must appear in the url (this query is similar to Google's allinurl: query).

Examples of how to use the inurl: command:

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

[inurl:microsoft] finds all pages containing “microsoft” anywhere in the url

[inurl:microsoft downloads] finds all pages containing both the terms “microsoft” and “downloads” anywhere in the url.

- **inbody:** restricts results to pages containing search term(s) in the body of a webpage. Can be used with or without other search terms.

Example of how to use the inbody: command:

[inbody:amazon -inurl:amazon] finds all pages containing the term “amazon” anywhere in the body (text) of a webpage but which do not contain the term “amazon” in the url of the page.

- **intitle:** restricts results to pages containing search term(s) in the webpage’s title. Can be used with or without other search terms.

Examples of how to use the intitle: command:

[intitle:amazon inbody:brazil] will find pages that contain “amazon” in the title of the webpage and “brazil” in the body text of the webpage.

- **contains:** restricts results to pages that have links to specific the file type(s). Can be used with or without other search terms.

Examples of how to the contains: command:

[music contains:mp3] finds webpages that contain links to MP3 files and have the keyword “music” in them.

[“final report” contains:pdf] finds webpages that contain links to PDF files that have the phrase “final report” in them.

- **link:** Restricts results to pages containing links to a specific url. Can be used with or without additional keywords.

Advanced Search > Links to returns results for pages that currently link to a specific url.

Examples of how to use the link: command:

[link:jpl.nasa.gov] finds all pages containing links to the specific domain jpl.nasa.gov.

[link:jpl.nasa.gov asteroid] finds all pages containing links to any page in the jpl.nasa.gov domain and the keyword “asteroid” anywhere on the linking webpage.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

- **linkdomain:** Restricts results to pages that link to any page within the specified domain. This is a broader search than the link: query. You can use this option to determine how many links there are to a specific page from sites indexed by Live Search. Can be used with or without additional keywords.

Examples of how to use linkdomain:

[linkdomain:jpl.nasa.gov] finds all pages containing links to any page at jpl.nasa.gov, including echo.jpl.nasa.gov, voyager.jpl.nasa.gov, etc.

[linkdomain:jpl.nasa.gov cassini] finds all pages containing links to any page at jpl.nasa.gov and that also include the term "cassini" at the linking website.

star [linkdomain:jpl.nasa.gov -site:jpl.nasa.gov] will find all pages containing links to any page at jpl.nasa.gov from sites other than jpl.nasa.gov (this eliminates internal links from the overall results).

- **linkfromdomain:** Restricts results to pages that are linked from the specified domain. This query only works with second-level domains, e.g., [domain.com]. You can use this option to determine how many links there are from a specific page. Can be used with or without additional keywords.

Examples of how to use linkfromdomain:

[linkfromdomain:nasa.gov] finds all the pages the nasa.gov domain links to, i.e., links from nasa.gov to site x.

[linkfromdomain:nasa.gov standards] finds all pages the nasa.gov domain links to that contain the term "standards" on their webpage, i.e., links from nasa.gov to site x where site x contains the keyword "standards."

- **Results ranking:** allows users to emphasize different factors to get a different set of results for the same search.

1. Type your search terms into the search text box, and then click Advanced Search.
2. Select Results ranking, and then move the equalizer slider(s) in the direction you want.

Live Search Help explains Results ranking in this way:

"You can put emphasis on different factors to get a different set of results for the same search. The sliders control:

- Updated recently: To modify your search to add emphasis to sites that have been recently added to the search index, move the left slider up.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

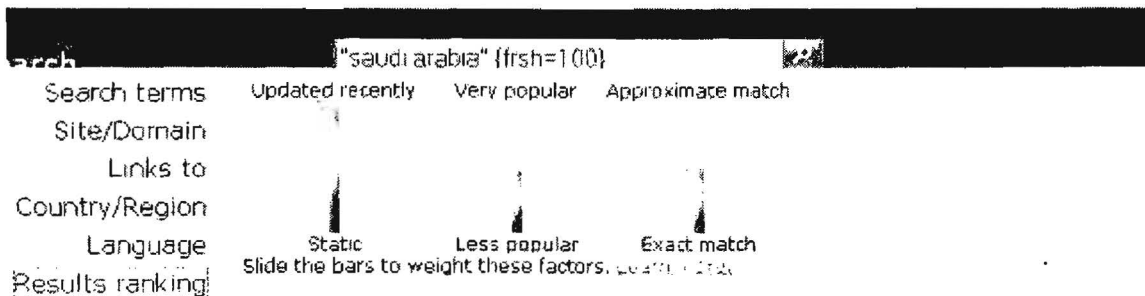
~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

- Very popular: To add emphasis to sites by the number of other sites that link to them, move the middle slider up.
- Approximate match: To put the most emphasis on the match between your exact search words and your results, move the right slider down.

## Notes

- Approximate match overrides the first two slider rankings.
- Results ranking applies to web searches only."

It is easier to visualize how to use results ranking by looking at an example. In this case, the search on ["saudi arabia"] has been reranked to emphasize pages that have been recently updated {frsh=100} means the "freshness" ranking of these pages is 100 or the most recently updated pages in the Live Search database:



- **filetype:** restricts results to a specific filetype. Can be used with or without additional keywords. The file types Live Search will search for include the major Microsoft file types and a few others:

Microsoft Excel (xls)

Microsoft PowerPoint (ppt)

Microsoft Word (doc)

Portable Document Format (pdf)

Rich Text Format (rtf)

Text (txt)

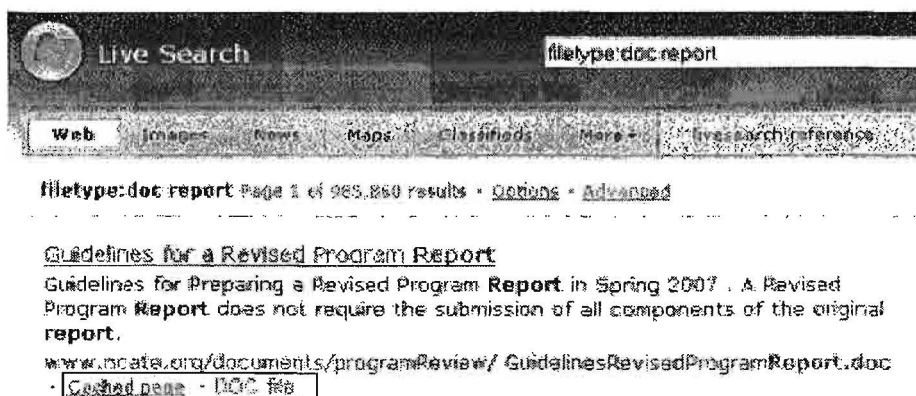
Examples of how to use the filetype: command:

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

[filetype:doc domain:nasa.gov] finds all Word files at the NASA domain in Word format.

[filetype:xls "financial data"] finds all Excel spreadsheets that contain the phrase "financial data."

Live Search **does offer safe previewing of non-HTML file types, and this is especially useful for Microsoft file types, such as Word documents and PowerPoint slides.** In order to access the safe HTML versions, users must select the "Cached page" on the results page:



- **IP:** finds all the sites on a specific host computer. Can be used with or without additional keywords.

Examples of how to use the ip: command:

[ip:66.218.77.68] finds all the sites on this specific host computer.

[ip:66.218.77.68 "computer security"] finds all the sites on this specific host computer containing the phrase "computer security."

- **feed:** one of two RSS search options; similar to the filetype: command. It limits searches to text within a feed. Feeds are specially formatted brief descriptions of content with a link to the full version of that content. RSS (and the competing Atom) feeds are in XML format. These feeds are usually used for syndicating web content such as blogs and news. The feed: command only searches the text of the feed, which is often a very condensed description of the full web content.

Example of how to use the feed: command:

[feed:"trojan horse"]

Each of the results represents an XML feed that includes the phrase "trojan horse." There is no point in clicking on the link in a browser because that brings

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

up the XML page that most browsers are not designed to parse. The cached copy shows the search terms as they appeared in the feed.

- **hasfeed:** shows the pages that offer feed links and, if you add a keyword (something I'm pretty sure Live intended you to do), the pages with feed links and that also have that keyword somewhere on the page.

Examples of how to use the hasfeed: command:

```
[hasfeed:"trojan horses"]
```

The results are webpages that offer news feeds and contain the phrase "trojan horses" on the webpage. This does not guarantee, however, that the news feed will be about Trojan horses, but the chances are good that if you are looking for sites with newsfeeds about this topic, you can find them using this query.

```
[hasfeed:encryption site:microsoft.com]
```

This query should find the pages at the Microsoft website with feeds about encryption. What this query actually finds are pages at the Microsoft site that contain both XML feeds and the word encryption in the text, so a little research will reveal which of these Microsoft newsfeeds are the most appropriate to the topic of encryption.

This command is listed at the Live.com but is not working properly:

- **inanchor:** restricts results to pages containing search term(s) in the webpage anchor.

## Live Search Special Features

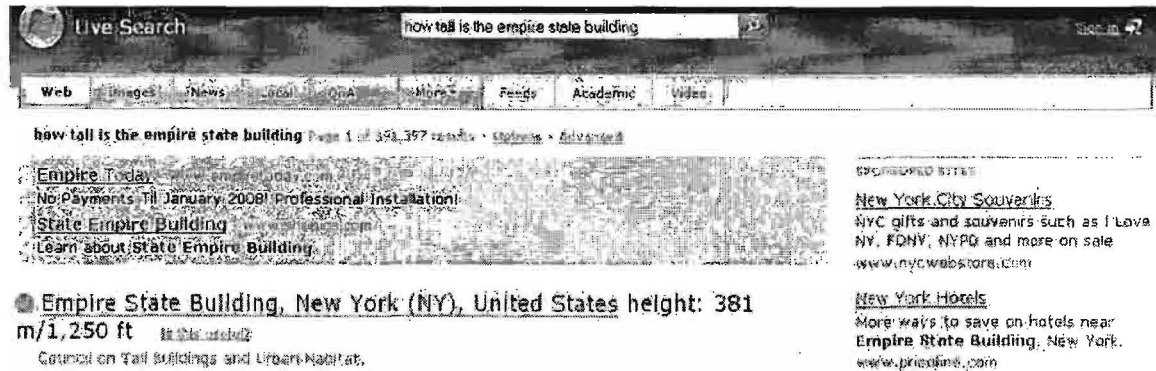
**Spell Checker:** Live Search has a very good spell check option. When you input a query, Live checks to see if you are using the most common spelling of the keyword. If not, just like Google, Live nicely asks, \* Were you looking for x, where x is the most common spelling. The Live Search dictionary also includes some proper names.

**Dictionary Definitions:** as with Google and Yahoo, Live Search offers the define option. To use it, type [define] then a word or brief phrase, e.g., [define king cobra]. Live's define option is more limited than some others because it only refers to Encarta.

**Encarta:** Microsoft's encyclopedia and general reference source Encarta provides answers to questions and facts about a topic. Users can type questions and (sometimes) get direct answers to them by simply entering a question and clicking on Search. Live Search does a much better job of correctly answering questions than MSN Search did (unlike its predecessor, it correctly identified Chirac as the

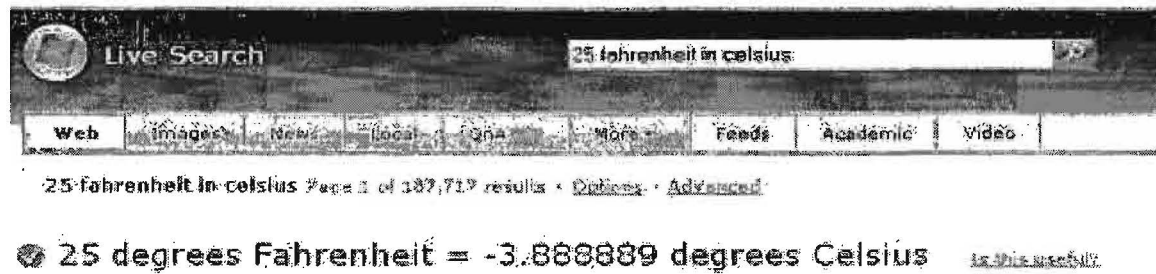


French president). Live Search can also directly answer certain specific questions, such as [how tall is the empire state building]:



Live Search no longer has the Encarta option that used to exist on MSN Search. For now, the easiest way I have found to invoke Encarta from Live Search and to take advantage of the Encarta Free Pass is to limit your search to Encarta using the site: syntax, e.g., [site:encarta.msn.com keyword]. This will give you two hours of free Encarta research.

**Measurement Conversions:** Live Search uses Encarta Answers to convert distance, weight, time, volume, and temperature. The conversions may be stated as questions, e.g., [how many seconds in a year?], or as a simple phrase:



**RSS Results:** when added to the end of any search result, the **&format=rss** parameter will provide users those search results via RSS. "When you subscribe to this RSS feed from Live Search, you'll get the top ten search results for this query delivered to your RSS Reader or personalized site. You can subscribe to any number of RSS feeds of Live Search results and view them all in your RSS Reader without re-running your search queries." To use this option, first search for your terms, e.g., [tsunami relief]. On the results' page, add **&format=rss** to the end of the url in the address bar and hit return:

<http://search.live.com/results.aspx?q=tsunami&mkt=en-US&form=QBRE&go.x=0&go.y=0&go=Search&format=rss>



UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

**Calculator:** Live Search uses the Encarta Calculator and Equation Solver to perform mathematical functions using “operators, exponents, and roots, factorials, modulo, percentages, logarithms, trig functions, and mathematical constants.” The Encarta calculator appears to be the most sophisticated of all those offered by major search engines because it will even solve complex algebraic equations, such as  $4x^3 - 2x + .9 = 0$

The Live Search calculator uses the following symbols:

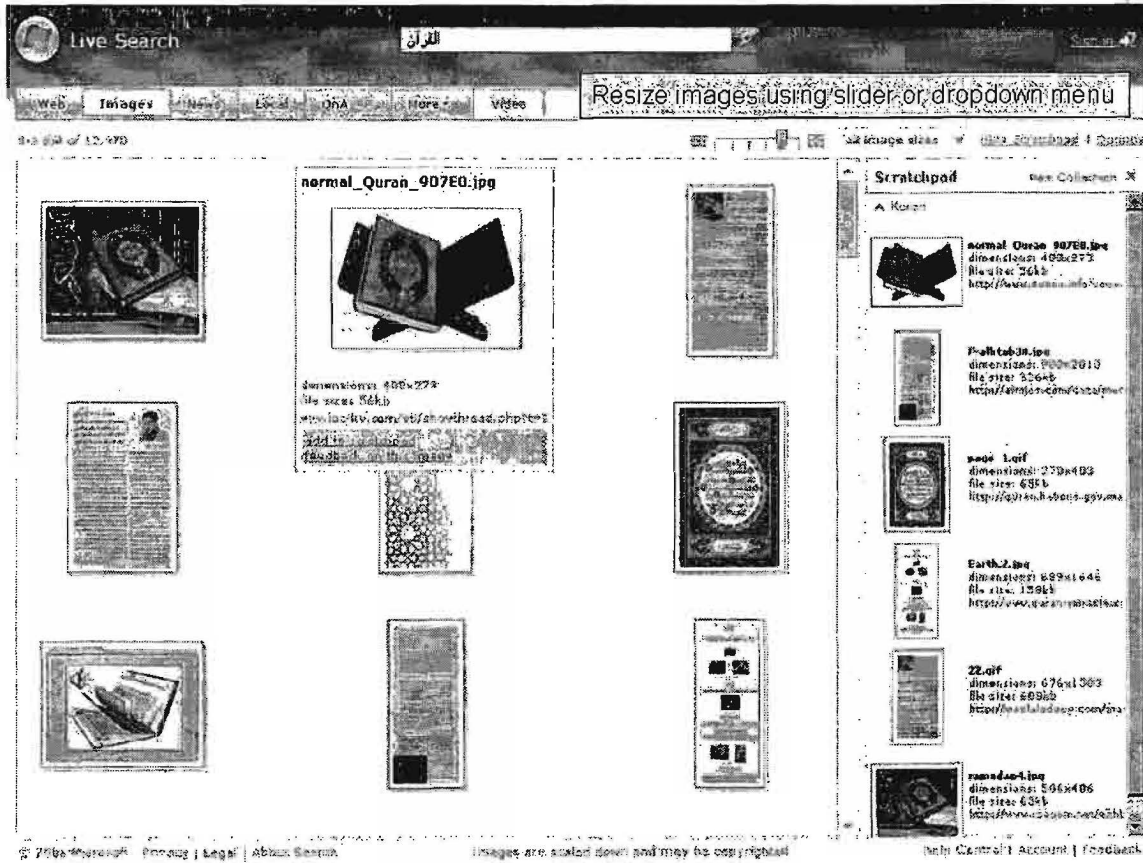
Add	+
Subtract	-
Multiply	*
Divide	/
Raise a number to an exponent (For example, $3^2$ is 3 squared)	^
Specify the order of operation	( )
Find a percent of a number	% of
Find the square root of a number	sqrt
Find the sine of an angle	sin
Find the cosine of an angle	cos
Find the cosine of an angle	!

[http://search.world.msn.com/docs/help.aspx?t=SEARCH\\_PROC\\_FindFactsNStatistics.htm](http://search.world.msn.com/docs/help.aspx?t=SEARCH_PROC_FindFactsNStatistics.htm)

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

### Live Search Services

**Images:** the Live Search image database is no longer Picsearch. Instead, Live Image Search uses Microsoft's own proprietary image database. Images are displayed as thumbnails (small versions of the original images), and the user can resize the thumbnails either using the slider or the dropdown "all image size" menu. One of the other changes to image search is the addition of a Scratchpad, which lets users drag and drop images onto a collection of images on the right-hand side of the screen. At this time, you do not have to have an account with Live in order to retrieve your image collections (they are retrieved based upon a cookie set by Live). When you mouse over an image, it zooms to a slightly larger size and moves toward the center and a box appears that shows the image source and size, and a link to the page where the image resides. If you click on the link, the linked page appears on the right with a "show image" in the top left corner. At present there are no advanced search options for images.



Also, when you search for a famous person using Live image search, look for the "Related People" window to appear on the right side of the screen. This can be an

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

extremely useful tool in finding relationships between people in the news or historical figures.

The Live image search respects some but not all of the web search syntax and some of it is not really very useful for image search:

- **site/domain:** restricts results to images from a specific website or domain, including a specific top-level domain (com, gov, dell.com, a country digraph, etc.). May be used with or without keywords.

Examples of how to use the site: command in image search:

[site:amazon.com "twelfth night"] finds images of "twelfth night" that are from amazon.com; note that the images from amazon.com may reside on another website (amazon.com is in the image's url).

[site:ir] finds all the image pages from the Iranian (.ir) top-level domain indexed by Live Search.

- **inurl:** restricts results to images that contain the term in the url of the image itself. Can be used with or without other search terms.

Examples of how to use the inurl: command in image search:

[inurl:amazon "rain forest"] finds all pages containing "amazon" in the url of the image and "rain forest" anywhere on the webpage.

- **intitle:** restricts results to images that appear on pages containing search term(s) in the title of the webpage. Can be used with or without other search terms.

Examples of how to use the intitle: command in image search:

[intitle:amazon brazil] will find pages that contain "amazon" in the title of the webpage and "brazil" anywhere on the webpage.

[intitle:amazon inurl:brazil] will find pages that contain "amazon" in the title of the webpage and "brazil" in the image's url.

**Video Search:** Live video search is clearly trying to be competitive in the video search market. In October, Microsoft announced a new partnership with Blinkx to power its video search. This looks like a very good move for Microsoft. "Blinkx already powers video search on sites ranging from AOL to ITN, Lycos and Times Online. It also indexes video from the likes of BCC, Fox, MTV, Sky News, Reuters and YouTube and makes and makes videos on those sites searchable on Blinkx or partner sites. To date, the company has indexed more than six million hours of audio, video, and TV programming to make it searchable."<sup>55</sup> However, as of this writing, *the Live video search has not yet been updated to reflect this partnership.*

---

<sup>55</sup> Eric Auchard, "Blinkx Signs Microsoft Pact," Reuters via Yahoo, 9 October 2006, <[http://news.yahoo.com/s/nm/20061009/wr\\_nm/media\\_blinkx\\_dc\\_3](http://news.yahoo.com/s/nm/20061009/wr_nm/media_blinkx_dc_3)> (17 October 2006).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

As of now, the Live video search results include a thumbnail image from the video with the title, source, length, and format. All videos are viewed at the originating site, as shown below with the Newsweek On Air interview with Iranian President Ahmadinejad.



You can use some of the web search syntax for video search. Note the difference between these two searches:

[site:reuters.com iran]

[reuters iran]

The first query returns only those videos on Iran from the Reuters website; the second query returns queries from any site that includes the keywords "reuters" and "iran." We will have to wait and see how these query options change once the results come from Blinkx.

**News Search:** as of now, the Live news search is only a list of stories listed by relevance. MSN Newsbot <<http://newsbot.msnbc.msn.com/>> remains Microsoft's premier news page. However, if you want to search for news stories, MSN Newsbot takes you directly to the new Live news search. **Most of the web search commands work for news search.** Especially useful is the site/domain: syntax, which lets users limit a news query to a specific source:

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

[[site:washingtonpost.com](http://site:washingtonpost.com) iran] finds pages at the Washington Post website that contain the keyword "iran." One big drawback of the Live news search is its inability to list the results by date.

**Feed Search (Beta):** This search is virtually identical to the feed: websearch. It limits searches to text within a feed. Feeds are specially formatted brief descriptions of content with a link to the full version of that content. RSS (and the competing Atom) feeds are in XML format. These feeds are usually used for syndicating web content such as blogs and news. The feed search only searches the text of the feed, which is often a very condensed description of the full web content.

Example of how to use the feed: command:

[feed:"trojan horse"]

Each of the results represents an XML feed that includes the phrase "trojan horse." There is no point in clicking on the link in a browser because that brings up the XML page that most browsers are not designed to parse. The cached copy shows the search terms as they appeared in the feed.

**Live Book Search (beta):** Microsoft added its own proprietary book search in late 2006. Details are in the [Book Search section](#) below.

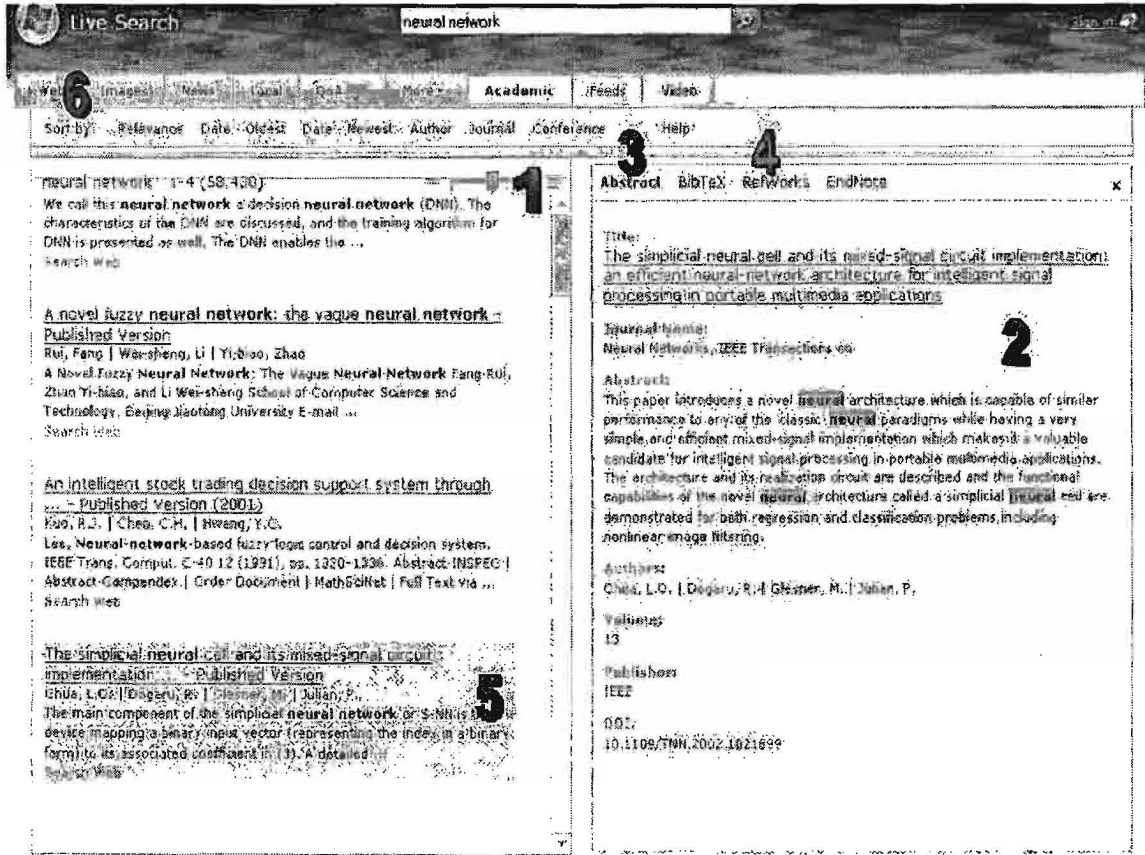
**Academic (Beta):** Microsoft introduced Academic Search Beta for scholarly search earlier this year, and it is now also a Live search option. Academic search still has a separate website at the [Windows Academic Live Beta Homepage](#). Clearly, Academic search is intended to compete with Google Scholar and other scholarly search sites. Unlike Google Scholar, Academic search focuses on computer science, physics, medical, and electrical engineering publications. As with Amazon and Google Scholar, Academic search has partnered with the **Online Computer Library Center (OCLC)**. "OCLC's involvement in Windows Live Academic is part of the Open WorldCat Find in a Library program,"<sup>56</sup> and also provides metadata from [WorldCat](#) to Academic search to give researchers access to the resources in library collections around the world.

As with almost anything, Academic search has good features and weaknesses. Here is a snapshot of the first page of results on the search [neural network]. When you execute a query, you will be presented with an interface that looks like this. One of the first things you notice is the split screen, which I actually like.

---

<sup>56</sup> "WorldCat live in Windows Live Academic search tool," OCLC Newsletter, Issue 2, 2006, <http://www.oclc.org/nextspace/002/updates.htm> (17 October 2006).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~



On the left you see the results; on the right-hand side of the screen is more detailed information that appears automatically as you click on different results. You have the option to view the abstract or properly formatted citations:

1. Slider bar: This allows you to expand or contract the amount of information contained in the search result
2. Preview pane: This pane allows you to obtain more information on the result that you are hovering over with your mouse on the results pane
3. Abstract: one of the options in the preview pane - choosing this option will allow you to see the abstract of the article that you are hovering over with your mouse on the results pane
4. BibTeX/RefWorks/EndNote: citation options in the preview pane - choosing one of these options will allow you to see the formatted citation (BibTeX, RefWorks, or EndNote format) on the preview pane for the search result that you are hovering over with your mouse on the results pane. BibTeX, RefWorks, and EndNote are different formats that allow users to create citations automatically. The EndNote RIS



~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

format is compatible with EndNote, Reference Manager, and ProCite programs.

5. Search result: the actual search result; this includes links to the full text of the paper, link to search the web for that paper and potentially links that allow you to search your library for access to the full text from their subscription
6. Sort by options: allows you to sort the search results by relevance (default), oldest or newest date, author of paper, journal, or conference.

The best things about Academic search are:

- a list of journals it searches (something Google Scholar sorely needs); still, the list is too general (for example, IEEE Computer Society encompasses a huge number of journals and publications):  
<<http://academic.live.com/AcademicJournals.htm>>
- the preview pane is a good idea—no need to open new windows.
- the slider to view more or less information.
- the ability to extract citations (if you need to cite the information, this is a big benefit).
- the “find it in a library near you” search: [worldcatlibraries keyword].

Academic search needs to improve:

- lack of citation search (everyone seems to agree this is the biggest problem that simply must be rectified).
- no advanced search (may come later).
- not enough content.

#### **Edit Macros:**

<http://search.live.com/macros/default.aspx>

This new feature allows users to “create their own search engine,” so to speak. Of course, you are not really making a new search engine. In fact, what you are really doing with a basic macros’ search is automatically generating a “site:” search. A basic macros search for [“north korea” “nuclear test”] on CNN, Reuters, and USA Today is equivalent to:

[“north korea” “nuclear test” (site:www.cnn.com OR site:www.reuters.com OR site:www.usatoday.com)]

The advantage of the macros is that they are much simpler to create, especially if you want to search 30 sites, and you can easily save and retrieve your macros, but ***you must sign in to Live.com in order to save and retrieve your macros.***

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~



UNCLASSIFIED//FOR OFFICIAL USE ONLY

Live Search

Web Images News Local Answers More [livesearch.reference](#)

uranium-235 macro:livesearch.reference Page 5 of 226 results - [Options](#) - [Advanced](#)

[Uranium-235 - Wikipedia, the free encyclopedia](#)  
 Uranium-235 is an isotope of uranium that differs from the element's other common isotope, uranium-238, by its ability to cause a rapidly expanding fission chain reaction; i.e., it is fissile.  
[en.wikipedia.org/wiki/Uranium-235](http://en.wikipedia.org/wiki/Uranium-235) - [Cached page](#)

[Uranium: Definition and Much More From Answers.com](#)  
 The most abundant (greater than 99%) and most stable is uranium-238 (half-life 4.5x10 9 years); also present are uranium-235 (half-life 7x10 8 years) and uranium-234 (half-life 2.5x10 5 years).  
[www.answers.com/topic/uranium](http://www.answers.com/topic/uranium) - [Cached page](#)

[Enriched uranium - Wikipedia, the free encyclopedia](#)  
 These pie-graphs showing the relative proportions of uranium-238 (blue) and uranium-235 (red) at different levels of enrichment.  
[en.wikipedia.org/wiki/Uranium\\_enrichment](http://en.wikipedia.org/wiki/Uranium_enrichment) - [11/14/2006](#) - [Cached page](#)

[uranium -- Encyclopædia Britannica](#)  
 Fission occurs with slow neutrons in the relatively rare isotope uranium-235 (the only naturally occurring fissile material), which must be separated from the plentiful isotope uranium-238 for its ...  
[www.britannica.com/eb/article-907442/uranium](http://www.britannica.com/eb/article-907442/uranium) - [Cached page](#)

[HighBeam Encyclopedia - Free Online Encyclopedia for Reference ...](#)  
 The most abundant (greater than 99%) and most stable is uranium-238 (half-life 4.5x10 9 years); also present are uranium-235 (half-life 7x10 8 years) and uranium-234 (half-life 2.5x10 5 years).  
[www.encyclopedia.com/doc/1E1-uranium.html](http://www.encyclopedia.com/doc/1E1-uranium.html) - [Cached page](#)

[uranium: Discovery and Uses](#)  
 Uranium-235 is the only naturally occurring nuclear fission fuel, but this isotope is only about 1 part in 140 of natural uranium; the balance is mostly uranium-238.  
[www.infoplease.com/ce6/sci/A0861224.html](http://www.infoplease.com/ce6/sci/A0861224.html) - [Cached page](#)

SPONSORED LINKS  
[Uranium 235 Ringtone](#)  
 We Have Every **Uranium 235** Ringtone. Get Them Now.  
[mytoneallstars.com/Uranium-235](http://mytoneallstars.com/Uranium-235)

I believe there are too many results from Wikipedia in the reference search, but you can easily eliminate the Wikipedia results by adding [-site:wikipedia.org] to any query (conversely, you could limit your search to Wikipedia by adding [site:wikipedia.org] to your query. Live Search Macros are only the latest in a number of "create your own search engine" options, all of which are variations on complex queries of already existing search engines. For comparison, see the section on [Custom Search Engines](#) below.

**QnA:** Live Search's new QnA (question and answer) search is mostly fluff, at least for now. You can look at the questions and responses to see what I mean (typical questions: "How can i get my Space Cadet Pinball that was preinstalled in Windows XP back in Windows Vista?" "Do you think the Internet is contributing to 'Intellectual Laziness'?"). Lots of opinion, not a lot of fact. Let us hope this is not all that "Web 2.0" portends.

**Live Platform:** In September 2005 Microsoft announced it would begin offering APIs for Live Search, Virtual Earth, Spaces (weblogs), Messenger, Gadgets, and Expo classified ads database. These have begun to rival Google in terms of innovation

UNCLASSIFIED//FOR OFFICIAL USE ONLY

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

and shared technology. To keep abreast of these changes, I recommend the MSN Developer Center.

MSN Developer Center's Windows Live Platform and Services for Web Mashups  
<http://msdn.microsoft.com/live/default.aspx>

Microsoft subsequently opened **Windows Live Dev (Beta)**, a “one-stop shop for the Windows Live Platform, including information on getting started with Windows Live services, latest documentation and APIs, samples, access to community areas and relevant blogs, and announcements of future releases and innovations.”<sup>57</sup> Microsoft is trying to make it easy for users to integrate their products with Live regardless of platform, browser, or language. Certainly the first two are a departure for Microsoft, which in the past had made the requirement of a Windows platform and an Internet Explorer browser a necessity in most cases in order to “play ball” with the software giant. A further example of Microsoft’s reluctant openness is the fact that Microsoft’s Internet Explorer 7+ browser will not default to Live Search, something other search engines had objected to.

Windows Live Dev (Beta) <http://dev.live.com/>

Microsoft is working very hard to improve and expand its search properties, so much so that at times one feels as if we can see them working under the hood as we watch. Clearly, there are many things that need improvement and many things that are very good about Live.com. It will continue to be one of the top search sites on the Internet. If you are interested in keeping up with news about and changes to Live Search, there is a blog devoted to it; the blog offers RSS and Atom syndication. Also, all the Windows Live Beta projects are accessible through one webpage if you want to see what Microsoft is planning.

Windows Live Ideas Beta <http://ideas.live.com/>

Live Search Weblog <http://blogs.msdn.com/livesearch/>

---

<sup>57</sup> Windows Live Dev, Live Dev News, 8 June 2006,  
<<http://dev.live.com/blogs/devlive/archive/2006/05/19/15.aspx>> (17 October 2006).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

---

## Gigablast

---

The Gigablast search engine, which has been around since 2002, is still not quite in the same league as powerhouses Google, Yahoo, and Live Search, but it is well on its way to becoming one of the best search engines. That's something of a surprise given Gigablast's humble origins and unique status among major search engines. In case you're not familiar with Gigablast, it is different from its major competitors most notably because it is still owned and largely run by the guy who first wrote its C++ code in 2000. Matt Wells is still the very hands-on proprietor of Gigablast. Its database *now indexes over 2 billion pages*, up from 650 million in late 2004. While this falls short of the size of the Google, Yahoo, and Live Search databases, it's not bad, especially considering a lot of the "stuff" in those databases is dross and the numbers are not verified independently.

How does Gigablast stack up to the big boys? Gigablast has some very nice features, some of which are unique to it, such as the IP range search (something AlltheWeb once offered).

### Gigablast

<http://www.gigablast.com/>

#### Strengths

- simple interface
- cached copies with date indexed [archived copies]
- cached copies of webpages without images [stripped]
- links to Internet Archives [older copies]
- clusters results by default (can be turned off)
- no limit on number of search terms
- file types indexed include Microsoft Word, Excel, and PowerPoint, as well as PDF, PostScript, HTML, and text; syntax is:
  - **type:pdf** for Adobe Acrobat PDFs
  - **type:doc** for Microsoft Word documents
  - **type:ppt** for PowerPoint presentations
  - **type:xls** for Excel spreadsheets
  - **type:ps** for PostScript files
  - **type:text** for ASCII text files
  - **type:html** for HTML Web pages

UNCLASSIFIED//FOR OFFICIAL USE ONLY

- **unique feature: IP range**; Gigablast adds the ability (unique as far as I know) to search on an IP address range. [ip:216.239.41] will find all IP addresses that begin with 216.239.41

This query finds all the sites in the Gigablast database that begin with the IP address 66.218.77:



Results 1 to 10 of about 70,997 for ip:66.218.77 .

#### Yahoo! GeoCities

us.geocities.yahoo.com/goview?member=batman\_927 - 33.1K - [archived copy] - [stripped] - [older copies] - indexed: Jul 26 2005 - modified: Jul 27 2005

#### Roces Guitars

us.geocities.yahoo.com/gb/sign?member=rocehogan - 2.3K - [archived copy] - [stripped] - [older copies] - indexed: Jul 26 2005 - modified: Jul 28 2005

[More results from this site.]

This query finds all the sites in the Gigablast database residing on the specific host whose IP address is 66.218.77.68:



Results 1 to 10 of about 49,152 for ip:66.218.77.68 .

#### SMScheerleading

Description: The official cheerleading page for SMS in Manassas, Virginia, provides tryout information, team news, and contacts.

Category: Sports: Cheerleading: Youth and Recreation

www.geocities.com/sabrescheercoach/SMScheerleading.html - 30.8K - [archived copy] - [stripped] - [older copies] - indexed: Oct 09 2005 - modified: Feb 18 2005

- other special syntax includes **link:**, **site:**, **title:**, and **suburl:**, which searches for webpages that have the keyword anywhere in the url
- although Gigablast will ignore stop words in a long query, users can search on any word or number by itself
- default operator is AND; OR and AND NOT also work; nested queries (with parentheses) are supported
- **unique feature: indexes and displays of generic meta tags**; only search engine that will display the metatags in the results list, but the syntax for this query is very complex. Please see the Gigablast review at Search Engine Showdown for details on this type of query:

**"Meta Tag Searching and Display:** Gigablast is the only search engine indexing meta tags beyond just the meta description and meta keywords that some others index. It is the only search engine that can also display meta

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

tags in the results list. Gigablast claims to be indexing all “generic” meta tags. In addition, it can display the meta tags in the results list. Doing this requires adding commands to the URL of the results list. At the end of the url, add a `&dt=` followed by the word(s) for the meta tags, followed by a colon, and then a number to represent how many characters from each meta tag should be displayed. So, for example, adding `&dt=keywords+author+generator+description:30` will display the meta tag content for meta keywords, meta author, meta generator, and meta description tags for any records retrieved. Use a `+` between meta tag words. It seems that this “generic” meta tag approach excludes more complex meta tags like Dublin Core, which use a syntax like `DC.Creator`. The dot syntax will not work for the display command, although Gigablast does index some of the content of these tags.”<sup>58</sup>

### Sample Output of Meta Tag Search

The screenshot shows a search results page from Gigablast. The search URL is `http://www.gigablast.com/search?k1z=134827&q=dublin+core&dt=k...`. A callout box with a black border and white background contains the text: "add string to the end of resulting url in address". Below the search results, there are three entries:

**DC-dot**  
 ..DC-dot now conforms with the Expressing **Dublin Core** in HTMLXHTML meta and I...  
 ..Now you can click on the DC-dot button, wherever you are, to create **Dublin Core** me...  
 about... This service will retrieve a Web page and automatically generate **Dublin Core**...  
 metadata, either as..  
 Description: Give DC-dot a URL and see the Dublin Core it generates.  
**keywords:** Dublin Core, DC; generator; editor; Warwick Framework; SOIF; TEI; USMARC; XML; GILS; ROADS; RDF; IMS  
**generator:** HTML Tidy, see [www.w3.org](http://www.w3.org)  
**description:** A CGI based Dublin Core  
 Category: Reference: Libraries: Library and Information Science: Technical Services: Cataloguing: Metadata: Dublin Core  
[www.ukoln.ac.uk/metadata/dcdot/](http://www.ukoln.ac.uk/metadata/dcdot/) - 8.8k - [archived copy] - [stripped] - [older copies] - indexed: Oct 05 2005 - modified: Dec 11 2001

**Dublin Core Metadata Template**  
 ..When the list of Qualifiers for **Dublin Core** elements is finally decided upon, this template...  
 will... You may include my name and email-address in a list of those using **Dublin Core**.  
 Additional DC... **Dublin Core** Metadata Template.. This service is provided by the "Nordic...  
 Metadata Project" in..  
 Description: from the Nordic Metadata Project  
 Category: Reference: Libraries: Library and Information Science: Technical Services: Cataloguing: Metadata: Dublin Core  
[www.lub.lu.se/cgi-bin/nmdc.pl](http://www.lub.lu.se/cgi-bin/nmdc.pl) - 40.5k - [archived copy] - [stripped] - [older copies] - indexed: Oct 05 2005

**Dublin Core/MARC/GILS Crosswalk**  
 ..For conversion of MARC 21 into **Dublin Core**, many fields may be mapped into a single...  
**Dublin Core**... In the **Dublin Core** to MARC mapping, two mappings are provided,  
 one for unqualified **Dublin Core**... The following is a crosswalk between the fifteen elements...  
 in the **Dublin Core** Element Set and MARC..  
 Description: Library of Congress  
**keywords:** MARC Dublin Core GILS Crosswalk  
**author:** Library of Congress Network Development and MARC Standards Office  
**description:** Crosswalk from Dublin Core  
 Category: Reference: Libraries: Library and Information Science: Technical Services: Cataloguing: Metadata: Crosswalks  
[web.loc.gov/marc/dc/cross.html](http://web.loc.gov/marc/dc/cross.html) - 18.6k - [archived copy] - [stripped] - [older copies] - indexed: Oct 06 2005 - modified: Dec 31 2002

- clearly displays date webpage was indexed and, in some cases, modified
- search query spellchecker (Did you mean? option)

<sup>58</sup> Greg R. Notess, “Review of Gigablast,” Searchengineshowdown, 17 September 2006, <http://www.searchengineshowdown.com/features/qigablast/review.html>> (14 November 2006).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

- **undocumented feature:** will search in some specific languages, but I don't know how many; use **language:de** to search for webpages in German, for example.

### Weaknesses

- most obviously, the Gigablast index is still smaller than those of Google, Yahoo, or Live Search
- no truncation
- is not case sensitive
- no wildcard
- limited file type searches
- limited language options
- poor documentation

### Gigablast Options & Services

**Custom Topic Search:** Gigablast offers some special options, the most important of which is a Custom Topic Search, which I discuss in detail under the Custom Search Engines section below. ***If you don't read anything else about Gigablast, please take a look at this innovation.***

**Directory:** As with Google and Yahoo, Gigablast's web directory uses the Open Directory Project's collection but Gigablast use a "hypertechnology for searching the directory that allows its users to perform searches over websites, not just the actual pages, under any topic in the directory, in effect, instantly creating over 500,000 vertical search engines. Additionally, all directory searches are enhanced by the massive amount of link information from Gigablast's multi-billion page index." So a Gigablast directory search returns not only DMOZ categories but "Giga Bits" and website listings as well.

**XML Search Feed:** Gigablast also offers an XML Search Feed that will run up to 1000 queries per day with a maximum of ten results each. But remember, you must have XML parsing software to read XML feeds, so this new feature isn't an option for all users.

XML Search Feed

<http://www.gigablast.com/searchfeed.html>

**Giga Bits:** Gigablast has its own *refine* option called "Giga Bits." Giga Bits are terms that appear in a blue box at the top of a results page to help refine and focus your search.

**Related Pages:** Gigablast's Related Pages were introduced in March 2005. Related Pages are "relevant search results which do not necessarily contain the searcher's



UNCLASSIFIED//FOR OFFICIAL USE ONLY

query terms." Related Pages are results that are contextually related to the query terms without having a direct connection to them. The Related Pages appear in the yellow box on the results page.

**GIGABLAST** "artificial intelligence" 10 Search

Results 1 to 10 of about 2,640,799 for "artificial intelligence"

<b>Giga Bits (more)</b>	26% <a href="#">Artificial Life</a>	21% <a href="#">Artificial Intelligence Resources</a>	20% <a href="#">Distributed Artificial Intelligence</a>
30% <a href="#">CMU Artificial Intelligence Repository</a>	23% <a href="#">Artificial Intelligence Laboratory</a>	21% <a href="#">Artificial Intelligence Research</a>	20% <a href="#">John McCarthy</a>
28% <a href="#">Collection of Computer Science Bibliographies</a>	23% <a href="#">robotics</a>	21% <a href="#">Artificial Intelligence Depot</a>	20% <a href="#">Modern Approach</a>

**Reference Pages** 10% [Psychology Links](#) 5% [AI on the Web](#)

**Related Pages (more)**  
 100% [sigart.acm.org](#)  
 80% [The Multi-Agent Systems Lab](#)  
 75% [IEEE Computer Society](#)  
 The IEEE Computer Society is one of the major international professional bodies for IT professionals.

American Association for **Artificial Intelligence** (AAAI)

Welcome to the American Association..for **Artificial Intelligence**! Founded in 1979, the.. American Association for **Artificial Intelligence** (AAAI) is a nonprofit...aims to increase public understanding of **artificial intelligence**, improve the teaching...  
 Purpose: "Nonprofit scientific society devoted to advancing the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines."  
 Website: [Computers: Artificial Intelligence Associations](#)  
[Computers: Organizations: Associations](#)  
 WWW: aaai.org - 103k - [archived copy] - [stripped] - [older copies] - indexed: May 15 2005 - modified: Mar 31 2005

MIT Computer Science and **Artificial Intelligence** Laboratory

Computer Science and **Artificial Intelligence** Laboratory. About

Gigablast still "runs on eight desktop machines, each with four 160-GB IDE hard drives, two gigs of RAM, and one 2.6-GHz Intel processor. It can hold up to 320 million Web pages (on 5 TB), handle about 40 queries per second and spider about eight million pages per day. Currently it serves half a million queries per day to various clients, including some metasearch engines and some pay-per-click engines." We are not talking about a huge "server farm" here. Interestingly, despite keeping his search engine "small," Gigablast creator/proprietor Matt Wells says "I am a firm believer that bigger is better," and toward that end he is hoping to get the Gigablast index up to 5 billion pages. For more on Wells and Gigablast, read his interview with his former boss at Infoseek in the April 2004 edition of *AMC Queue*:

"A Conversation with Matt Wells: Steve Kirsh of Propel Software Interviews Gigablast Designer," *ACM Queue*, vol. 2, no. 2, April 2004, <http://www.acmqueue.com/modules.php?name=Content&pa=showpage&pid=135> (15 November 2006).

---

## Exalead

---

The French search engine Exalead, which introduced a new look in 2006, has features that make it worth special mention. Exalead offers both proximity searches and truncation, two options no other major search engine offers anymore. In addition, Exalead presents thumbnail images of websites in the results list (if you want them) and related search terms, directory categories, website locations, and filetypes. Exalead now claims to index more than eight billion pages. Although this is far smaller than some major search engines, it is a respectable number and one that is sure to increase.

While the new version of Exalead did away with one of its best features—the safe page preview—Exalead offers a number of other unusual or unique features designed to create a very powerful search tool:

- Exalead refreshes its index continuously, not on a schedule (this is a good thing).
- default operator is AND; users may use OR.
- Exalead does not publish a search term limit; it handled some very long searches perfectly while it had trouble with others.
- truncation, proximity, phonetic, and true wildcard searches.
- as of now, Exalead has no sponsored links.



Welcome to your exalead homepage. [Add a shortcut](#) to personalize it.



[Add more shortcuts](#) - [Hide edit buttons](#)

Notice the images below the query box. Exalead lets users put “shortcuts” here by entering a title and url for your favorite websites.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

Exalead is in the process of updating its help pages; thus far, you can find various types of help at these pages:

Exalead <http://www.exalead.com/search>

Exalead Refine Your Search <http://www.exalead.com/search/?action=kourou&id=49>

Exalead Advanced Search Help  
<http://www.exalead.com/search/?action=kourou&id=24>

Exalead Search Syntax Help  
<http://www.exalead.com/search/C?definition=querySyntaxReference>

### **Customizing Exalead Preferences**

Exalead currently offers these **Search** Preferences settings:

1. Interface Language: English, French, or German.
2. Search language: any or any combination of most languages.
3. Adult content Filtering: on or off.
4. Display: Open results and shortcuts in new window?
5. Number of search results: up to 100 for web and up to 60 for image.
6. Number of shortcuts per row: 4 up to 12.
7. Display view on results page: text only; text and thumbnail; text thumbnail and extra

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//FOR OFFICIAL USE ONLY

## The Exalead Results Page

Once you have entered your search term(s) and clicked the Exalead search button, Exalead will present you with a complex results screen. Depending on the search you are running, you will see some or all of the following for a Web search:

The screenshot shows the Exalead search interface. At the top, the search term 'phenomenology' is entered in the search box. Below the search bar, the results are displayed as a list of search results. Each result includes a thumbnail image (labeled C), a title (labeled A), a brief description (labeled B), and a directory path (labeled D). The results are:
 

- Research Phenomenology at Questia**: A search engine for research papers, journals, and articles.
- SPEP**: Professional organization supporting philosophy inspired by continental European traditions.
- Indo-Pacific Journal of Phenomenology**: A journal focusing on the supernatural, anthropology, and the limitations of scientific rationalism.
- THE WORLD PHENOMENOLOGY INSTITUTE**: World Institute for Advanced Phenomenological Research and Learning.
- Seminars on Phenomenology and Hermeneutics**: A forum for discussing the work of Edmund Husserl.

 On the right side of the page, there is a 'Refine your search' sidebar. It includes sections for 'Related terms' (F), 'Multimedia' (G), 'Languages' (H), 'Directory' (D), and 'File types' (I). The 'Related terms' section lists: Merleau-Ponty, Continental Philosophy, Edmund Husserl, Phenomenology of Spirit, and Phenomenological Research. The 'Multimedia' section lists: Audio, Video, and RSS. The 'Languages' section lists: English and German. The 'Directory' section lists: Society and Culture, Phenomenology, and Continental Philosophy. The 'File types' section lists: Acrobat (.pdf), Word (.doc), and Text (.txt).

- **A Matching Documents:** the best results for the query with the page title listed first; Exalead clusters results, showing only the “best” page for each website.
- **B Webpage Description:** a brief summary of the website.
- **C Page preview and thumbnail image:** The biggest disappointment of the new Exalead is that it no longer offers the safe page preview option for webpages. Instead it has chosen to give a thumbnail image of the cached copy of the webpage; users can click on “Preview” to see the cached copy, complete with highlighted search terms and the date cached. Fortunately, *Exalead does offer safe previewing of non-HTML file types, and this is especially*

UNCLASSIFIED//FOR OFFICIAL USE ONLY

**useful for Microsoft file types, such as Word documents and PowerPoint slides.**

The screenshot shows a search results page for 'Search Strategies - Exalead Fact Sheet'. The page includes a search bar, navigation links like 'Back to results', and a list of search results. The first result is 'FACT SHEET: EXALEAD EXALEAD URL http://www.exalead.com/ Key features • ..... 151 FACT SHEET: EXALEAD Similar pages Limit by domain/site No ... http://www.rba.co.uk/search/exaleadsumr.pdf - 22k - Add to shortcuts'. Below the search results, there is a section titled 'Search Strategies - Exalead Fact Sheet' which contains the following text:

FACT SHEET: EXALEAD

EXALEAD  
URL  
http://www.exalead.com/

Key features  
..... wildcards for stemming words pattern matching ("regular expressions") phonetic search approximate spelling search NEAR proximity operator full Boolean search thumbnails of pages displayed in results related terms and categories displayed on the results page user specified shortcuts (Smart Bookmarks) to other search engines on the home page

Search options  
Default search type Case sensitive? Wildcard/Truncation All of your words No Yes. Asterisk (\*) at the end of words, for example  
pollut\*. Also pattern matching/regular expressions for internal wildcards, for example /psych \*st/ or /mpg(1|2|3)?/  
Phrases and proximity Phrases "....." For example "climate change". NEAR operator to search for terms within sixteen words of one another. Specify maximum number of words using NEAR/n, for example climate NEAR/5 change Plus sign (+) before stop words such as "the", "of". The plus sign can also be used to disable automatic stemming if set up by the user under preferences. Minus sign (-) before the word, for example

Mandatory search terms

Exclude pages containing a term Word in the URL

branson -balloon inurl. for pages with the term in their URL, for example inurl.chocolate intitle. for pages that contain the adjacent word in the title, for example intitle:chocolate link. for example linkrba.co.uk

- **D Directory link**: opens the related categories folders from The Open Directory Project, which are also listed to the right. You can completely alter the results by selecting a different related category, e.g., in this example, *continental philosophy* instead of *phenomenology*. Clicking on "More choices" will greatly expand the related terms and related categories lists.
- **E Add to shortcuts**: selecting this link will make the current site one your shortcuts that appears on the Exalead homepage.
- **F Related Terms**: clicking on a related term runs a new search on that term and displays a new results page with new and different related terms, related categories, etc. Clicking on "More choices" will greatly expand the related terms and related categories lists.

UNCLASSIFIED//FOR OFFICIAL USE ONLY

- **G Multimedia:** selecting this option causes Exalead to restrict the search to webpages that have links to audio, video RSS content. You can select one, two, or all three multimedia options. If you click on RSS, any feeds available at any of the sites in your results' list will become visible.

The screenshot shows the Exalead search engine interface. At the top, there is a search bar with the text 'nasa' and buttons for 'Web Search' and 'Advanced search'. Below the search bar, it says 'Web Results: 1-35 of about 75 for nasa'. On the right side, there is a 'Refine your search' panel with 'Your refinements' section. Under 'Multimedia', there are three options: 'Audio' (remove), 'Video' (remove), and 'RSS' (remove). Under 'Languages', there are two options: 'English' and 'Spanish'. Below the search results, there are two search results listed. The first result is 'NASA - Ares: NASA's New Rockets Get Names' with a thumbnail image and a preview. It includes a 'Site Help & Preferences + Home + NASA Home > Mission Sections > Exploration > Spacecraft Print ... Credit: NASA + View Expanded Views of Ares-I, ...' link and a URL 'www.nasa.gov/mission\_pages/exploration/spaceraft/ares\_naming.html - 20 Jul 2006 - 5k - add to shortcuts'. Below the URL, there are three file type options: 'Audio file: RealPlayer Low (Ares logo256K\_Stream.ram) - 0.3 Kb', 'Video file: Windows High (Areslogo768K\_Stream.wmv)', and 'RSS Feed: Ares, NASA's New Rockets'. The second result is 'Matador Records | Guided By Voices' with a thumbnail image and a preview. It includes a 'The Matador records website for Guided by Voices, the label releasing all material up to 1997 and since 2002 for the band. Includes contest, recording, ...' link and a URL 'www.matadorrecords.com/guided\_by\_voices/ - 01 Feb 2006 - add to shortcuts'. Below the URL, there are three file type options: 'Directory: Arts and Entertainment > Music > ... > Guided By Voices', 'Audio file: My Kind of Soldier (gbv\_my\_kind\_of\_soldier.mp3) - 3.6 Mb', and 'Video file: gbv\_ref.mov - 0.2 Kb'. There is also an 'Unofficial RSS Feed: Matador Records' link.

- **H Languages:** limit results to a specific language.
- **I Document Type:** clicking on a specific file type will only return matching documents in that specific file type, e.g., PDF, TXT, DOC, PPT, RTF, and XLS (remember: do not open the Microsoft file types on the Internet; use the page preview option in the thumbnail image to view these files).
- **J Image Search:** Clicking on image search will automatically run the web search against the image database.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

## Exalead Basic Search

Exalead assumes as its default that multiple search terms are joined by the **AND** operator, so that a search on the keywords [windows explorer] will find all the webpages that contain ***both*** search terms. However, unlike Google, Exalead does not search first for phrases, then the terms anywhere on a webpage.

Exalead ***will not return any results*** if there is no webpage containing all the search terms. Try this query to see what I mean:

[rollerskate handshake buckyball]

However, remember you can use the **OPT** (optional) operator to make a term desirable but not required.

Unlike Google, Exalead ***does not limit the number of search terms to 32 keywords***. Exalead will try to match all the keywords you enter.

Exalead is ***not case sensitive***.

Exalead ***automatically clusters*** search results. If you want to see more pages from a specific site, the only way I know to do so now is to run a site search. For example, to see the pages at Amazon UK search for [site:amazon.co.uk].

Exalead permits the use of ***the OR operator*** in simple search. The **OR** must be capitalized.

Exalead recognizes ***double-quotes*** as enclosing a phrase.

Exalead ignores certain ***stop words***, i.e., when searched alone or with other stop words. If you include a stop word such as *a*, *an*, *the*, *in*, or *be* in a search, Exalead searches for it. If you need to search for stop words by themselves or with other stop words, you must either enclose them in double-quotes or put the plus sign (+) in front of them. Compare [to be or not] to ["to be or not to be"] and compare [fire and ice] to ["fire and ice"].

Using the ***minus sign (-)*** in front of a keyword ensures that Exalead excludes that term from the search. For example, the results for the search [phenomenology -philosophy] are very different from the results for [phenomenology].

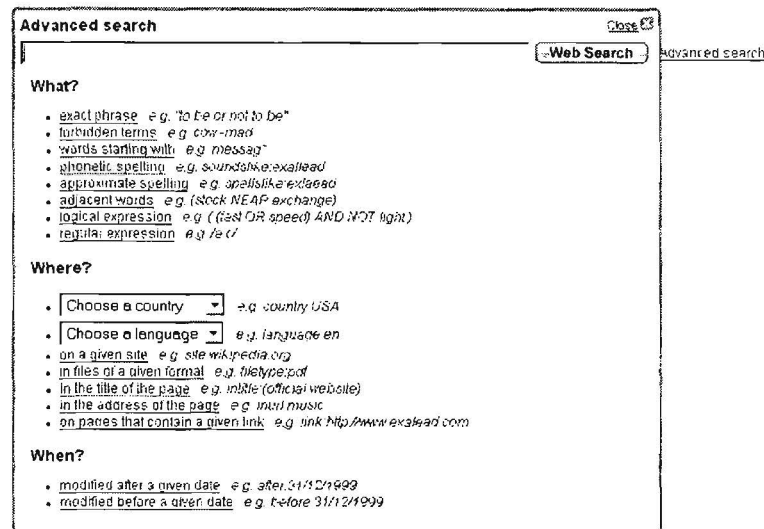
~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

## Exalead Advanced Search

Exalead has a unique and very appealing way of presenting advanced search features. Clicking on the “Advanced search” link on the main page brings up a window that displays and explains the advanced search options. In every case, these options work in the simple search screen by using the correct syntax.

exalead



Two features Exalead offers that have almost vanished from search elsewhere are *proximity searches* and *truncation/wildcards*.

Exalead's proximity search uses **NEAR**. The default setting is for *terms that are within sixteen terms of each other*, but users can change the proximity by adding a number, e.g., [empire NEAR/5 building]. With the NEAR operator, order is almost irrelevant as this query demonstrates. A query using the name of an 18th Century French foreign minister, Charles Jean-Baptiste Fleuriau, comte de Morville, shows how the NEAR operator works: the query [comte de Morville NEAR Fleuriau NEAR Charles NEAR Jean-Baptiste] finds any indexed page containing all these terms within sixteen words of each other, regardless of the order in which they appear either in the query or in the text.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~



UNCLASSIFIED//FOR OFFICIAL USE ONLY

Web Images Sign in | Preferences

comte de Morville NEAR Fleuriau.NEAR Charles NEAR J Web Search advanced search

Web Results: 10 of about 193 for comte de Morville NEAR Fleuriau NEAR Charles NEAR Jean-Baptiste

Did you mean: [comet de Morville NEAR Fleuriau NEAR Charles NEAR](#)

View: [HTML](#) [PDF](#) [Image](#)

**Related terms**

- Jean-Baptiste
- Minister to France
- Louis Michel
- Pierre Henn
- Tout d'Auvergne

**Languages**

- English
- French

[More choices >](#)

**MORVILLE (Charles-Jean-Baptiste de Fleuriau d'Armenonville, comte ...**  
 140 ministers of Foreign Affairs' Gallery - Archives diplomatiques - Ministère des Affaires  
 Étrangères ... MORVILLE (Charles-Jean-Baptiste de Fleuriau ...  
[www.diplomatie.gouv.fr/archives\\_gb/dossiers/140ministers\\_gb/louis15a03.html](#) - 15 Sep 2004 - 2k - [Add to shortcuts](#)

**Charles Jean Baptiste Fleuriau de Morville - Wikipédia**  
 Charles Jean Baptiste Fleuriau de Morville Un article de ... Fleuriau d'Armenonville,  
 comte de Morville est un homme d'État français né à Paris le 30 octobre ...  
[fr.wikipedia.org/wiki/Charles\\_Jean\\_Baptiste\\_Fleuriau\\_de\\_Morville](#) - 22 Jul 2006 - 14k - [Add to shortcuts](#)

**Jean Baptiste de Boyer, Marquis d'Argens - explanation-Guide.info**  
 August 1723 Charles Jean-Baptiste Fleuriau, comte de Morville 16 August 1723 19 August  
 1727 ... ... Églises Baptistes de la RCA) Baptist Community of ...  
[explanation-guide.info/meaning/Jean-Baptiste-de-Boyer,-Marquis-d'Argens.html](#) - 18k - [Add to shortcuts](#)

**Minister of Foreign Affairs (France): Facts and details from ...**  
 Encyclopedia subject: Minister of Foreign Affairs (France) ... February 1680 28 July 1696  
 Jean-Baptiste Colbert, Exception Handler. No article summary ...  
[www.absoluteastronomy.com/ref/minister\\_of\\_foreign\\_affairs\\_france](#) - 13 Mar 2006 - 75k - [Add to shortcuts](#)

Also, the presence or absence of parentheses does not appear to affect the NEAR search. **Proximity operators can be extremely useful in finding pages with search terms that may not be in a precise order while excluding a lot of irrelevant hits.**

Exalead **supports both limited and true wildcard** searching.

Exalead supposedly offers both **automatic truncation** (word stemming) and the **wildcard**, which are welcome features discarded by other search engines. As of now, Exalead is the only major search engine to offer truncation or a wildcard. On a search with two or more words, stemming is supposed to be automatic. However, I find that the automatic truncation feature is so capricious as to be useless: sometimes it works, usually it doesn't. In a search for [child play toy], Exalead does not find *children*, *plays/played/playing*, or *toys*.

However, when I search on [child\*], Exalead will return pages with *children* highlighted as a search result. The wildcard also can be **used inside a search term**, e.g., [kazak\*stan]. However, this search will also find *kazakh* and *kazak* as well as *kazakstan* and *kazakhstan*. The wildcard option is listed in the Advanced search window as **words starting with**, but keep in mind the asterisk can be used inside words as well.

Exalead has a number of other interesting features. For example, in the advanced search window, users can choose among these search method options: **exact**

UNCLASSIFIED//FOR OFFICIAL USE ONLY

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

*search*, *forbidden terms*, *phonetic search*, and *approximate spelling*. Exact search is what you would expect, i.e., phrase searching inside double-quotes. "Forbidden terms" is a different way of saying NOT or using the minus sign.

The *phonetic search* sounds great, but I am often frustrated by it because so many websites misspell so many words, Exalead is going to find those misspelled words first (try: [geneology] to see what I mean). However, the phonetic search successfully figured out that [criptografy] meant [cryptography]. The phonetic search has genuine utility.

The screenshot shows the Exalead search engine interface. At the top, there is a search bar with the text "soundslike:criptografy" and a "Web Search" button. To the right of the search bar are links for "Sign in" and "Preferences". Below the search bar, the results are displayed. The first result is titled "CS851/551: Cryptography Applications Bistro" and is described as the homepage for a seminar offered during Spring 2004 at the University of Virginia. The second result is titled "Cryptography And Network Security" and is described as a site about cryptography and network security. The third result is titled "CryptoSys cryptography software tools for Visual Basic and C/C++ ..." and is described as software tools for developers. On the right side of the results, there is a "Refine your search" section with various filters: "Related terms" (Public key cryptography, Strong cryptography, Quantum cryptography, Applied Cryptography, Elliptic curves in cryptography), "Multimedia" (Audio, Video, RSS), "Languages" (English, German), "Directory" (Computers, Science and Environment, Math), and "File types" (Acrobat (.pdf), Text (.txt), Word (.doc)).

The *approximate spelling* option can be similarly frustrating. A search on [programme] will find a few sites containing *program*, *programmen*, or *programs*, but usually the results are for the actual term searched, in this case [programme]. However, it worked very well with [colour], finding a good mix of *color* and *colour* and the approximate search on [geneology] found *genealogy*.

What I like much, much better is Exalead's *regular expression patterns* option, which amounts to a *true wildcard search*. Here's how it works:

Use a forward slash (/) at the beginning and end of the term; use a period (.) to indicate one missing term; if you are not sure how many letters are missing, use the wildcard (\*) after the period. For example, the query [/crypt.\*c/] will find *cryptographic* and *cryptologic*:

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

Web | Images

Sign in | Preferences

Exalead

Web Search Advanced search

Web

Results 1-100 of about 1,269,173 for 'crypt.\*/'

**C Products**  
www.ebayexpress.com - Get it new on eBay Express. Happy Shopping.

**West Memorials - Headstones & Markers**  
www.westmemorials.com - Crypt - we offer affordable cemetery markers, headstones, monuments and cemetery memorials. Free ...

**International Association for Cryptologic Research**  
further research in cryptology and related fields. ... Crypto 2007, August 19-23, 2007, Santa Barbara, California, USA. ... Workshop on Cryptographic Hardware ...  
www.iacr.org - 4k - [load to site faster](#)

**Directories:**

- Science and Environment > Math > ... > Communication Theory > Cryptography
- Science and Environment > Math > Organizations
- Computers > Hacking > Cryptography

**CryptoLogic Inc - software development company specializing in ...**  
CryptoLogic is an Internet software development company with leading proprietary technologies that enable secure, high-speed financial transactions over ...  
www.cryptologic.com/ - 12 Oct 2005 - 13k - [load to site faster](#)

**Directories:**

- Computers > Software > Business and Economy > E-Commerce (Toronto)
- Business and Economy > Computers and Internet > Software (Toronto)

**Refine your search**

**Related terms**

- Cryptographic algorithms
- Cryptographic software
- Cryptographic keys
- Cryptographic protocols
- Cryptographic systems

**Multimedia**

Audio Video RSS

**Languages**

- English
- German

**Directory**

- Computers
- RFCs
- Science and Environment

**File types**

- Acrobat (.pdf)
- Text (.txt)
- Word (.doc)

More choices >

Here are the basic rules for pattern matching (wildcard) searches:

The first character is always a **slash ( / )**. This tells Exalead a special pattern will follow.

Within the pattern, the **period ( . )** is a special character that can represent any character.

The **asterisk ( \* )** stands for character repetition, i.e., any number of characters.

The **pipe ( | )** stands for OR, and **parentheses** are used to group characters.

A **question mark ( ? )** is placed at the end of a character group to make that group optional.

The last character is always a **slash ( / )**. This tells Exalead this is the end of the query.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED#FOR OFFICIAL USE ONLY~~

In this example—[/mpg(1|2|3)?/]—Exalead will search for any page containing the query term *mpg* and 1 or 2 or 3. It will also find pages containing only *mpg* because the ? makes the 1, 2, and 3 optional. Without the ? Exalead will only find pages containing *mpg1*, *mpg2*, or *mpg3*.

Exalead will handle **complex boolean queries** in the simple search screen or from the Advanced search window. The boolean operators Exalead supports are AND, OR, NOT or AND NOT (in caps). A typical boolean query would be:

[(baseball OR football) NOT cardinals]

In addition, there are two other operators that can be used in a boolean query: NEAR and OPT. NEAR finds search terms within 16 words of each other and OPT makes a query term preferable but does not require it. For example:

[(football NEAR cardinals) OPT "st louis"]

This is nice to know because most search engines use AND as their default, and will not return results unless all terms are found. Check the difference between the results for these two searches in Exalead: [buckyball skateboard OPT flyswatter] and [buckyball skateboard flyswatter].

Exalead will search in all or one of most **languages**. Use either the syntax *language:* followed by the language digraph or the pulldown menu in the Advanced search window. Also, Exalead offers a country search option either from the Advanced search window or using the syntax *country:* followed by the country digraph.

Exalead does not recognize **diacritical marks at this time**. This means that a search on [façade] finds both *façade* and *facade*. However, Exalead will handle some **non-Latin character sets**. Exalead officially supports Unicode (UTF), Windows encodings, and miscellaneous encodings (Arabic, Chinese, Korean, Japanese, and Russian).

Exalead offers **limited field searching**, i.e., special search terms to restrict searches and make them more effective. These special operators can be used in both simple search and in the Advanced search window.

- **language:** restricts results to pages in a specific language. The language syntax uses the obsolete two-letter ISO language codes (639-1). Must be used with additional keywords.

#### **Advanced Search > Where? > Choose a language**

Example of how to use the **language:** command:

[language:de welt] finds all the pages indexed by Exalead that are written in

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

German and contain the keyword “welt,” which has a very different meaning in German than in English.

- **country:** restricts results to pages in a specific country. The country syntax uses the two-letter ISO country codes. Must be used with additional keywords.

#### **Advanced Search > Where? > Choose a country**

Example of how to use the **country:** command:

[country:de wissenschaft] finds all the pages indexed by Exalead that are purportedly in Germany and contain the term “wissenschaft.” It will not limit the search to the German TLD “de.”

- **site:** restricts results to a specific website or domain, *excluding* specific top-level domains. You must search on a second-level domain for site to work. May be used with or without keywords.

#### **Advanced Search > Where? > on a given site**

Examples of how to use the **site:** command:

[site:amazon.com] finds www.amazon.com, cards.amazon.com, www.amazon.com/dvd/. However, it will not find www.amazon.com.br.

[site:ir] *does not* find the pages from the Iranian (.ir) top-level domain. However, [site:gov.ir] does find all the pages from the Iranian government domain indexed by Exalead.

[site:federalreserve.gov “statistical data”] finds all the pages at the Federal Reserve website indexed by Exalead containing the phrase statistical data.

- **filetype:** restricts results to PDF, MS Word, and other filetypes. May be used with or without keywords. **Exalead converts these other types of files to HTML, making them safe to view.** Select **PREVIEW** to see the HTML version.

#### **Advanced Search > Where? > in files of a given format**

To search by specific type of file, use the syntax **filetype:** plus one of these abbreviations:

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

html or htm	standard webpage
pdf	Adobe Acrobat
xls	MS Excel Spreadsheet
ppt	MS PowerPoint
doc	MS Word
wpd	Corel WordPerfect versions 6 & 7
rtf	Rich Text Format
swf	Macromedia Flash text & hypertext link
txt	text

Examples of how to use the **filetype:** command:

[filetype:xls] finds all pages indexed by Exalead that are in Excel spreadsheet format.

[filetype:pdf "white paper"] finds all pages indexed by Exalead that are in PDF format and contain the phrase "*white paper*" anywhere in the text, title, or url.

- **intitle:** restricts results to pages containing a specific word or phrase anywhere in the webpage's title, which usually appears in the browser's title bar and is the HTML <title> tag. May be used with or without additional keywords.

**Advanced Search > Where? > in the title of the page**

Examples of how to use the **intitle:** command:

[intitle:amazon] finds all pages that include the word *amazon* in their title

["rain forest" intitle:amazon] finds all pages that include the word *amazon* in their title and mention the phrase "*rain forest*" anywhere in the document (title or text or anywhere in the document)

- **inurl:** restricts results to pages containing a specific word or phrase *anywhere* in the webpage's url, that is, the webpage address. May be used with or without additional keywords.

**Advanced Search > Where? > in the address of the page**

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

Examples of how to use the **inurl:** command:

[inurl:amazon] finds all pages that include the word *amazon* anywhere in their url.

["cosmic ray" inurl:spacecraft] finds all pages that include the exact phrase "*cosmic ray*" anywhere in the document (title or text or anywhere in the document) and include *spacecraft* anywhere in the site's url.

- **link:** restricts the results to documents that have links to a specific website. Will work without the full url (absent the http://) but the preferred syntax is [link:http://www.domain.com]. Also, the link: command does not work beyond the top level of a site, so the query [link:www.noaa.gov/wx.html] treats the "wx.html" as a keyword. May be used with or without keywords.

**Advanced Search > Where? > on pages that contain a given link**

Example of how to use the **link:** command:

[link:http://www.noaa.gov] finds all pages linking to the NOAA homepage.

[link:http://www.noaa.gov drought] finds all pages linking to the NOAA site that contain the keyword *drought*.

## Exalead Search Services and Tools

Exalead does not offer any special services or tools such as news, maps, reference tools, except for a browser toolbar that works with both Internet Explorer and Firefox. At present, the two types of specialized Exalead search are the multimedia (audio, video, and RSS) refinement option and image search.

Image Search: Exalead offers some nice options with its image search. You can look for images of specific sizes (small, medium, large), computer wallpaper by resolution, image color, layout, or filetype. Exalead's advanced search options work in image search as well.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//FOR OFFICIAL USE ONLY

lead [Web](#) [Images](#) [Sign in](#) | [Preferences](#)

[Image Search](#) [Advanced search](#)

Results 1-24 of about 19,307 for cassini

Background:

**Start your search**

Size: [Small](#), [Medium](#), [Large](#)

**Wallpapers**

- [800x600](#) (93)
- [1024x768](#) (54)
- [1280x1024](#) (16)
- [1600x1200](#) (8)
- [1920x1200](#) (1)

**Image color**

- [Color](#) (67%)
- [Grayscale](#) (32%)
- [Black & White](#) (0.1%)

**Layout**






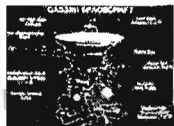


- [Landscape](#) (78%)
- [Portrait](#) (21%)

**File types**

- [Jpeg](#) (82%)
- [Gif](#) (17%)
- [Png](#) (0.6%)

Search within results

[Less choices](#)

 <p><b>Cassini I</b> Cassini I Bild... 139 x 198 - 22.2 Kb - gif www.geophys.lu-bis.de</p>	 <p><b>Cassini picture</b> 213 x 150 - 13.7 Kb - gif s.amadhl.jpl.nasa.gov</p>	 <p><b>Cassini launch</b> 200 x 145 - 8.8 Kb - jpeg edition.cnn.com</p>	 <p><b>The Cassini ...</b> 200 x 148 - 8.1 Kb - jpeg edition.cnn.com</p>
 <p><b>Cassini Resources at ECSL</b> 400 x 287 - 16.5 Kb - jpeg www.ecsl.su.se</p>	 <p><b>Cassini</b> 250 x 187 - 10.6 Kb - jpeg www.askusurf.com</p>	 <p><b>ISSUE ON CASSINI</b> 631 x 480 - 57.4 Kb - jpeg www.globe.net</p>	 <p><b>Giovanni Domenico Cassini</b> 423 x 588 - 25.7 Kb - jpeg www.sterrenkunde.nl</p>

**The Bottom Line**

Exalead is not in the Google and Yahoo class yet, but because it offers unique and important features dealing with truncation, wildcards, proximity searching, etc., **it is one of the top-tier search services**. In addition, Exalead offers the option to preview non-html files (e.g., Microsoft file types) safely, which is extremely important given the security dangers that plague Internet users. Exalead is a valuable addition to the world of Internet search.



---

## Ask

---

During 2006 Teoma and Ask Jeeves ceased to exist as separate search sites and merged under the Ask.com umbrella. I had never been impressed with Ask Jeeves, which was one of the few sites that continued to try to respond to users questions, though not very successfully. Teoma was always an “also ran” in the world of search. However, when Barry Diller, former Chairman and CEO of Paramount Pictures and Fox, Inc.’s, and his IAC/Interactive Corp. acquired Ask Jeeves this year, things changed dramatically. The name was shortened to Ask, the annoying butler icon was gone, along with the ubiquitous ads and usually unfulfilled promise of answers to natural language queries. Ask incorporated Teoma’s search algorithm, ExpertRank, and the Teoma site went away. Now, Ask.com has become a major player.

One of the most striking differences is obvious as soon as you run a search. Instead of a list of sponsored links, which Google, Live Search, and Yahoo all display, Ask shows “zoom related search” links, designed to help users either narrow or expand a search. Of course, Ask still serves up ads with its search results, but the search company is putting the primary focus on free search results and not on sponsored results.

### Customizing Ask’s Settings



























Ask offers six general **Settings**:

1. Locations: you may enter a specific location, including a street address or a city, state, and zip code for the US. This is an optional feature and you can sign up for an account if you want to enter multiple locations. This information is used to provide tailored search results relevant to your location.
2. Displaying results: Ask lets you see as few as 10 and up to 100 results per page. There is also an option to open results in a new window.
3. Content filtering: Unlike most search engines, Ask automatically filters adult content; the two options are to alert the user when content is filtered and provide a link to it or to minimize adult content and not link to it.
4. Interface language: if you are more comfortable working in another language, Ask can display in dozens.
5. Make Ask your Default Search Engine: In this case, you are telling your browser to use Ask as the default search engine from the browser address bar.

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

- 6. Default Ask Site: You can chose one location from a list including the US, France, Germany, Italy, Netherlands, Spain, and UK, or no default site. Your results will vary depending on the default site.

The other setting option is similar to Yahoo's feature that lets users edit the search tools. Here are the options Ask offers; you can select only the ones you want to appear on your Ask main search page.

<b>Search Tools</b> 	<b>Search Tools</b> 	<b>Search Tools</b> 
 Web	 Advanced Search	 Toolbar
 Images	 Bloglines	 Unit Conversion
 News	 Currency Conversion	 White Pages
 Maps & Directions	 Desktop	
 Local	 Mobile Content	
 Weather	 Movies	
 Encyclopedia	 MyStuff	
 Ask for Kids	 Shopping	
 Dictionary	 Stocks	
 Blogs & Feeds	 Thesaurus	
Edit                      Next »	Edit                      « Back Next »	Edit                      « Back

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

### The Ask Results Page

Once you've entered your search terms and selected the Ask Search button, Ask will present you with a list of results (hits). For each result returned you may see:

The screenshot shows the Ask.com search results for the term "cardinal". At the top, there is a navigation bar with "Web", "Images", "News", "Blogs & Feeds", "Shopping", and "More" links, along with a search box containing "cardinal" and "Search" and "Advanced Search" buttons. The search results are displayed in a list format. The first result is from an encyclopedia, titled "Encyclopedia: Cardinal (bird)", with a small image of a cardinal bird. The text describes the family of passerine birds. Other results include "Cardinal Health", "Northern Cardinal", and "Illinois State Bird". On the right side, there is a "Narrow Your Search" section with a list of related terms like "Cardinal Birds", "Northern Cardinal", "Red Cardinal Bird", etc. The bottom of the screenshot shows the search bar for "rwanda" and the top of the results for "rwanda", including a result for "Rwanda" with details like capital, population, and location.

- **A Smart Answers:** Ask's best guess about what you want, Smart Answers provides quick access to encyclopedias (Wikipedia, Houghton Mifflin, or Columbia), weather, dictionary results, translations, conversions, etc. Note that "other matches" will try to disambiguate a search term with multiple meanings such as [cardinal]. This is an extremely useful way to find information about commonplace topics, such as [Rwanda]:

The screenshot shows the Ask.com search results for the term "rwanda". At the top, there is a navigation bar with "Web", "Images", "News", "Blogs & Feeds", "Shopping", and "More" links, along with a search box containing "rwanda" and "Search" and "Advanced Search" buttons. The search results are displayed in a list format. The first result is from an encyclopedia, titled "Rwanda", with a small image of the Rwandan flag. The text provides details about the country, including its capital (Kigali), population (8,440,820), location (Central Africa, east of Democratic Republic of the Congo), and chief of state (President Paul Kagame). Other results include "Rwanda Genocide", "Facts about Rwanda", "Rwanda People", and "Rwanda Map".

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

- **B Webpage Title & Description**: the title and a brief summary of the website.
- **C Binoculars Site Preview**: Ask's Binoculars Site Preview are periodic screen captures of the browser navigating a page. To view the site preview, users should only move the mouse over the binoculars because clicking on the binoculars takes you to the site. The mouseover is of a static image, so it is safe to view, but I find it too small to be very useful beyond revealing the general nature of a site.
- **D Cached**: a link to the version of the site stored by Ask with the date and time the page was indexed.
- **E Save**: Ask offers this service for web and image searches. When users click on a "save" link on either a web or image search, for web searches Ask will save the title of the result, the url, the description, the binoculars icon, and the query used to find that result. For image searches, Ask will save the name and location of the picture, as well as the query used to find the image. Also, everything saved is fully searchable so all saved content is easy to find again later. However, for the save feature to work properly, users need to allow search history to be enabled (the default). *If you do not want Ask to save your search history, go to My Stuff | Settings and uncheck "Record all my searches into my 'Search History.'"*
- **F Zoom Related Search**: This is a popular feature retained from Teoma that helps users either narrow or broaden a search "with possible alternative search terms which appear on the right hand side of the Ask results page.
  - **Narrow Your Search**: helps you to drill down into topics that are specifically related to your search
  - **Expand Your Search**: allows you to explore topics that are conceptually related to your search
  - **Related Names**: presents a list of names that are conceptually tied to topic options within the 'Narrow Your Search' and 'Expand Your Search' lists.<sup>59</sup>
- **G More Search Types**: Selecting any of these other search options causes Ask to search automatically for images, news stories, blog entries, etc., with your search term(s).

---

<sup>59</sup> Ask.com Site Features, "Zoom Related Search,"  
[http://help.ask.com/en/docs/about/site\\_features.shtml#relatedsearch](http://help.ask.com/en/docs/about/site_features.shtml#relatedsearch) (14 November 2006).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~**Ask Basic Search**<http://www.ask.com/>

Ask assumes as its default that multiple search terms are joined by the **AND** operator, so that a search on the keywords [windows explorer] will find all the webpages that contain **both** search terms.

Ask **will not return any results** if there is no webpage containing all the search terms. Try this query to see what I mean:

[kong spektioneer synecdoche]

Ask **does not appear to limit the number of search terms**.

Ask is **not case sensitive**. There does not appear to be anything you can do about this.

Ask does not offer **word stemming** or **truncation**, i.e., searching for variations of search terms. Ask searches for exactly the term as you enter it, e.g., a search for [window] will not search for [windows].

Ask automatically **clusters search results**. Multiple hits from the same site are indented and there is usually an option to see more results from a specific site.

Ask **permits the use of the OR operator** in simple search. The OR needs to be capitalized.

Beyond the use of the OR operator in its simple search, **Ask does not support boolean search**.

Searchers can delimit phrases using double-quotes. For example, if I search on:

[the last king of france]

without double-quotes, Ask will ignore the "the" and the "of" in its search. I noticed that the results from this search are more relevant than the ones I received from Google for the same query. If I enclose the same query in **double-quotes**, Ask will search on exactly the phrase ["the last king of france"], and the first hit links to a site that lists all the Kings of France, where Louis XVIII can be readily identified. Enclosing searches in double-quotes is much more effective for finding precise results than relying on automatic phrase searching.

Ask appears to ignore **stop words** outside double quotes only when other search terms are used. These two searches will return identical results:

[the last king of france] [last king france]

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

However, if I search for [the], Ask returns over 2 billion hits. If I add another search term, e.g., [the france], that query is identical to searching for [france], so the stop word is ignored. Nonetheless, it appears that if you search only for stop words, Ask will find pages containing them all, e.g., [i a an the].

Ask does not seem to like the plus sign (+) because it returns an error message when I try to use it. By default Ask searches for all keywords except stop words. However, there are many times when searchers need to exclude certain terms that are commonly associated with a keyword but irrelevant to their search. That's where the *minus sign* (-) comes in. Using the minus sign in front of a keyword ensures that Ask excludes that term from the search. For example, the results for the search ["pearl harbor" -movie] are very different from the results for ["pearl harbor"].

Ask treats most *punctuation marks* the same way, as links in a search string. For example, Ask handles a search for [c-span], [c.span], ["c span"], and [c?span] basically the same way. However, a search for [cspan] with no space or mark is treated differently.

## Ask Advanced Search

Ask has a number of "query modifiers" to restrict searches and make them more effective in many cases. These query modifiers can be used in simple search using the following syntax or on the advanced web search page using the appropriate menu options. Interestingly, Ask using the "must exclude" minus sign differently from other search engines: the minus sign goes after the command syntax, for example, [inurl:nasa site:-gov]

The query modifiers Ask supports are:

- site: restricts results to websites in a given domain. *This syntax requires a keyword.*

### **Advanced Web Search > Domain or Site**

Examples of how to use the site: command:

[shuttle site:www.nasa.gov] finds pages about the space shuttle at the NASA website.

["bulletin officiel" site:fr] finds pages in the French top-level domain about official bulletins.

["bulletin officiel" site:-fr] finds pages containing the phrase "bulletin officiel" that are not in the French top-level domain. Note that the minus sign goes after the site: syntax.

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

- **title:** or **intitle:** restricts the results to documents containing the keyword in the title.

### Advanced Web Search > Location of words or phrases > In page title

#### Examples of how to use the **title:** command:

[title:amazon] finds all pages that include the word *amazon* in their title

[intitle:amazon jungle rainforest] finds all pages that include the words *amazon*, *jungle*, and *rainforest* in their title. Using **intitle:** makes this search function the same as Google's `allintitle: query`. Note: use a hyphen to search for phrases using the intitle: syntax because the double-quotes do not work.

[-books title:amazon] finds all pages that contain *amazon* in the title and do not contain the term *books* anywhere on the page. Note that you must put the excluded term before the intitle: syntax.

[title:galileo site:-nasa.gov] finds all pages that contain the term *galileo* in the title but are not at any *nasa.gov* website.

- **inurl:** restricts the results to documents containing the keyword in the url.

### Advanced Web Search > Occurrences

#### Examples of how to use the **inurl:** command:

[inurl:nasa] finds all pages that include *nasa* anywhere in the url (address)

[inurl:nasa site:-gov] finds all pages that include *nasa* anywhere in the url of sites that are *not* in the *.gov* top-level domain. Note that the minus sign goes after the site: syntax.

[inurl:shuttle inurl:-nasa] finds all pages that include *shuttle* in the url but exclude *nasa* from the url. Note that the minus sign goes after the site: syntax.

[inurl:nasa shuttle sts-90] finds all pages that include both *nasa* and *shuttle* in the url of a site. Used this way, Ask's `inurl:` command functions the same as Google's `allinurl:` command, that is, all terms must be in the url.

[-shuttle inurl:nasa] finds all pages with *nasa* in the url but do not include the term *shuttle* anywhere on the page. Note that you must put the excluded term before the intitle: syntax.

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

## Ask's Services and Specialty Searches

Ask offers a number of special features designed to help users find specific kinds of information faster and more easily.

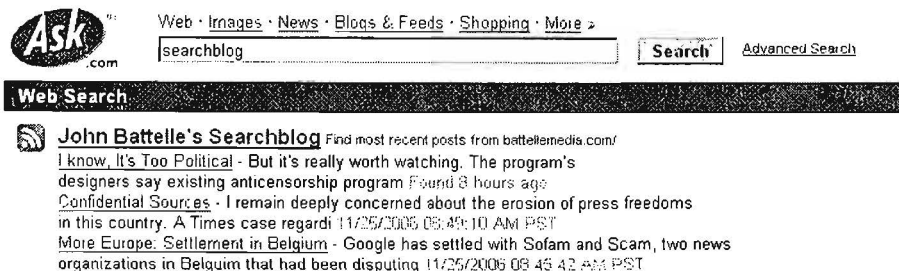
**Blog Search:** Ask is partnered with Bloglines, the most popular (and my favorite) RSS feed reader, to create blog and RSS feed search. The blog search options are:

- sort by date, popularity, or relevance (which combines date and popularity).
- sort by posts, feeds, or news.
- binoculars preview last five posts from a feed.
- options to subscribe and/or post to a feed using several different applications.

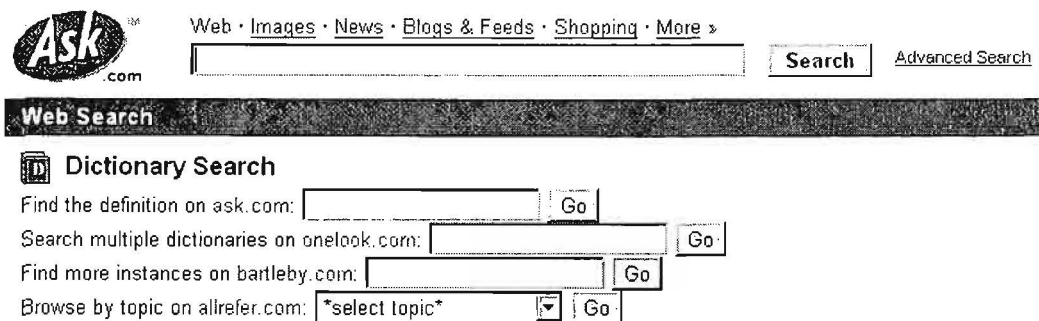
The screenshot displays the Ask.com interface for 'Blogs & Feeds'. At the top, there are navigation links for 'Web', 'Images', 'News', 'Blogs & Feeds', 'Shopping', and 'More'. A search bar contains the text 'mashup'. Below the search bar, there are buttons for 'Search Blogs', 'Search the Web', and 'Advanced Search'. The main content area is titled 'View Posts, Feeds, or News Results' and shows search results for 'mashup'. A specific result for 'UK Google Maps Mashup Roundup Part 1/2' is highlighted. A callout box labeled 'Binoculars preview last 5 posts' points to a map preview of the mashup. Below the main content, there are pull-down menus for social media sharing, including 'MY AOL', 'digg', 'del.icio.us', 'reddit', and 'YAHOO!'. A callout box labeled 'Subscribe or Post to Blogs from pull-down menus' points to the 'MY AOL' and 'digg' options.



- **RSS Answers** will display the three most recent entries at a blog. Obviously, only a limited number of blogs work in RSS Answers, but it is a quick way to see what is new at your favorite blog site. Here is an example of an RSS Answers for John Battelle's Searchblog:



**Definitions:** Ask will present a dictionary or encyclopedia definition of a term if you phrase the query as [define keyword], [definition of keyword], [the meaning of keyword], or [dictionary], which brings up the Dictionary Search option:



**Local Search:** search for services or businesses by US zip code or city.

**Maps:** to map a US or Canadian location, search on the street address, city and state or the word *map* and a location. Some international maps are now available. See the section on maps for details.

**News:** links to news stories appear when a search term matches current news stories. Sort news by date or relevance. A separate Ask News page is available at <http://news.ask.com/>

**Travel Shortcuts:** To find arrival and departure information, flight delays, airport status, and weather conditions at a US or Canadian airport, enter the airport's three-letter code and the word *airport*. For example, to information about Baltimore-Washington International, enter [bwi airport].

**White Page Search:** search for US phone numbers and addresses for people, businesses, government offices, doctors, and schools in the U.S.



**Image Search:** the Ask image search uses “authoritativeness” to rank its results and also accesses a proprietary image index. ***It is one of the best image search tools available.*** The image search appears as one of the default search tools on the right-hand side of the main search page. There is no advanced image search and no special image search options. However, when you search for an image, zoom related search terms to expand or narrow the search appear. If you select the “save” option, this link will save the image to your personal “stuff,” which can later be accessed via <http://mystuff.ask.com/>. If you select “info” about an image, you will then see detailed information about the image, including copyright information, and its source homepage will appear in a frame in the bottom portion of the screen.

The screenshot shows the Ask.com search interface. At the top, there are navigation links for 'Web', 'Images', 'News', 'Blogs & Feeds', and 'Shopping'. The search bar contains the text 'define ontology'. Below the search bar, there are two tabs: 'Search' and 'Advanced Search'. The search results are displayed in a list format. The first result is 'Definitions of 'ontology'' with a brief description. Below this, there is a 'Free Online Dictionary' link. The 'define ontology' link is highlighted with a red box, and a red arrow points from it to the 'define ontology' link in the search results. A red circle highlights the 'define ontology' link in the search results, and a red arrow points from it to the 'define ontology' link in the search results. A red arrow points from the 'define ontology' link in the search results to the 'define ontology' link in the search results.

Ask Image Search “Info” Page

<http://pictures.ask.com/>

**Number Search:** Ask offers many types of number searches. The numbers Ask will search for are:

- **UPS tracking:** enter the UPS tracking number [1Z9999X999999999], or enter [ups tracking] to bring up the UPS tracking query option.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

- **USPS** tracking: enter USPS plus the tracking number with or without spaces [usps 99999999999999999999], or enter [usps tracking] to bring up the USPS tracking query option.
- **FEDEX** tracking: enter FEDEX plus the tracking number [fedex 99999999999999999999], or enter [fedex tracking] to bring up the FEDEX tracking query option.
- **DHL and Airborne Express** tracking: enter DHL plus the tracking number [DHL 9999999999], or enter [DHL tracking] to bring up the DHL tracking query option.
- **ZIP codes:** enter a US ZIP code, either five or nine digits
- **ISBN:** enter any International Standard Book Number
- **VIN** Information: to find information about a vehicle's history, search on its 17-character Vehicle Identification Number (VIN)

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

---

## More Help: Internet Guides and Tutorials

---

For anyone who wants additional help in learning how to use the Internet more effectively, many excellent resources are available for free via the Internet. Also, there are more and more sites appearing to help new Internet users get started with searching the web. Some help you choose the right search engine, others how to formulate a query, and others are step-by-step tutorials.

The Internet Detective Tutorial is a free online tutorial that is part of the Intute: Virtual Training Suite, a set of "free Internet tutorials to help you learn how to get the best from the Web for your education and research...[created by] a national team of subject specialists based in universities and colleges across the UK."<sup>60</sup> Not familiar with Intute? It is the newly evolved face of the Resource Discovery Network, a carefully selected and evaluated set of academic research resources. The Internet Detective tutorial focuses on how to evaluate Internet sources for quality and authoritativeness, how to avoid wasting time on questionable websites and searches, and how to avoid violating copyright laws and plagiarism. The tutorial includes a set of practical exercises to try your Internet research skills. Although the tutorial is aimed at university research, I highly recommend it for all readers. The tutorial requires about an hour to complete, but it is designed so you can do it in more than one sitting.

The Internet Detective Tutorial <http://www.vts.intute.ac.uk/detective/index.html>

All the Intute tutorials are available at:

Intute: Virtual Training Suite <http://www.vts.intute.ac.uk/>

The following are tutorials, guides, and search-oriented sites available on the Internet:

BrightPlanet's Guide to Effective Searching of the Internet

<http://www.brightplanet.com/deepcontent/tutorials/search/index.asp>

Finding Information on the Internet: A Tutorial

<http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html>

Internet Tutorials from University of Albany Libraries <http://www.internettutorials.net/>

Internet Scout Report

<http://scout.wisc.edu/Projects/PastProjects/toolkit/searching/index.html>

---

<sup>60</sup> Intute: Virtual Training Suite, <<http://www.vts.intute.ac.uk/>> (12 September 2006).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

Pandia's Goalgetter <http://www.pandia.com/goalgetter/index.html>  
Phil Bradley's Searching the Internet <http://www.philb.com/searchindex.htm>  
Search Engine Watch Tutorials (old but still useful)  
<http://www.searchenginewatch.com/resources/article.php/2156611>  
Web Search Guide <http://www.websearchguide.ca/tutorials/tocfram.htm>

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

---

## Specialized Search Tools & Techniques

---

This section, which first appeared in the 2006 edition, was born of the rapid growth of both unconventional search techniques such as Google hacking and the wildfire spreading of such tools as online maps. This year, I have added a new section on Wikipedia and expanded the maps and mapping section.

---

### “Google Hacking”

---

This topic has received a great deal of attention in the world of Internet search in the past few years. While this activity is generically referred to as “Google hacking,”<sup>61</sup> this is a double misnomer. First, to limit this practice to “Google” is a mistake because many of these kinds of searches can be run using any search engine, though they are clearly going to be most effective with a large, powerful search tool that offers many search options, such as Google. Second, this is not hacking in the sense that most people use the term, i.e., gaining access to a computer or data on a computer illegally or without authorization. Nothing I am going to describe to you is illegal, nor does it in any way involve accessing unauthorized data. **“Google (or search engine) hacking” involves using publicly available search engines to access publicly available information that almost certainly was not intended for public distribution.** In short, it’s using clever but legal techniques to find information that doesn’t belong on the public Internet.

To understand how this information has found its way into search engine databases, we need a quick overview of how search engines work. Very simply, search engines deploy “spiders” (aka crawlers or bots), which is actually software that “crawls” websites looking for new sites, updating old ones, following links, and dumping all that data into search engine databases where it is stored, sorted, and eventually accessed by users. There is nothing illegal, immoral, or even fattening about search

---

<sup>61</sup> Let's talk about the term *hacking* for a minute. A hacker is someone who is proficient at using or programming a computer; in short, a computer expert. While there is no universal agreement on a preferred term for someone engaged in illegal/illicit computer or network activity, I will call these “black hat” hackers “malicious hackers” to distinguish them from “white hat” or neutral “hackers,” meaning proficient or expert computer users.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

engine spiders. Indeed, without them, we would have little or no idea what is “out there” and available to us. The problem for webmasters is that it is their responsibility to keep the search engine spiders out of any parts of their websites they do not want to be accessed and indexed by a search engine. The spider is not smart; it simply knows that if a “door” is open, it can—and will—go in and crawl around. Webmasters must tell spiders “do not enter” (primarily) by the use of the Robots Exclusion Protocol.

Robots Exclusion<sup>62</sup> comes in two basic flavors: either a metatag that can be inserted into the HTML of a web page (usually used by an individual) or a Robots Exclusion Protocol (robots.txt) file, a specially formatted file inserted by the website administrator to tell the spider which parts of the website may and may not be indexed by the spider. If a robots exclusion is missing or improperly configured, the spider will index pages that the website owner may not have wished to have been accessed.

The whole problem of keeping information on the Internet private dramatically worsened almost overnight a couple of years ago when Google quietly started indexing whole new types of data. Originally, most of what got spidered and indexed was HTML webpages and documents, with some plain text thrown in for good measure. However, the ever-innovative Google decided this wasn't good enough and started to index PDF, PostScript, and—most importantly—a whole range of Microsoft file types: Word, Excel, PowerPoint, and Access. Problem was, lots of folks had assumed these file types were “immune” to spidering not because it couldn't be done but because no one had yet done it. As a result, many companies, organizations, and even governments had quite a lot of egg on their faces when sensitive documents began turning up in the Google database.

That was then, this is now. You might think people would have learned, but judging by the amount of “sensitive” information still available, many have not. Even though search engines now routinely index many non-HTML file types, many individuals and organizations still do not protect these files from the long reach of search engine spiders. Furthermore, there are many ways for sensitive information to end up in search engine databases. An improperly configured server, security holes, and unpatched software can give search engine spiders unintended access. Quite frankly, most of the problems boil down to one thing: human error, either through ignorance or neglect.

What kinds of sensitive information can routinely be found using search engines? The types of data most commonly discovered by Google hackers usually falls into one of these categories:

---

<sup>62</sup> For additional information, see: <<http://www.robotstxt.org/wc/exclusion.html>> (14 November 2006).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~



~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

- personal and/or financial information
- userids, computer or account logins, passwords
- private, confidential, or proprietary company data
- sensitive government information
- vulnerabilities in websites and servers that could facilitate breaking into the site

Now, you may be thinking to yourself, “I use Google all the time and I’ve never encountered this type of information.” That’s not surprising. It’s not usually the kind of thing you would stumble across inadvertently. Normally, one would have to be actively looking for this type of information. Of course, many of the documents Google hackers find using these techniques are not sensitive and indeed are intended for the public Internet. Only a tiny fraction of the over eight billion pages in the Google index were not meant to be made available to the public. *And, it so happens, these techniques are excellent unconventional ways of finding useful information that might not be discovered using routine search engine queries.* Here are some of the typical techniques used in Google hacking:

- search by file type<sup>63</sup>, site type, and keyword: many organizations store financial, inventory, personnel, etc., data in Excel spreadsheet format and often mark the information “Confidential,” so a Google hacker looking for sensitive information about a company in South Africa might use a query such as:

[filetype:xls site:za confidential]

a similar but more specific search could involve use of a keyword such as *budget* to search for Excel spreadsheets at Indian websites; for example:  
[filetype:xls site:in budget]

- one of the most popular Google hacking technique is to employ **stock words and phrases** such as *proprietary, confidential, not for distribution, do not distribute*, along with a search for specific file types, especially Excel spreadsheets, Word documents, and PowerPoint briefings.
- search for files containing **login, userid, and password** information; note, even at international sites, these terms usually appear in English. This type of information is typically stored in spreadsheet format, so a typical search might be:  
[filetype:xls site:ru login]

---

<sup>63</sup> It is critical that you handle all Microsoft file types on the Internet with extreme care. Never open a Microsoft file type on the Internet. Instead, use one of the techniques described [here](#).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

- ***misconfigured web servers*** that list the content of directories not intended to be on the web often offer a rich load of information to Google hackers; a typical command to exploit this error is:

[intitle:"index of" site:kr password]

- **numrange search**: this is one of the least known and (formerly) one of the scariest searches available through Google. Numrange uses two number separated by two periods (dots) and no spaces. While "***legitimate***" numrange users probably will want to indicate what the numbers mean, e.g., weight, money, pixels, etc. Google does not require any special words or symbols to run a successful numrange search; hence its power. *Numrange* can be used with keywords and other Google search options, such as:

[site:www.jordanislamicbank.com 617..780]

How is numrange typically used in Google hacking? It used to be extremely effective in finding credit card numbers and social security numbers. Because of the publicity about criminals using Google to look for private data, this particular search no longer works for credit card and Social Security numbers, which is not a bad thing.

The disabled "hack" was:

[numrange:4567000000000000..4567999999999999 visa] or

[numrange:222000000..250999999 ssn]

Now if you try these searches, you will see this message:



## Not Found

The requested URL  
/sorry/?continue=http://www.google.com/search%3Fnum%3D100%26hl%3Den%26lr%3D%26newwindow%3D1%26safe%3Dof%3D%26q%3Dnumrange%25  
was not found on this server.

Lest you think I am spilling the beans here, I assure you I am not revealing anything that is not already widely known and used on the Internet both by legitimate and illicit Google hackers. I am fully indebted to Johnny (johnnyihackstuff) Long for many of the "Google hacking" techniques<sup>64</sup> I have learned. Please use the information he provides judiciously because many of the Google *hacking* techniques he discusses are really designed for *cracking*, i.e., breaking into websites and servers. That is not

<sup>64</sup> Johnny Long, *Google Hacking for Penetration Testers*, Syngress: Rockland, MA, 2004.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

something I encourage or advocate. I do encourage you to "hack" your own website to see what kinds of information is being revealed inadvertently via Google and other search engines.

Also, a lot of the best information Johnny offers is for his site members only, and I do not want to suggest you register there. Nonetheless, Johnny's briefing slides from the 2004 Black Hat and Defcon12 conferences are available at the official Black Hat Briefings website and elsewhere (so much for registration). I have also found his excellent white paper "The Google Hacker's Guide" at other sites that do not require registration; there is another very good briefing on the dangers of Google by Sebastian Wolfgarten.

There was a fair amount of sniping following Long's talks at Black Hat and Defcon, mostly of the "big deal" variety, i.e., it is not "real" hacking and therefore not worthy of presenting at Defcon. However, this is a very shortsighted point of view when one considers the kinds of information that is so very easily available via Google, et al. How would you like to see your Social Security Number, credit card number, and that very handy little three digit number on the back of your credit card used for "verification," bank routing information, mother's maiden name, etc., in the next Google hacking briefing? Yes, all this kind of information is readily available (I know...I've uncovered quite a bit of it myself). And this doesn't even take into consideration all the other website weaknesses, such as multiple vulnerabilities with IIS 6.0 Web-based administration, that can be exposed using Google.

Johnny Long's Googledorks Page <http://johnny.ihackstuff.com/ghdb.php>

Johnny Long's "The Google Hacker's Guide"  
[http://www.securitymanagement.com/library/Google\\_Hacker0704.pdf](http://www.securitymanagement.com/library/Google_Hacker0704.pdf)

Johnny Long, "You Got That With Google?" Black Hat Briefings and Defcon12, July 2004.  
<http://www.blackhat.com/html/bh-media-archives/bh-archives-2004.html#USA-2004>

Johnny Long, "Google Hacking Mini-Guide," *Informit.com*, 7 May 2004  
<http://www.informit.com/articles/printerfriendly.asp?p=170880>

Sebastian Wolfgarten, "Watch Out Google"  
[http://www.wolfgarten.com/downloads/Watch\\_out\\_google.pdf](http://www.wolfgarten.com/downloads/Watch_out_google.pdf)

Joe Barr, "Google Hacks are for Real," *Newsforge.com*, 6 August 2004  
<http://www.newsforge.com/article.pl?sid=04/08/05/1236234>

Taken all together, the information Johnny Long has found using Google (he sticks with this one search engine), combined with the techniques he details at his website, provide an excellent tutorial on using Google to find stuff that really should not be on the public Internet or easily accessible via a search query. Furthermore, the greatest value of his efforts may not be in finding useful information but in demonstrating the vulnerabilities of any given website and the necessity of taking strong measures to

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

ensure the information that gets into Google (as well as other search engine databases and the Internet Archive) is only that which is intended.

Given the large amount of "sensitive" or private data readily available via Internet search engines, people naturally wonder why companies and individuals do not actively try to remove this information. Sometimes they do, but much still remains accessible. Why? ***Getting private information "back" is harder than preventing its disclosure in the first place.*** There are steps you can take to remove your data, but as hacker Adrian Lamo says, "removing links after the fact isn't a very elegant solution." Nor is it likely to be terribly effective. There are a number of reasons for this, but what it boils down to is: it's very hard to put the genie back in the bottle.

First of all, you have to find out if your data is "out there" in order to ask search engines to remove it and, clearly, many people and organizations are not playing defense, that is, they are not routinely checking to see what is indexed from their websites. Let's say you find something on Google that shouldn't be on the public Internet. The first thing you have to do is to protect the sensitive pages on your site or remove them entirely. However, even when you have removed those pages from your website, this doesn't mean they can't be accessed. Once documents are indexed in a search engine database, a publicly available copy of those documents (usually referred to as the cache copy) may remain behind for days, weeks, even months.

The next step is to ask Google to remove your sensitive pages from its database. However, even when Google removes your data, there are literally hundreds of other search engines around the world, and who knows what they have indexed from your site. It will not be an easy task finding out. And I'll hazard a guess that not all of them will be quite so accommodating as Google in removing pages.

To make matters worse, if something really "juicy" shows up in a search engine, chances are someone will find it and copy it to another website. Once this happens, you can forget about removing that information from the Internet. To further complicate matters, even if no individual comes across your sensitive data, the Internet Archive<sup>65</sup> spider is almost certainly going to find that webpage and index it in the Archive, and there it will remain until and unless you find it first and ask the Archive to remove it. As you can see, the genie is running amuck! Prevention is much easier (though certainly not easy) than curing this particular disease, so it's vital to pay close attention to anything you put on a website, especially something you do not want the whole world to see.

---

<sup>65</sup> The Internet Archive is a non-profit organization that was founded to "build an 'Internet library,' with the purpose of offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format. Based in San Francisco, the Internet Archive has been harvesting the World Wide Web since 1996, to create one of the largest data collections in the world. The Internet Archive's web archive contains over 100 terabytes of data, and the collection is growing at a rate of 12 terabytes per month." <<http://www.archive.org/>> (14 November 2006).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

Because of the vast amount of information available using public search engines, it's relatively easy to find lots of interesting, amusing, shocking examples of sensitive information. While this is all fine and good for entertaining yourself and impressing your friends, what we are really after is useful, meaningful, and actionable information. Put succinctly:

*It's Easier to Find Anything Than It Is to Find Something*

So how do you find "something" useful? While it isn't easy to do so, I can make some suggestions that might help. The most valuable assets you have are your subject matter knowledge and your creativity. Add these to a few search engine strategies, and you can probably find many relevant and genuinely useful pieces of information. The strategies I recommend for finding "something" rather than just "anything" are:

Limit the search by site

This can be as broad as a county [site:fr] or as specific as an individual server on a company website [site:office.microsoft.com].

Try to be as specific as possible

You will have a lot more success searching for information within the Chinese Ministry of Foreign Affairs [site:fmprc.cn.gov] than looking at all the sites indexed for China [site:cn] or even for the government of China [site:gov.cn]

Add keywords

Here's where your subject matter knowledge and creativity really help. You are the best source of information about what words are most likely to yield the best quality and quantity of useful information. As a general rule, more uncommon words work best (consider using unusual proper names).

Limit the search by file type

Most of the best information found by Google hackers is not on webpages (HTML) but in other types of files. Try all or most of the file types one at a time (these are not the only searchable file types; check the particular search engine's documentation (*Help* page) for others):

filetype:pdf—good for large documents of all types; widely used in academia, government, and business; many PowerPoint briefings are also made available in PDF at the same website

filetype:doc—good for internal working documents, reports, etc.

filetype:xls—good for personnel data, computer records, financial information

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

filetype:ppt—good for briefings, which often contain company or government plans for the future

Back File Edit View Insert Format Tools Window Help

http://www.apnic.net/meetings/16/programme/docs/apops-pres-gaur...

88%

Bookmarks

Thumbnails

# ICT in Afghanistan

PowerPoint briefings may contain useful and/or unique information

Muhammad Aslam  
af ccTLD Manager

Presentation by: Gaurab Raj Upadhaya  
at APOPS Forum, 16 APNIC Open Policy Meeting  
August, 21, 2003. Seoul, Korea

AFGNIC

And, often, PowerPoint files are also available in PDF(safer/easier to read)

1 of 14 11.69 x 8.26 m

### Use Google hacking techniques to search inside websites requiring registration

You will frequently encounter a website, perhaps a database, that requires registration to view its contents. On occasion, you can use Google to get at that data without registering. For example, let's say you find a database of international companies that requires *free* registration. Without registering, you may be able to use Google to list all the companies and even get a look at the individual entries. Try this series of queries or something similar:

[site:www.companyname.com inurl:database] or

[site:www.companyname.com inurl:directory] or

[site:www.companyname.com inurl:index]

Then, look for keywords, such as *companies*, and move to the next level query:

[site:www.companyname.com inurl:companies]

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

You may be able to browse through the list of companies and get names, addresses, phone numbers, etc.

### Search in the native language

I cannot emphasize strongly enough how important it is to use keyword search terms that are in the native language of the entity you are researching. The Internet is becoming much less dependent upon English, and sites written in languages that do not use the Latin alphabet are growing by leaps and bounds. For example, a search term written in the native language and encoding is far more likely to yield interesting, useful results than the same word transliterated into English. Most good quality search engines now correctly render non-Latin search terms regardless of how the term is transliterated in English. A search on the Arabic **محمد** returns very different results than searching on [muhammad], [mohamet], [mohammed], etc.

[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [Desktop](#) [more »](#)  
**Google** محمد   [Advanced Search](#) [Preferences](#)

---

**Web** Results 1 - 100 of about 25,800,000 for 0.06 seconds

[Sheikh Muhammad Jebri | الشيخ محمد جبريل](#) | [Translate this page](#)  
 ... by Sheikh Muhammad Jebri's spectacular voice. مع صوته الشيخ محمد جبريل في رحاب القرآن الكريم ولأداء مع صوته الشيخ محمد جبريل  
[www.jebri.com/](http://www.jebri.com/) - 6k - [Cached](#) - [Similar pages](#)

[الصفحة الرئيسية](#) | [Translate this page](#)  
 ... صفحات الشيخ محمد جبريل الموسيقية والأدعية والتسجيلات القرآنية بالإضافة لمعلومات عن  
[www.jebri.com/af/index.html](http://www.jebri.com/af/index.html) - 33k - [Cached](#) - [Similar pages](#)

[Mohammad Estahani Official Web Site](#)  
 Iranian singer. Profile, discography, and pictures.  
[www.mohammad-estahani.com/](http://www.mohammad-estahani.com/) - 10k - [Cached](#) - [Similar pages](#)

[مهرجانكم في موقع الموسوم محمد المناخي](#)  
 photographing landscape,Portrait. ... All works of art copyright © MOHAMED MANNAI. All rights reserved. Copyright © 2000-2006 MOHAMED MANNAI.  
[www.mmannai.com/](http://www.mmannai.com/) - 7k - [Cached](#) - [Similar pages](#)

[MUHAMMAD ALI - The Greatest Of All Time](#)  
 This is the Official website of Muhammad Ali, the greatest of all time.  
[www.ali.com/](http://www.ali.com/) - 22k - [Cached](#) - [Similar pages](#)

[Welcome to His Highness Sheikh Mohammed bin Rashid Al Maktoum's ...](#)  
 Official site of the Ruler of Dubai, who is also the UAE Vice President and Prime Minister. Contains news, his poetry and other information in Arabic and ...  
[www.sheikhmohammed.co.ae/](http://www.sheikhmohammed.co.ae/) - 2k - [Cached](#) - [Similar pages](#)

[... الموقع الشخصي الرسمي للشيخ محمد بن](#) | [Translate this page](#)  
 ... هذا هو الموقع الشخصي الرسمي لصاحب السمو الشيخ محمد بن راشد آل مكتوم نائب رئيس الدولة  
[www.sheikhmohammed.co.ae/arabic/index.asp](http://www.sheikhmohammed.co.ae/arabic/index.asp) - 2k - [Cached](#) - [Similar pages](#)

[Al-Hammadi.com - A Website for All](#)  
 Al-Hammadi.com is a website with information on Qatar, Islam, Arabic music, and more. Come on in and enjoy what we have to offer.  
[www.al-hammadi.com/](http://www.al-hammadi.com/) - 12k - [Cached](#) - [Similar pages](#)

### Remember that Diacritics Also Affect Searches

Most search engine algorithms are now set up to “read” accented search terms differently from those without accents. It’s easy to test this by searching first for a term without any diacritical marks and then the same word with the marks, e.g., resume vs. résumé.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

## Types

Some common types of diacritical marks:

- ◆ acute accent ( ´ )
- ◆ ring<sup>1</sup> above ( ° ) used for angstrom ( Å ), aka krouzek
- ◆ breve ( ˘ )
- ◆ caron or háček ( ˇ )
- ◆ cedilla ( ¸ )
- ◆ circumflex ( ^ )
- ◆ umlaut<sup>1</sup> or diaeresis ( ¨ )
- ◆ double acute accent ( ˆ )
- ◆ grave accent ( ` )
- ◆ macron ( ¯ )
- ◆ ogonek ( ˛ )
- ◆ spiritus asper
- ◆ spiritus lenis

<sup>1/</sup> Strictly taken not diacritics but parts of the character.

66

### Look for Misspellings (Intentional or Accidental)

I am constantly amazed by the frequency of misspelled words, urls, file names, etc., I encounter on the Internet. By far, most appear to be simple mistakes, often made by non-English speakers trying to cope with our confusing language. These mistakes tend to propagate as users copy and paste them again and again, which is what I believe happened here:

---

<sup>66</sup> Fact Index, <<http://www.fact-index.com/d/di/diacritic.html>>



UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~**Web**Results 1 - 10 of 10 from [www.chinadaily.com.cn](http://www.chinadaily.com.cn) for **enlgish**. (0.30 seconds)jobs

Chinadaily.com.cn Recruitment 中国日报网站招聘. 网站设计部-高级制作师(1名)(北京市). 要求 有良好的美术基础 ...  
[www.chinadaily.com.cn/enlgish/doc/2004-03/16/content\\_315314.htm](http://www.chinadaily.com.cn/enlgish/doc/2004-03/16/content_315314.htm) - 23k - [Cached](#) - [Similar pages](#)

jobs

Chinadaily.com.cn Recruitment 中国日报网站招聘. 网站设计部-设计师(1名)(北京市). 要求 有专业美术设计基础 ...  
[www.chinadaily.com.cn/enlgish/doc/2004-03/16/content\\_315316.htm](http://www.chinadaily.com.cn/enlgish/doc/2004-03/16/content_315316.htm) - 23k - [Cached](#) - [Similar pages](#)

jobs

Chinadaily.com.cn Recruitment 中国日报网站招聘. 市场部-项目经理 (发行推广业务) (2名)(北京市). 要求: 30 ...  
[www.chinadaily.com.cn/enlgish/doc/2004-03/16/content\\_315317.htm](http://www.chinadaily.com.cn/enlgish/doc/2004-03/16/content_315317.htm) - 23k - [Cached](#) - [Similar pages](#)

jobs

Chinadaily.com.cn Recruitment 中国日报网站招聘. 《21世纪少年英文报》诚聘各地发行代理. 《21世纪少年英文报》 ...  
[www.chinadaily.com.cn/enlgish/doc/2004-04/06/content\\_321050.htm](http://www.chinadaily.com.cn/enlgish/doc/2004-04/06/content_321050.htm) - 9k - [Cached](#) - [Similar pages](#)

jobs

Chinadaily.com.cn Recruitment 中国日报网站招聘. 英语学习栏目编辑(2-3名)(北京市). 工作所在地: 北京 职责 ...  
[www.chinadaily.com.cn/enlgish/doc/2004-03/16/content\\_315312.htm](http://www.chinadaily.com.cn/enlgish/doc/2004-03/16/content_315312.htm) - 30k - [Cached](#) - [Similar pages](#)

Finally, the enormity of the task of finding meaningful and useful information on the Internet is both daunting and comforting: daunting because we know we can only scratch the surface of all the data and comforting because there is an almost limitless pool of possibilities. I find it useful to keep the challenge in perspective by recalling that a study published in 2000 showed "*the sixty known, largest deep Web sites contain data of about 750 terabytes (HTML-included basis) or roughly forty times the size of the known surface Web.*"<sup>67</sup> In short, there is just so much data and information available via the Internet that no institution, no government, no computer, and certainly no individual can possibly grasp more than a small portion of all there is.

<sup>67</sup> Michael K. Bergman, "The Deep Web: Surfacing Hidden Value," *BrightPlanet .com*, July 2001, <<http://www.brightplanet.com/technology/deepweb.asp>> (14 November 2006), Introduction.

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

---

## Custom Search Engines

---

This topic is new this year and expands upon the entries on Rollyo and Gigablast's Custom Topic Search from last year's edition. During 2006 there was an explosion in the number of custom search engines, including entries from Google, Yahoo, and Live Search, so you know the powerhouses think this is worth a try. Whether this trend catches on remains to be seen.

The phrase "custom search engine" is very misleading. None of these sites permits users to create a new search engine. What each site does in its own way is to let users customize an existing search engine to search specific sites in specific ways and return results in a personalized fashion. Thus, a better name for these services would be customizable searching, but that moniker is clearly unappealing. Just remember that you are not creating a new search engine any more than customizing a car is building a new automobile from the tires up.

Most of the custom search sites operate on a simple principle: they automate a long "site" search, e.g., the search is equivalent to [keyword(s) AND (site 1 OR site 2 OR site 3...OR site n)], where n stands for the maximum number of sites you are allowed to search.

In short, the proliferation of customizable search means that companies, educational institutions, government agencies, and individuals can easily put the power of the big search engines such as Google, Yahoo, and Live Search with its search Macros to work creating tailored and specialized search services in a way that has never before been possible. Customizable search may be "the next big thing," and I believe it is one of the most positive examples of that vague but ubiquitous concept called Web 2.0.

### Gigablast's Custom Topic Search

<http://www.gigablast.com/cts.html>

Gigablast's Custom Topic Search was one of the first "create your own search engines" to appear, although Gigablast's creator Matt Wells never claimed it was anything other than a way to customize Gigablast. The beauty of the Gigablast CTS is that it requires no software installation but is very, very simple HTML code, so simple anyone can edit and understand it. No registration is required.

Many of Gigablast's features were primarily designed for webmasters instead of users, but this one is potentially valuable to both: "**Build Your Own Topic Search Engine.**" Gigablast "allows you to create a list of up to 200 web sites (or subsites) and a search box that searches just those sites." **Custom Topic Search** even lets you decide if you want Gigablast to cluster the results for you. The concept behind

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

topic search is that you, and not some anonymous marketer, choose the sites you want to search. This “tool” (for want of a better word) is amazingly easy to use and powerful. As someone whose eyes glaze over at the mere sight of code, let me put this in “user” language. If you are familiar with Google’s **site:** syntax, imagine being able to have a “canned” query that runs against up to 200 websites of your own choosing and lets you run it whenever you like and use whatever keyword(s) you want at any time. The query on Google would look something like this:

[keyword site:cnn.com OR site:dmoz.org OR site:amazon.com OR  
site:usatoday.com OR site:cia.gov (etc.)]

The problem with Google is that multiple site/domain searches are cumbersome at best, and they quickly run up against Google’s 32-word limit. Enter Matt Wells and Gigablast. As the creator and sole proprietor of his own search engine, Matt has the luxury of being able to add new options easily. I think CTS is his best innovation yet. Even if you are as HTML-averse as I am, this code is so easy to edit that it’s a piece of cake. To make things even easier, I have done the basics for you. First, however, I highly recommend you read through the Gigablast pages below on the concepts behind CTS.

Build Your Own Topic Search Engine of Custom Topic Search

<http://www.gigablast.com/byose.html>

<http://www.gigablast.com/cts.html>

Now you’re ready to take a look at, edit, and try the CTS. Copy and paste this HTML code into an application such as Notepad.

```
<head>
  <title>Gigablast Custom Search</title>
</head>
<body>
  Search News Websites
  <form method="post" action="http://www.gigablast.com/search">
  <input type="text" name="q" size="60">
  <input type="submit" value="search" border="0">
  <input type="hidden" name="sc" value="1">
  <input type="hidden" name="sites" value="cnn.com news.yahoo.com
  news.google.com usatoday.com foxnews.com">
  </form>
</body>
```

This is a bare bones version of the CTS code. Now you can play with the code and make it into your own custom topic search page. I should mention that I set the “site clustering” option to ON `<input type="hidden" name="sc" value="1">` but you can reset it to OFF by changing 1 to 0. Once you **save as an HTML file**, all you have to

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

do to use it is to **open the file in your browser**, insert keyword(s), and go. Obviously, you will want to add more sites to search (I only put in a few) and change the topic to something of interest to you (I chose the rather bland News topic for demonstration purposes). Also, you can enter sub-sites or more specific sites, such as [cnn.com/WORLD](http://cnn.com/WORLD) or [dir.yahoo.com](http://dir.yahoo.com).

One thing to keep in mind that is **you are searching Gigablast's database of pages from these websites, not the sites themselves**. The "work" that goes into creating a CTS is mostly up front because once you create your list of sites, it is not a complicated matter to add to or subtract from it. I can easily imagine creating a set of these search forms on a variety of topics using existing bookmarks.

### Rollyo

<http://rollyo.com/>

Rollyo stands for "Roll your own" search engine, meaning that you select the sources you want to search. Rollyo is powered by Yahoo, so results will come from Yahoo only. Rollyo lets users search up to 25 sites (not a huge number) and also try out and use other people's "Searchrolls." In order to save, share, and use your Searchrolls on other computers, you must register with an email address and a user-created name and password.

Rollyo has some unusual features. For example, Rollyo permits users to upload their bookmarks to create Searchrolls, edit someone else's Searchroll to make it your own, keep your Searchrolls private or share them. Rollyo searches entire sites or you can limit your search to a subdomain; however, you cannot limit your search to directories within a site, e.g., in this case, everything after the slash is ignored: [security.news.com/library](http://security.news.com/library).

Rollyo has a nice little bookmarklet called Rollbar that "gives you access to all of your Searchrolls wherever you are.

- Search any site you visit, from the same spot on your browser, without having to dig around for every site's search page.
- Add sites to your Searchrolls on the fly.
- Create a new Searchroll from anywhere."  
<<http://rollyo.com/bookmarklet.html>>

One of the most attractive features of Rollyo is the ability to share Searchrolls. Here is an example of a Searchroll named "Muslim World Views." The sources searched are on the left side:

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//FOR OFFICIAL USE ONLY

The screenshot shows the ROLLYO website interface. At the top left is the ROLLYO logo with the tagline 'GET IT FIRST'. To the right is a search bar with a 'Try it out' button and a dropdown menu set to 'Muslim World Views'. Below the search bar are three columns of results:

- MUSLIM WORLD VIEWS:** Includes an 'Add to my Rollyo' button and a list of sites searched: irna.com, english.aljazeera.net, asharqalawsa.net, adnk1.com, islamrepublic.com, islamgenerations.com, al-bawaba.com, arabicnews.com, epilive.net, nst.com.my, english.daralhayat.com.
- Latest Custom Blog Results:** Features a news item from the New Straits Times (Dec 18, 2006) about 'Housemen's working conditions: Making it more bearable for junior doctors'.
- Latest Custom News Results:** Features a news item from the New Straits Times (Jan 17, 2007) about 'Baghdad university bombing kills 70'.
- Custom Web Results:** Lists search results for 'Al Jazeera English - Archive' and 'Al Jazeera English - File Not Found'.

On the right side, there is a 'CHECK IT OUT!' section with links to 'Get your RollBar' and 'Add a Searchbox', and a 'DO STUFF' section with options like 'Add to my Rollyo', 'Edit this Searchroll', 'Link to Searchroll', 'Add to Firefox™', and 'Share with a friend'. At the bottom right, there is an advertisement for 'Soundflavor DJ' with the text 'Your Personal DJ for iTunes Free Download!'.

Rollyo has added blog and news searches (again, from Yahoo) to the results. Rollyo makes it very easy to create, save, and edit custom searches.

**Google Custom Search Engine** <http://www.google.com/coop/cse/overview>

Google got into the custom search game rather late. In October 2006 Google announced its own version of a custom search engine. In the announcement, Google said,

“When we say we’re letting people build a custom search engine, we mean the whole thing: choosing which pages they want to include in their index, how the content should be prioritized, whether others can contribute to the index, and what the search results page will look like...Here’s how a Custom Search Engine works: organizations or individuals simply go to [www.google.com/coop/cse](http://www.google.com/coop/cse) and select the websites or pages they’d like to include in their search index. Users can choose to restrict their search results to include only those pages and sites, or they can give those pages and sites higher priority and ranking within the larger Google index when people search their site. Users can then customize the look, feel and functionality of their search engine.”<sup>68</sup>


<sup>68</sup> Google Press Release, “The Power of Google Search is Now Customizable,” 23 October 2006, <[http://www.google.com/press/annnc/custom\\_search.html](http://www.google.com/press/annnc/custom_search.html)> (17 January 2007).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

After a telephone conference with Google's Marissa Mayer and the Google product managers, search and Google expert John Battelle shared his comments, which I think are excellent insights:

"While similar to Rollyo's innovative custom roll, the Google CSE adds the benefit of allowing users to roll an unlimited number of sites together and display the results on their own site, with personalized presentation. Someone on the call described this as the fragmentation of search. The ability to build verticals will allow experts to build specialized engines. But while the engines will be individual, the collaborative element of tagging the domains encourages communities of knowledge to create together. So while each will stand apart from the amazing all-in-one answer box, the Custom Search will also allow a thickening or deepening of intelligent tags in Co-op, which feeds the one box that unites them all." <<http://battellemedia.com/archives/003006.php>>

Not surprisingly, **you must have a Google account to use this service**. Also, Google Custom Search includes AdSense sponsored links alongside search results, but government sites, non-profits, and educational institutions are exempt from the advertising requirement. To see the Google Custom Search in action, take a look at Real Climate.org's internal search: <<http://www.realclimate.org/>> Even better, check out **Customsearchguide**, a directory of Google Custom Search Engines that others have created but you can use. Here is an example of general science and technology custom searches.



**Control Center**  
[Home](#) • [Register](#) • [Login](#) • [Suggest](#)

**General Science And Technology Search Forms**

You Are Here: [Home](#) > [Technology](#) > General Science And Technology Search Forms

[Business and Finance](#) • [Health](#) • [Media](#) • [Reference](#) • [Shopping](#) • [Society](#) • [Sports](#) • [Technology](#) • [Travel](#)

Search Form	Editor Rating	Visitor Rating	Description
<a href="#">Technology Search</a>	n/a	●●●●	Searches science and technology resources.
<a href="#">Science and Engineering Search</a>	n/a	n/a	Find info on science and engineering.
<a href="#">Science Wikis</a>	n/a	n/a	Searches science-oriented wiki sites

**Webmasters & Bloggers:** Link here & encourage people to use and vote for your favorite CSEs

```
<a href="http://www.customsearchguide.com/categories/technology/general-science-Science And Technology Search Forms">
```

[home](#) • [link to us](#) • [about](#) • [legal](#) • [privacy](#) • [press](#)  
 Copyright © 2006, Moving Traffic, Inc. All Rights Reserved.

Customsearchguide

<http://www.customsearchguide.com/>

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

### **Yahoo Search Builder**

<http://builder.search.yahoo.com/>

Yahoo's custom search option requires registration and is very similar to others create your own search sites. Rollyo predates the Yahoo Search Builder and also searches the Yahoo database, giving you a good idea of what you can do with this tool.

### **Live Search Macros**

<http://search.live.com/macros/default.aspx>

I discuss creating and finding search macros in the section devoted to Microsoft's Live Search.

### **Alexa Web Search Platform**

<http://websearch.alexa.com/welcome.html>

What has been for many the holy grail of search is now a big step closer to reality. With little fanfare, Amazon's Alexa subsidiary announced in December 2005 that it was opening up its search tools and index to the world in a new project named the Alexa Web Search Platform (AWSP)—and for a very modest price.

According to its website, "The Alexa Web Search Platform provides public access to the vast web crawl collected by Alexa Internet. Users can search and process billions of documents and even create their own search engines using Alexa's search and publication tools. Alexa provides compute and storage resources that allow users to quickly process and store large amounts of web data. Users can view the results of their processes interactively, transfer the results to their home machine, or publish them as a new web service."

What exactly is Alexa offering to the user? In essence, Alexa gives the user, whether an individual or organization, access to the same kind of powerful technology used by Google, Yahoo, and Live Search. "Alexa spiders 4 billion to 5 billion pages a month and archives 1 terabyte of data a day. The new platform will allow developers to build their own search engines." The goal? To democratize web search by taking it out of the hands of giants like Google and putting it into the hands of literally anyone and everyone. The implications are enormous. And it appears it is a hit. In fact, within a very short time of its initial opening, Alexa had to cut off new applications temporarily because it was overloaded with customers wanting to sign up for the new service, but the site soon reopened registration.

The Alexa Web Search Platform (AWSP) offers the user the capability to:

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

- define (search): AWSP has a much more robust set of search options, syntax, and APIs than other search engines and also permits the use of stored (canned) queries; the AWSP “data store” contains text, html, music, video, images, and more types of files.
- process: users can search the entire Alexa data store and “are able to process both the raw content and the metadata extracted by Alexa’s internal processes.”
- publish: the output of the search can be anything from one result to an entirely new vertical search engine, for example a new video search engine or a new search engine for automotive parts. Quite literally, “by making use of these utilities, a user might introduce a great new search service to the world with nothing more than a home computer.”<sup>69</sup>

The costs are modest and are based on consumption (you pay for what you use and not for a subscription or service contract):

\$1 per cpu hour (\$0.50 for reserved but unused hours)

\$1 per GB/year of user storage

\$1 per 50 GB processed

\$1 per GB uploaded/downloaded

\$1 for every 4,000 user-published web service requests

In case you’re curious, Alexa has a long history. Now owned by Amazon, Alexa was created by Bruce Gilliat and Brewster Kahle (of Internet Archive fame), and until now has been both famous and infamous as the technology behind the controversial web traffic and website statistics “What’s Related” toolbar feature in both Netscape and Internet Explorer. The new AWSP is actually integrated into Amazon’s web services platform, something no one has done before.<sup>70</sup>

Simply stated, Alexa/Amazon are “renting” their huge database (“data store”) to any and all takers for a remarkably reasonable price and, what is more, offering detailed

---

<sup>69</sup> Alexa Web Search Platform User Guide, Introduction: What Can I Do with the Platform? <[http://pages.alexaweb.com/awsp/docs/WebHelp/AWSP\\_User\\_Guide.htm](http://pages.alexaweb.com/awsp/docs/WebHelp/AWSP_User_Guide.htm)> (17 January 2007).

<sup>70</sup> There is one example of something similar, which came to my and some others' minds. If you are familiar with IBM's WebFountain and its proprietary implementations for specific customers, you may see some similarities. WebFountain also spidered the web and then let IBM's customers run queries against that data set in more sophisticated ways than simple querying (something akin to datamining). However, the problem with WebFountain and its progeny was that IBM had to write the programs, and thereby hangs a tale of woe. For more, I recommend Jeff Dalton's blog entry on this topic (I think he nails it). Jeff Dalton, “Alexa Web Search Platform: IBM WebFountain 2.0,” Jeff's Search Cafe, <<http://searchcafe.blogspot.com/2005/12/alexaweb-search-platform-ibm.html>>



~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

user support on how to maximize the effectiveness of this data to get the most out of it. The customer is empowered to write his own program to run against the Alexa/Amazon data, download the results (metadata), and even create his own private search engine on their platform. Perhaps I am wrong, but this could be a huge development, perhaps even a major change in the way we use the web.

Amazon Web Services Platform

<http://www.amazon.com/gp/browse.html/104-1308416-9976726?node=3435361&>

Alexa Web Search Platform (beta)

<http://websearch.alexa.com/welcome.html>

Alexa Web Search Platform Users Guide

[http://pages.alexa.com/awsp/docs/WebHelp/AWSP\\_User\\_Guide.htm](http://pages.alexa.com/awsp/docs/WebHelp/AWSP_User_Guide.htm)

### **More Custom Search Sites**

There are other sites offering customized search that you may want to experiment with to find one that best suits your needs. Search expert Phil Bradley reviews some of these custom search sites in a two-part article on Searchenginewatch.com:

“Search Your Own Way,” Part I,

<http://searchenginewatch.com/showPage.html?page=3623434> and Part 2,

<http://searchenginewatch.com/showPage.html?page=3623482>

Eurekster's Swicki

<http://swicki.eurekster.com/>

PSS

<http://www.pssdir.com/>

---

## **Fagan Finder**

---

The Fagan Finder site has been a boon to searchers for some time not so much because of its basic interface, which is a good but unexceptional megasearch tool, but because of the many other “useful tools” site creator Michael Fagan has made available.

### **Fagan Finder File Format Search**

Instead of having to visit a number of different search engines to search for files in a variety of formats, users can now go to the Fagan Finder “search by File Format” page, which is still in beta testing but appears to be running just fine. By selecting a specific file format, e.g., Microsoft PowerPoint, Fagan Finder automatically shows

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//FOR OFFICIAL USE ONLY

which of the search engines is capable of searching for that particular type of file. Not every search engine on the list searches for every file type.

Also, keep in mind that the Fagan Finder file type search for XML is less precise than going directly to Google or Yahoo and searching by *filetype:* in Google and by *originurlextension:* in Yahoo. If you use one of these search engines, you can specify that you only want to search for, say, those files that are .rss by entering the query [filetype:rss] or [originurlextension:rss]. These queries will return only those documents in RSS format, not those in XML or RDF. So I recommend using the Fagan Finder search by file type for files types other than XML, RSS, or RDF.

**Fagan Finder > Search by File Format (beta)**

Sponsors: [Search Engine Optimization](#) [Buy Text Links](#) [URL Directory](#)

**options**

<b>File Format:</b>		<b>Search Engine</b>
<input checked="" type="checkbox"/> Adobe Portable Document Format	<input type="checkbox"/> Lotus 1-2-3	<input checked="" type="checkbox"/> Google <a href="#">info</a>
<input type="checkbox"/> Adobe PostScript	<input type="checkbox"/> Lotus WordPro	<input type="checkbox"/> Yahoo!
<input checked="" type="checkbox"/> Microsoft Excel	<input type="checkbox"/> Star Office	<input type="checkbox"/> Gigablast
<input type="checkbox"/> Microsoft PowerPoint	<input type="checkbox"/> MacWrite	<input type="checkbox"/> Teoma
<input checked="" type="checkbox"/> Microsoft Word	<input type="checkbox"/> Macromedia Flash	<input type="checkbox"/> Exalead
<input type="checkbox"/> Microsoft Works	<input type="checkbox"/> Text	<input type="checkbox"/> Scirus
<input type="checkbox"/> Microsoft Write	<input type="checkbox"/> XML	<input type="checkbox"/> Sensis
<input checked="" type="checkbox"/> Rich Text Format	<input type="checkbox"/> AutoCAD	
<input checked="" type="checkbox"/> Corel WordPerfect		

**information**

**About this Tool**  
This tool uses enables easy access to searching for various non-HTML (standard web page) file formats. Certain documents are commonly used for different purposes; for example many academic papers are in Adobe Portable Document Format. Because this tool makes use of other tools, it is limited by their functionality. Searching in Google for XML files, for example, uses the file extensions xml, rdf, and rss; which means that not all XML files are included, and some non-XML files may be included.

**File Viewing**  
Different file formats require different software to view those files. Adobe Portable Document Format, for instance, requires [Adobe Reader](#).

**Scirus and Sensis**  
Scirus is a search engine for scientific information; it includes Adobe Portable Document Format files in addition to standard web pages. It is powered by [Fast Search & Transfer](#), the former owner of the [AlltheWeb](#) search engine. Sensis is a search engine which has both "world" and "Australia" options. Restricting by file format is not perfect yet, as some results returned may not be of the requested type.

Fagan Finder Search by File Type

<http://www.faganfinder.com/filetype/>

## URLInfo

<http://www.faganfinder.com/urlinfo/>

The indefatigable Michael Fagan also introduced a beta version of a new tool, URLInfo, in mid-2004. URLInfo fills a void created when AlltheWeb effectively shut down and took with it the useful "url investigator." While Yahoo now offers [Site Explorer](#) and Google a lame version [info:domain.com], Fagan's URLInfo provides many more options for exploring a site. As with everything he does, Fagan has gone all out with URLInfo, almost to the point of providing too many options! However, he has done a smart thing in keeping the main URLInfo page simple, "hiding" the nearly 85 investigative tools in his toolkit behind a variety of tabs. I think URLInfo is important and valuable enough to spend time looking at most of the options in some detail.

UNCLASSIFIED//FOR OFFICIAL USE ONLY

Here is a snapshot of the URLinfo main page.

The screenshot shows the URLinfo main page with the following sections:

- Search by File Format (beta)**: A search bar with a search button and a clear button.
- Options**: A section with three columns of file formats and search engines.
 

File Format	Search Engine
<input type="checkbox"/> Adobe Portable Document Format	<input type="checkbox"/> Google <a href="#">info</a>
<input type="checkbox"/> Adobe PostScript	<input type="checkbox"/> Yahoo!
<input type="checkbox"/> Microsoft Excel	<input type="checkbox"/> Gigablast
<input type="checkbox"/> Microsoft PowerPoint	<input type="checkbox"/> Teoma
<input type="checkbox"/> Microsoft Word	<input type="checkbox"/> Exalead
<input type="checkbox"/> Microsoft Works	<input type="checkbox"/> Scirus
<input type="checkbox"/> Microsoft Write	<input type="checkbox"/> Sensis
<input type="checkbox"/> Rich Text Format	
<input type="checkbox"/> Corel WordPerfect	
<input type="checkbox"/> Lotus 1-2-3	
<input type="checkbox"/> Lotus WordPro	
<input type="checkbox"/> Star Office	
<input type="checkbox"/> MacWrite	
<input type="checkbox"/> Macromedia Flash	
<input type="checkbox"/> Text	
<input type="checkbox"/> XML	
<input type="checkbox"/> AutoCAD	
- Information**: A section with three sub-sections:
  - About this Tool**: This tool enables easy access to searching for various non-HTML (standard web page) file formats. Certain documents are commonly used for different purposes; for example many academic papers are in Adobe Portable Document Format. Because this tool makes use of other tools, it is limited by their functionality. Searching in Google for XML files, for example, uses the file extensions xml, rdf, and rss, which means that not all XML files are included, and some non-XML files may be included.
  - File Viewing**: Different file formats require different software to view those files. Adobe Portable Document Format, for instance, requires [Adobe Reader](#).
  - Scirus and Sensis**: Scirus is a search engine for scientific information; it includes Adobe Portable Document Format files in addition to standard web pages. It is powered by [Fast Search & Transfer](#), the former owner of the [AlltheWeb](#) search engine. Sensis is a search engine which has both "world" and "Australia" options. Restricting by file format is not perfect yet, as some results returned may not be of the requested type.

Note the eleven tabs at the top, behind each of which is a range of investigatory options. For help using URLinfo simply click on the dark blue **[info]** link on the far right. The first step in using URLinfo is to enter a url (address) in the search box at the top of the page. Keep in mind that ***if you enter a url in the search box and simply hit return, you will be taken to that webpage, not to information about it.***

Entering a url can prove to be more problematic than you might think because not every URLinfo tool can handle the same format. For example, in the General tab, the one most users are likely to use most frequently, you will get very different results depending on the type of url entered. For basic .com, .org, .net, .info, .biz, and .us domains, Domain Tools is great. However, for any other top-level domain, you must use Global Whois, and it will not search on anything but first-level domain names. This means that neither Domain Tools nor Global Whois can look up [www.duma.gov.ru]. Global Whois, however, will find first-level domains such as [www.feb-web.ru]. This does not mean you cannot find information about [www.duma.gov.ru].

Take a look at the results from the first tab, Alexa.

As you can see, you get lots of data about the Russian Duma website. Note that there are many additional useful links from the Alexa page, including one to the Internet Archive's Wayback machine.

UNCLASSIFIED//FOR OFFICIAL USE ONLY

UNCLASSIFIED//FOR OFFICIAL USE ONLY



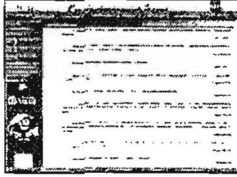
Web Search **Traffic Rankings** Web Directory

duma.gov.ru

Get Traffic Details

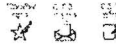
Top 500 - Moves & Shakes

Category: Top > World > Russian > Страны и регионы > Европа > Россия > Государство > Законодательная власть > Государственная Дума



Государственная Дума  
Федерального Собрания  
Российской Федера...

Официальный сервер. История и регламент Государственной Думы, информация о ее законодательной деятельности. Депутатский корпус. Законодательство РФ. Информация пресс-службы.



EXPLORE THIS SITE

- ◆ Overview
- ◆ Traffic Details
- ◆ Related Links
- ◆ Sites Linking in

Sponsored Links

**Raid The Bank!**  
Money you never knew you were missing!

**Free Congo Netpass**  
Free Access to the Web's Top Subscription sites!

Traffic Rank for gov.ru: 14,249

Your Ad Here

Share your thoughts

Write a review on Amazon.com..

E-mail a friend about this site

Quick Pick

President.kremlin...



People who visit this page also visit:

- Akdi.ru/gd - Site info
- Prodem.ru - Site info
- Minist - Site info
- Government.gov.ru - Site info
- www.vokrugz - Site info
- Lithuania - Site info
- Cabinet of Ministers of Ukraine - Site info
- Governul Rom - Site info
- Government.gov.sk - Site info

The Alexa database contains site statistics, contact information, similar pages, and more.

**What is Alexa?** Many things, but most interesting and useful is Alexa's Site Information:

"Alexa has built an unparalleled database of information about sites that includes statistics, related links and more. All of this information can be found on Alexa's Site Overview pages, Traffic Detail pages and Related Links pages. To access these pages, simply type the URL of any site into the Alexa Search box."

Alexa Site Information, <http://www.alexa.com/site/company>

Fagan Finder's URLInfo <http://www.duma.gov.ru> view page clear

General Links Similar Cache Search Blogs/Feeds Translate Track Post Develop Misc

Alexa Whois Source Global Whois SurfWax Stumble Upon 7Search metaEUREKA Furl Delicio.us Gibeo Semantic data extractor FyberSearch [Info](#)

---

**Travel Back in Time!**  
INTERNET ARCHIVE

Use the Wayback Machine to see how Государственная Дума Федерального Собрания Российской Федера... looked in the past.

Site Stats for gov.ru:

- **Traffic Rank for gov.ru:** 58,371 (★852)
- **Speed:** Very Fast (95% of sites are slower), Avg Load Time: .4 Seconds (what's this?)
- **Other sites that link to this site:** 172
- **Online Since:** 10-Jul-1997

[▶ See Traffic Detail...](#)

---

Contact Info for gov.ru:

**RUSSIAN GOVERNMENT INTERNET NETWORK (RGIN)**

+7 095 2062863, Fax: +7 095 2067355  
webadmin@duma.gov.ru

---

**Track Your Website Statistics!**  
The NEW HitsLink tracks even more live data on your visitors, with advanced keyword analysis

Take A Free Trial

User Reviews for gov.ru

Be the first person to write a review of this site on Amazon.com!

---

Look for similar sites by category:

- World / Russian / Страны и регионы / Европа / Россия / Государство / Законодательная власть / Государственная Дума
- World / Russian / Страны и регионы / Европа / Россия / Государство

Let's look at a different url for the **SurfWax** results. What you are seeing are "SurfWax SiteSnaps™, [which] count the number of links, images, words, and forms on a page, shows the meta description tag, and extracts 'key points' and 'FocusWords.'" This is a very useful way to analyze a website without actually visiting it, though the amount of information is considerably less for some sites than others, *cf.*, [www.fateh.net](http://www.fateh.net).

Fagan Finder's URLInfo <http://www.nla.gov.au> view page clear

General Links Similar Cache Search Blogs/Feeds Translate Track Post Develop Misc

Alexa Whois Source Global Whois SurfWax Stumble Upon 7Search metaEUREKA Furl Delicio.us Gibeo Semantic data extractor FyberSearch [Info](#)

---

**SurfWax SiteSnaps™** (patent pending) [\[ Help \]](#)

National Library of Australia  
<http://www.nla.gov.au>  
Links: 32 • Images: 29 • Words: 303 • Forms: 0

Author Summary

Our collections and services underpin Australian cultural life and intellectual pursuits. We are the preminent source for the documentary record of Australia and its place in the world. This site provides access to our catalogue and links to other information resources, details of our collections and services, upcoming events at the Library; the NLA Shop and information on our initiatives in the field of information management.

Site's FocusWords

<ul style="list-style-type: none"> <li>Q <a href="#">Australian Libraries Gateway</a></li> <li>Q <a href="#">ال مكتبة الوطنية الأسترالية</a></li> <li>Q <a href="#">collections of Australia's national</a></li> <li>Q <a href="#">combined catalogues of Australian libraries</a></li> <li>Q <a href="#">conferences and meetings</a></li> <li>Q <a href="#">exhibitions and publications</a></li> <li>Q <a href="#">Exhibitions Displays</a></li> <li>Q <a href="#">GUIDES INDEXES DATABASES</a></li> <li>Q <a href="#">HOME CATALOGUE ASK</a></li> <li>Q <a href="#">information journey</a></li> <li>Q <a href="#">history</a></li> <li>Q <a href="#">Libraries Australia</a></li> </ul>	<ul style="list-style-type: none"> <li>Q <a href="#">Library</a></li> <li>Q <a href="#">Libraries</a></li> <li>Q <a href="#">million items</a></li> <li>Q <a href="#">MultiAustralia</a></li> <li>Q <a href="#">National Library of Australia</a></li> <li>Q <a href="#">phone services</a></li> <li>Q <a href="#">phone system the main</a></li> <li>Q <a href="#">Picture Australia</a></li> <li>Q <a href="#">rare and limited editions</a></li> <li>Q <a href="#">rare of Australia's musical gems</a></li> <li>Q <a href="#">watercolourz</a></li> </ul>
--	---

UNCLASSIFIED//FOR OFFICIAL USE ONLY

The next tab is a link to **Stumble Upon**, a collaborative bookmarking service. If people who belong to the service have recommended a site, Stumble Upon will show who they are, any comments, the category it is in, and the page size. The tab for 7Search brings up some interesting results, including the website's traffic ranking (how many people visit it), number of links to the site, link popularity, language, area of service, contact information, Open Directory category, and "TrustGauge." **TrustGauge** is a commercial program that measures how "trustworthy" a site is in terms of such things as the amount and quality of contact information, secure billing, third party validations (e.g., Truste seal), and what people think of the site.


Fagan Funder's URLInfo <http://www.fateh.net> [view page](#) [clear](#) [About URLInfo](#)

[General](#) [Links](#) [Similar](#) [Cache](#) [Search](#) [Blogs/Feeds](#) [Translate](#) [Track](#) [Post](#) [Develop](#) [Misc](#)  
[Alexa](#) [Whois](#) [Source](#) [Global Whois](#) [SurfWear](#) [StumbleUpon](#) [7Search](#) [metaEUREKA](#) [Furl](#) [Delicio.us](#) [Gibco](#) [Semantic data extractor](#) [FyberSearch](#) [Info!](#)

**Seal of Approval from ValidatedSite.com**  
 Communicate the integrity of your online business, build trust with your site visitors, and improve your overall image so consumers can trust you!

**TrustGauge Domain Info for "fateh.net"**

Fateh  
 Palestian movement founded by Yasser Arafat.

<b>Contact Information:</b>	<b>Organization Rating Information:</b>
<b>Site Information:</b>	<b>Traffic ranking among all sites:</b>
Language: N/A	394,483 ●
Area of Service: N/A	<b>Ranking in its category:</b> 7
	provided free by <a href="#">Ranking.com</a>
<a href="#">See how fateh.net used to look</a>	<b>TrustGauge:</b> 
<b>Webmaster Information</b>	provided free by <a href="#">TrustGauge.com</a>
Increase your TrustGauge™ Score!	<a href="#">click to install in your browser</a>
<a href="#">Add your website information here</a>	<b>Links pointing to this site:</b> 1,738
	<b>Link popularity ranking:</b> 70,776
	provided free by <a href="#">LinkToYou.com</a>
	<b>Organization Reviews:</b> N/A
	<b>Website Reviews:</b> N/A
	<b>Reviews &amp; Complaints</b> <small>(new)</small> <a href="#">read / write</a>

Ranking.com ranks this site #7 in its category:  
\* [Sort by Government/Politics/Parties and Groups](#)

[Close](#)

The remainder of the **General** tab links are:

metaEUREKA

metaEUREKA shows information about the page (last modified date, page size), meta information (description, keywords, author), web server information, and the number of backlinks

Furl

Furl is a collaborative bookmarking system. This tool allows you to see the comments others have written about a webpage.

Delicio.us

Delicio.us is a collaborative bookmarking system. This tool allows you to see the comments others have written about a webpage.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

### Gibeo

Gibeo allows anyone to annotate any part of a web page, and others can comment on the annotation. Gibeo requires registration.

### Semantic data extractor

The Semantic data extractor finds information about a page (metadata, page outline) by looking at its HTML code.

The next tab is for **Links**. This is pretty straightforward. The first two links are to Yahoo, the first for the *link:* command (links to a specific page) and the second for the Yahoo Site Explorer or alternately *linkdomain:* command (links to a website). The next is the Live Search (MSN) *link:* search, and then the Google *link:* command, which no longer shows all links as it once did. Gigablast does not show all links to a page, either.

The links from **blogs** is a very useful service because it lets you check to see if a website is mentioned in a number of weblogs very quickly (I expect Technorati to give the best results).

### Blogpulse

Intelliseek's Blog search (was not working when I tried it)

### Bloglines

Backlinks from blogs known to Bloglines, an online RSS/Atom aggregator.

**Blogdex** is defunct.

### Technorati

Backlinks from blogs known to the Technorati blog indexer. Each result is shown with an extract containing the link.

### Feedster

Backlinks from blogs known to the Feedster RSS/Atom search engine.

### BlogDigger

Backlinks from blogs known to the BlogDigger RSS/Atom search engine.

### Waypath

Backlinks from blogs, known to the Waypath blog indexer, each is listed with the date that the link was first seen and an extract from the page. Unlike some other backlinks tools, Waypath lists the permalinks rather than blog home pages.

### Daypop

Backlinks from blogs and news websites known to the Daypop search engine.

### BlogRolling

BlogRolling is a service for bloggers to include blogrolls (lists of blogs) on their own blogs. This shows what users include the given site on their blogroll.

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~Popdex

Backlinks from blogs (as well as the date of linkage) known to the Popdex blog indexer.

The **Similar** tab is not entirely self-explanatory. Alexa, UCmore, Furl, and Google all try to show related or similar websites, though not in the same way. **Alexa** shows 'people who visit this page also visit...'; **UCmore** clusters related pages by topic; **Furl** is a collaborative bookmarking tool, so it only shows pages bookmarked by the same person (of dubious use); and Google's related pages is, in Fagan's and my opinion, of poor quality. Google News will show related news articles, but only if the original article has been indexed by Google News. The **Waypath** tool looks for blog entries about a website, and Waypath is showing no links to <http://www.google.com> and two hits on <http://www.microsoft.com>. There is obviously a problem with this specific search.

The **Cache** tab is much more useful at this time. Fagan has done us all the great service of bringing the search tools that cache webpages together so they can be searched from one convenient interface. Also, **URLinfo makes it possible to see Google's cached pages without images, style sheets, or forms with Google (plain)**. Openfind is an Asian search engine and does not yet have an English version. I was unable to figure out how their caching works because of the language barrier. For news and blogs Daypop caches each page it crawls. "Its cache is often the most up-to-date copy of the page, and it shows the exact time that the copy was made."

Here's the low-down on the other general cache tools at Fagan Finder:

Internet Archive

The Internet Archive has been crawling the web and caching pages since 1996. The Wayback Machine allows you to view the copies made during any of those crawls, and also to compare any two versions of the same page.

Google

When Google crawls the web, it stores a copy of each web page. This is the most recent copy. This can also be used as a means of viewing some non-HTML files converted to HTML.

Google (plain)

Google's stripped cache, with images, styles (style sheets), and forms removed.

Gigablast

Gigablast does not provide direct access to its cache. You must follow the link labeled [archived copy]. Gigablast's cache shows the date on which the copy was made.



~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

#### Openfind

Openfind is an Asian search engine; their English version is under construction.

#### Spurl

Spurl is a collaborative online bookmarking tool. Whenever someone using Spurl bookmarks a page, a cached copy is stored. So Spurl may contain many different copies of the same page on different dates and times, which can be accessed from a selection box at the top of any Spurl cached page.

#### IncyWincy

This is the cached version of a web page from when it was last crawled by IncyWincy. That date is shown at the top of the page.

#### Scrub The Web

Cached version of the page from the Scrub The web search engine.

#### Ay-Up

Cached version of the page from the Ay-Up search engine.

#### Objects Search

Cached version of the page from the Objects Search engine. Objects Search has a small index, so don't expect every page to be cached. After using this tool, follow the link below the page you want labeled 'cached.'

#### SearchSpider

This is the cached version of a web page from when it was last crawled by SearchSpider. Most pages appear to have been last cached during July 2003.

The **Search** section is pretty much self-explanatory, except that MSN searches Live and Teoma searches Ask. Fagan explains the **Blogs/Feeds** tab very well for those who are interested in searching weblogs and RSS or Atom news feeds. The **Translate** tab simply sends your request to Fagan Finder's superb Translation Wizard discussed in the online dictionary and translators' section. The **Track** and **Post** tabs are in general not going to be useful for most of you in your work environment. The **Develop** tab offers an excellent selection of web authoring resources such as validation, editing, spelling, cacheability, and keyword analysis tools. One tool users may not recognize and which could prove quite useful is **Traffic** from Alexa. Here's Fagan's description:

'Shows a (logarithmic) graph of a website's (not a web page) popularity over time, as determined by Alexa. Alexa gathers this data from users of their toolbar. The six-month graph is shown by default. You can also use this tool to compare the popularity of a second website. Also shown are popular subdomains, reach per million users, average page views per user, etc.'

I find the graphic representation is so much clearer than the results from a tool such as Google PageRank (which is not a Google product, by the way). Traffic also lets

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

users compare two sites and shows you "Where do people go" on the site. It's a gold mine of data about the sites in Alexa's top 100,000; unfortunately, most of the sites I wanted to research were not in that top group, so no statistics were available when a site fell below the 100,000 threshold.

In case the Google PageRank tool confuses you, it normally requires users to download and install the Google Toolbar. However, you can access the Google PageRank option from URLInfo without the Google Toolbar. The results look rather mysterious, but the PageRank is there. In the following example, AOL's home page has a page rank of 8 (where 10 is the highest...and Google gets a 10 ranking, by the way):

<http://www.aol.com>

PR Toolbar: 9

PR Actual: 9

Finally, under **Misc** you'll find the tools that didn't quite fit anywhere else. One word of caution about **BugMeNot**: this is a service for sharing login information for websites that require user registration and, as such, its ethics is questionable. I do not recommend using it. It may also violate organizational Internet usage rules.

I think URLInfo will prove to be a very useful if not indispensable tool for researchers, but I also think the key to using it effectively is not using every bell and whistle.

Fagan Finder's URLInfo beta

<http://www.faganfinder.com/urlinfo/>

---

## Wikipedia

---

Wikipedia

<http://en.wikipedia.org/>

The 2007 edition is the first to include a separate section on and discussion of Wikipedia and the entire "wiki" phenomenon. The extraordinary growth and success of Wikipedia demand recognition and comment. Although the numbers change constantly, in mid-2006, Wikipedia sites were the twelfth most visited Internet sites among US properties, *up over 300 percent from the previous year*.<sup>71</sup> On March 1, 2006, Wikipedia reached one million articles, and "the site receives as many as

---

<sup>71</sup> Safa Rashtchy, et al., "Silk Road: Solid Search Results Could Boost the Sector," PiperJaffray Industry Note, 10 July 2006, available at John Battelle's Searchblog, <http://battellemedia.com/archives/Rashtchy%20-%20Silk%20Road%20200710.pdf> [PDF] (14 November 2006).

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

fourteen thousand hits per second.”<sup>72</sup> Just what is the Wikipedia itself and the wiki concept in general that have led to a level of success that is nothing short of astounding? For an excellent overview, I turn to my colleague Diane White’s article from an internal publication many of you read, *The WorthWhile Web*. In the May 2006 edition, Diane wrote:

“In true Ouroborosian fashion, the Wikipedia defines itself as a ‘multilingual Web-based free-content encyclopedia...written collaboratively by volunteers, allowing most articles to be changed by anyone with access to a web browser and an Internet connection.’ It exists as a wiki, which again Wikipedia self-defines as ‘a type of website that allows anyone visiting the site to add, remove or otherwise edit all content very quickly and easily, often without the need for registration.’ Truly collaboration to the extreme, wikis are the latest trend in open-ended community involvement and public debate. But it also conjures fears of authority and validity run amok, and general mischief and vandalism. Wikis are popping up everywhere; but just what are they, and how did they become so ubiquitous? More to the point, can they be trusted, or are they just the work of a few people with big egos and lots of time?...The term wiki is a shortened form of the Hawaiian language term *wiki wiki*, which is commonly used as an adjective to denote something quick or fast. It is also sometimes interpreted as the backronym for *What I Know Is*. The invention of the wiki is credited to Ward Cunningham, author of the book, *The Wiki Way* (Addison-Wesley Longman, March 2001, ISBN 0-201-71499-X). The first wiki, WikiWikiWeb, was created in 1994 and installed on the web by Cunningham in 1995.”<sup>73</sup>

“Once begun, almost anyone can edit a wiki, often without actually registering to do so. Wikis can be on any subject, on every subject, and in multiple languages. The most famous wiki, Wikipedia, was begun in 2001, initially as part of a broader, peer-reviewed project and later as a stand-alone, ‘neutral point of view’ product. Guided from the beginning by Larry Sanger and Jimmy Wales, today it is available in over 100 languages, with over 1 million articles in the English edition alone...”

### “Questions of Validity and Reliability

“But can Wikis be trusted? From almost the beginning, people have questioned the wiki’s seemingly radical departure from traditional methods of scholarship; that is, the use of a community of interested parties instead of the work of appointed experts. In the December 2005 issue of *Nature*, there began a major debate over which site was more ‘right,’ Wikipedia or the fee-based (\$85/year) Britannica Online; with the conclusion being that ‘Wikipedia comes close to Britannica in terms of the

---

<sup>72</sup> Stacy Shiff, “Can Wikipedia Conquer Expertise?” *The New Yorker*, 24 July 2006, <[http://www.newyorker.com/fact/content/articles/060731fa\\_fact](http://www.newyorker.com/fact/content/articles/060731fa_fact)> (14 November 2006).

<sup>73</sup> “Wikipedia,” Wikipedia: The Free Encyclopedia, <<http://en.wikipedia.org/wiki/Wikipedia>> (23 August 2006).

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

accuracy of its science entries.<sup>74</sup> From there it has escalated, with refutations and calls for retraction from *Encyclopaedia Britannica* and heated responses from *Nature*. Wikipedia itself has steered clear of this particular fray; however, it does attempt to respond to criticism and has a page on its site for common criticisms. It also addresses issues such as copyright, vandalism, and authorship.

“So what’s the bottom line? The same as it’s always been. When performing thorough research, be it Internet-based or otherwise, *the onus is always on the researcher to check sources, validity, and authority*. The speed and relative ease at which changes can be made to a wiki, while good for consensus correction and corroboration, are not so good for measured and thoughtful debate. A number of articles in Wikipedia are sourced, but many are not, and just because it’s on the Internet, does not mean it is true. In addition, merely because it’s free does not mean Wikipedia is more suspect and Britannica is more reliable. There is an argument to be made for being so passionate about a topic that you feel the need to share that passion with the world. But one man’s passion is also another’s conceit. There is a counter to every argument, a rebuttal to every claim.

“Like it or not, wikis and wiki behaviors have entered the mainstream, just like blogs and MySpace and the iPod. Love it or hate it, if you are involved in open source research you need to know about wikis.”<sup>75</sup>

### **The Wikipedia Itself: The Good, the Bad, and the Dubious**

As Diane White clearly indicates, there are many, many wikis now available on the Internet, and their numbers continue to increase at present. I want to focus on Wikipedia itself because it remains the center of the wiki universe and thus far shows no signs of decline. Many Wikipedia critics mourn the decline of traditional encyclopedias because they are thinking of an encyclopedia such as *Britannica* in its current form, that is, “the most authoritative source of...information and ideas,” the “definitive source of knowledge.”<sup>76</sup> According to Tom Panelas, *Britannica’s* Director of Corporate Communications, “We can’t cover as many things as they [Wikipedia] do but we wouldn’t even try to. What they do is very different from what we do. We don’t have an article on extreme ironing, and we shouldn’t.”<sup>77</sup>

---

<sup>74</sup> Jim Giles, “Internet Encyclopaedias Go Head to Head,” *Nature*, 14 December 2005 (last updated 28 March 2006), <<http://www.nature.com/news/2005/051212/full/438900a.html>> (14 November 2006).

<sup>75</sup> Diane White, “Wikis and the Wikipedia,” *The WorthWhile Web*, May 2006, <<http://www.fggm.osis.gov/Worthwhile/archive/20060501.html>>.

<sup>76</sup> Paula Berinstein, “Wikipedia and Britannica: The Kids Are All Right (And So’s the Old Man),” *Information Today*, March 2006, <<http://www.infotoday.com/searcher/mar06/berinstein.shtml>> (11 September 2006).

<sup>77</sup> Berinstein.

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

However, *Britannica* today (and by extension any other encyclopedia) is very different from *Britannicas* of the past. In thinking about this controversy, I was reminded of a passage in *The Fatal Shore*, Robert Hughes' masterpiece about the founding of modern Australia. Hughes writes about one transported convict, Thomas Palmer, who finished his sentence and went into business with his close friend John Boston.<sup>78</sup> Neither man had much business experience, "but they possessed a singular advantage: ***the only encyclopedia in the colony***. With it, they taught themselves to make beer. Then they learned how to make soap. Next they looked up 'ship' and, after some trial and error, contrived to build a somewhat cranky but adequate small vessel for trading stores to Norfolk Island." [emphasis added]<sup>79</sup> Their lone encyclopedia probably made it possible for these men not merely to survive but to thrive in this perilous new world.

The modern encyclopedia is very different from the encyclopedias of earlier centuries, which bear rather more resemblance to the Wikipedia than to the current *Britannica* in that the older encyclopedias were not only "sources of knowledge" but also "practical" how-to guides and almanacs. In other ways, however, encyclopedias are and always were quite different from the Wikipedia. They have always relied upon paid experts whose work is reviewed and edited. And they have always been for-profit enterprises.

Wikipedia *relies almost entirely upon individual users* to create, edit, maintain, and often argue about its entries. It is free and carries no advertising; it is a nonprofit and has a tiny staff.

- For practical purposes, Wikipedia has *no physical limits*: it could conceivably continue to expand indefinitely, something no print encyclopedia could ever do.
- Its content is "open," that is, *almost any topic can be included*; traditional encyclopedias generally do not include "how-to" instructions ("How to draw a diagram with Microsoft Word"), new or transient popular culture ("24: The TV Series"), or breaking stories ("JonBenét Ramsey").
- Wikipedia's heavy emphasis on current events and popular culture bespeak a prejudice of the present at the expense of the important: it favors *the fashionable over the important*.

---

<sup>78</sup>In what must be one of the most profound examples of friendship since Damon and Pythias, Boston actually traveled voluntarily with his wife to New South Wales to "keep Palmer company." Anyone who has read about a sea voyage from England to Australia at that time knows the trip in and of itself was a major sacrifice. Robert Hughes, *The Fatal Shore* (New York: Vintage Books, 1988), 180.

<sup>79</sup> Hughes, 180.

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

- Wikipedias are available in 229 languages. These are not always just translations of the English language Wikipedia but often contain their own content.
- Wikipedia's eight-word self-description—"*neutral and unbiased compilation of previously written, verifiable facts*"—usually keeps out articles about "my funniest dreams and what they mean" (no original "research" allowed), but firefights over controversial topics and outright vandalism occur on a regular basis.
- In 2006 comedian Steven Colbert's amusing rant against "wikiality" and "truthiness," i.e., that *reality and truth are what the most people say they are*, and his charge to his viewers to change a Wikipedia article on African elephants caused the entire site to go down temporarily. His point is well taken: if enough Wikipedians agree that the earth is flat, then the Wikipedia will reflect that "wikiality." While that is an absurd example, people vehemently (and often violently) disagree over the most basic topics (try to think of anything that isn't controversial).
- Wikipedia "does not favor the Ph.D. over the well-read fifteen year old."<sup>80</sup> While the *democratization of knowledge and information* has a certain appeal, the fact that Wikipedia pages dealing with policies, rules, administration, coordination, and other metadata now comprise thirty percent of Wikipedia indicates that the free-for-all nature of Wikipedia is giving ground to the harsh reality of the need for "crowd control." There is a fine line between democracy and mob rule.
- There is *no "weighting" of the relative significance* of any topic: compare the Wikipedia entries on the Beatles v. Boethius. Judged by sheer quantity, articles on popular culture far exceed those of traditional scholarly topics. Given its potentially limitless size, this may not be a drawback, but if everything from "The Simpsons" to "The Nicomachean Ethics" is on an equal footing, then aren't we back to the Colbert criticism that all objective standards are obliterated?
- Diane White correctly identified Wikipedia's "ouroborosian" nature: it is *fiercely self-referential* in that all the works cited in this creature of the Internet are also on the Internet.
- Some critics maintain that emergent enterprises such as Wikipedia reflect an "online collectivism" that lead to a kind of group think and produce poor quality results that both appeal to and are a product of the lowest common

---

<sup>80</sup> Shiff, "Can Wikipedia Conquer Expertise?"

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

denominator. For more on this topic, read Jaron Lanier's now famous think piece "Digital Maoism" and the many responses to it on Edge.org.<sup>81</sup>

- Finally, Wikipedia has *no editorial quality review*. Traditional encyclopedias do not guarantee zero mistakes; what they do promise are "strong scholarship, sound judgment, and disciplined editorial review."<sup>82</sup>

All this being said, nothing is going to stop people from using Wikipedia as a reference, in many cases, their primary source for information. Some search engines—for example, Ask—now proudly display Wikipedia responses at the top of the results list. Most will return Wikipedia links near the top. The best advice I can give you vis-à-vis Wikipedia and related community generated resources is as follows:

- Use multiple sources: Do not as a rule rely on Wikipedia as your sole reference or source of information. Any Wikipedia entry that is not well sourced should raise a red flag.
- Trust but Verify: Look for verification of Wikipedia information from sources such as traditional references that have been through editorial review: encyclopedias, dictionaries, scholarly (peer-reviewed) publications, university websites, books, etc.
- Follow those links: The best thing about Wikipedia in my opinion are the *external links* from entries; with the virtual demise of web directories, Wikipedia fills that void by supplying excellent links to what are often the best websites on a topic.
- Be skeptical: The more controversial the topic, the more skepticism you need to apply to the Wikipedia entry. For example, the article "Asteroid" is quite well done, but there isn't quite the controversy about that topic that there is about, say, Hezbollah, an article that was locked because of vandalism.

Wikipedia has an internal search option, but as any Wikipedia user knows, it is not the best way to search Wikipedia. First, unlike virtually every search engine on the web, its default is OR not AND, meaning it searches for ANY of the terms you enter. To search Wikipedia content you are better off using a separate search engine, either one of the major search engines or a specialty search tool designed to search Wikipedia.

---

<sup>81</sup> Jaron Lanier, "Digital Maoism," Edge.org, June 2006, <[http://www.edge.org/3rd\\_culture/lanier06/lanier06\\_index.html](http://www.edge.org/3rd_culture/lanier06/lanier06_index.html)> (14 November 2006).

<sup>82</sup> "Britannica Rips Nature Magazine on Accuracy Study," Encyclopedia Britannica Corporate Website, 24 March 2006, <<http://corporate.britannica.com/press/releases/nature.html>> (14 November 2006).

~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

# How to Check Links from Wikipedia

Wikipedia has a special page that lets users easily check to see which Wikipedia pages have links to a specific webpage. It even includes a wildcard function. Here's how it works. You can search for a specific link or a very general one. Here are examples of both, starting with a general search using a wildcard [\* .nasa.gov]:

The screenshot shows the Wikipedia 'Special:Web Links' search interface. At the top, it says 'special page' and 'Your continued donations keep Wikipedia running!'. The search term entered is '\* .nasa.gov'. Below the search bar, it says 'Showing below up to 50 results starting with #1.' and 'View (previous 50) (next 50) (20 | 50 | 100 | 250 | 500)'. The results list 26 items, each with a number and a URL. Many of the URLs are followed by text indicating they were linked from a specific Wikipedia page, such as 'linked from User:Jarnescarr' or 'linked from User:talk:Awc002'. The results include various NASA.gov pages and links to NASA.gov pages from other websites.

## Search Web Links at Wikipedia

<http://en.wikipedia.org/w/index.php?title=Special%3ALinksearch>

There are literally hundreds of results for this query. However, you can limit your search to a specific page [leonid.arc.nasa.gov/meteor.html], which in this case returns one result:



~~UNCLASSIFIED//FOR OFFICIAL USE ONLY~~

This is a very useful tool if you need to find out what pages in Wikipedia link to a specific site. Be sure to follow these basic rules for using this feature:

1. a full domain name, e.g., [www.nasa.gov] (this will only find links to this specific domain) OR
2. a partial domain name with a wildcard, e.g., [\*.nasa.gov] (this will find links to any site at nasa.gov, such as ase.arc.nasa.gov) OR
3. a full domain name plus directory and/or webpage, e.g., [leonid.arc.nasa.gov/meteor.html] (this will find links only to this specific webpage)

Some Wikipedias other than the English language version have a similar page. For example, the German language Wikipedia link search page is:

<<http://de.wikipedia.org/wiki/Spezial:Linksearch>>.

If you use the English Wikipedia link below and substitute the appropriate language digraph for the "en," you can find these non-English language link search pages. See this page <[http://meta.wikimedia.org/wiki/List\\_of\\_Wikipedias](http://meta.wikimedia.org/wiki/List_of_Wikipedias)> for all the Wikipedias and the appropriate digraph.

Search English Wikipedia Web Links

<http://en.wikipedia.org/w/index.php?title=Special%3ALinksearch>

## Wiki Search Engines

FUTEF (Beta)

<http://fufef.com/>

FUTEF, which uses its own proprietary search engine, provides both a list of relevant articles but also a list of related categories that can be used to further refine a search. **FUTEF handles non-Latin searches, something not every Wikipedia search engine can do.** Try a search on Σμύρνη and you will see that FUTEF finds this term in the English language Wikipedia.

Qwika

<http://www.qwika.com/>

Qwika indexes [English](#), [German](#), [French](#), [Japanese](#), [Italian](#), [Dutch](#), [Portuguese](#), [Spanish](#), [Greek](#), [Korean](#), [Chinese](#) and [Russian](#) wikis; the original content is combined with machine translated content to/from English. However, when searching for a non-Latin term Qwika will only find that term in the international Wikipedia not the English language Wikipedia even if it is, e.g., Σμύρνη.

LuMirX

<http://wiki.lumrix.net/>

LuMirX uses [AJAX](#) technology and searches English, German, Japanese, French, Polish, Italian, Swedish, Dutch, Portuguese, Russian, Danish, Spanish, Finnish, Norwegian, Hungarian, Turkish, and Chinese Wikipedias. However, when searching for a non-Latin term LuMirX will only find that term in the international Wikipedia not the English language Wikipedia even if it is there, e.g., Çeşme.


Clusty's Wikipedia Search (English only)

<http://wiki.clusty.com/>

One of the best Wikipedia search engines, Clusty not only searches the Wikipedia, it clusters the results into easy to understand categories that make it possible to zero in on the appropriate subtopic. Its main drawback is that the search is limited to the English-language Wikipedia.

The screenshot shows the Clusty search interface. At the top, there is a navigation bar with links for Web+, News, Images, Shopping, Wikipedia, Blogs, Jobs, and Customizer. Below this is a search bar containing the text 'plutarch'. To the right of the search bar are buttons for 'Cluster' and 'Advanced Preferences'. Below the search bar, it says 'Cluster by: Topics' and 'Top 200 results of at least 568 retrieved for the query plutarch (Details)'. On the left side, there is a list of 'All Results (200)' with various categories and counts: Parallel Lives (36), Claudius (2), Alexander (25), Letter, Moralia (5), Plutarch crater (3), Cicero (11), Spartan (12), Numa Pompilius (10), Mele, Exhalus (5), Pompey, Marcus (6), and a link for 'more | all clusters'. The main content area displays the top search results:

- 1. Plutarch**

 **Mestrius Plutarchus** (ca. 46-ca. post 127) was a [Greek historian, biographer, and essayist](#). Born in the small town of Chaeronea, in the Greek region known as Boeotia, probably during the reign of the Roman Emperor [Claudius](#), Plutarch travelled widely in the [Mediterranean world](#), including twice to [Rome](#). He had a number of influential Roman friends, including [Socius Senecio](#) and [Fundanus](#), both important [Senators](#), to whom some of his later writings were dedicated. He lived most of his life at Chaeronea, and was initiated into the [mysteries](#) of the Greek god [Apollo](#). However his duties as the senior of the two priests of [Apollo](#) at the [Oracle of Delphi](#) (where he was responsible for interpreting the auguries of the [Pythia](#) or priestess/oracle) apparently occupied little of his time - he led a most active social and civic life and produced an incredible body of writings, much of which is still extant. en.wikipedia.org/wiki/Plutarch
- 2. Plutarch of Eretria**

**Plutarch** (in Greek [Πλούταρχος](#), lived [4th century BC](#)) was a [tyrant](#) of [Eretria](#) in [Euboea](#). Whether he was the immediate successor of [Therison](#), and also whether he was in any way connected with him by blood, are points which we have no means of ascertaining. Trusting perhaps to the influence of his friend [Meidias](#), he applied to the [Athenians](#) in [354 BC](#) for aid against his rival, [Callias of Chalcis](#), who had allied himself with [Philip of Macedon](#). The application was granted in spite of the resistance of [Demosthenes](#), and the command of the expedition was entrusted to [Phocion](#), who defeated Callias at [Tamyrae](#) in [350 BC](#). But the conduct of Plutarch in the battle had placed the Athenians in great jeopardy, and though it may have been nothing more than rashness, Phocion would seem to have regarded it as [treachery](#), for he thenceforth treated Plutarch as an enemy and expelled him from Eretria. en.wikipedia.org/wiki/Plutarch\_of\_Eretria
- 3. Plutarch (crater)**

**Plutarch** is a [lunar impact crater](#) that lies near the north-northeastern limb of the Moon, just to the south of the irregular [Seneca crater](#). To the southeast is the flooded [Cannon crater](#). The proximity of this crater to the limb causes it to appear foreshortened when viewed from the [Earth](#), but it is actually a circular formation. en.wikipedia.org/wiki/Plutarch\_(crater)
- 4. Parallel Lives**

**Plutarch's Lives of the Noble Greeks and Romans** is a series of biographies of famous men, arranged in tandem to illuminate their common moral virtues or . . . lives.As he explains in the first paragraph of his [Life of Alexander](#) , **Plutarch** was not concerned with writing histories, as such, but in exploring the influence . . . the historic figures, there are also links to several on-line versions of **Plutarch's Lives** , see also "Other links" section below O Dryden is famous for . . .

UNCLASSIFIED//FOR OFFICIAL USE ONLY

Wikiseek

<http://wikiseek.com/>

Launched in early 2007, Wikiseek was created with the assistance of Wikipedia, although it is not a part of Wikipedia. "The contents of Wikiseek are restricted to Wikipedia pages and only those sites which are referenced within Wikipedia, making it an authoritative source of information less subject to spam and SEO schemes. Wikiseek utilizes Searchme's category refinement technology, providing suggested search refinements based on user tagging and categorization within Wikipedia, making results more relevant than conventional search engines." <<http://www.wikiseek.com/>> Wikiseek uses AJAX technology to create changing "tag clouds" of possible terms as you type.

This is a good way to find articles within English language Wikipedia and to search sites referenced in Wikipedia, but it is by no means a substitute for a general search engine. Results from Wikipedia are identified by the **W** icon. The tag cloud that appears at the top of each successful search is designed to show related categories to help users either narrow or broaden a search. Keep in mind these are user-generated tags, so many of them, e.g., "Japanese terms," do not correspond to Wikipedia categories.

The screenshot shows the Wikiseek search interface. At the top left is the Wikiseek logo with the tagline "A better way to search Wikipedia". To the right is a search bar containing the text "tsunami" and a "searchme™" button. Below the search bar is a tag cloud with "History of Southeast Asia" as the largest and most prominent tag. Other visible tags include "Disasters", "Tsunami", "Waves", "Weather", "Natural disasters", and "Natural hazards". Below the tag cloud, the search results are displayed in two columns. The left column is titled "All Results" and shows several search results, each starting with a "W" icon and a title, followed by a brief description and a URL. The right column is titled "Sponsored Links" and lists several advertisements, including "Surplus Military Tents", "Fun Activity Patches", "Tsunami Relief", "Tsunami Relief Patch In Stock and ready to ship", "Tsunami Relief Medical Relief for Tsunami Victims BBB Wise Giving Alliance", "News in Pictures", "Tsunami Victims", "Visit Western Union Today", and "The Coming World War".

The drawbacks to Wikiseek are that it only searches the English language version of Wikipedia and it **cannot parse non-Latin languages**. It touts itself as an "authoritative source of information less subject to spam and SEO schemes,"

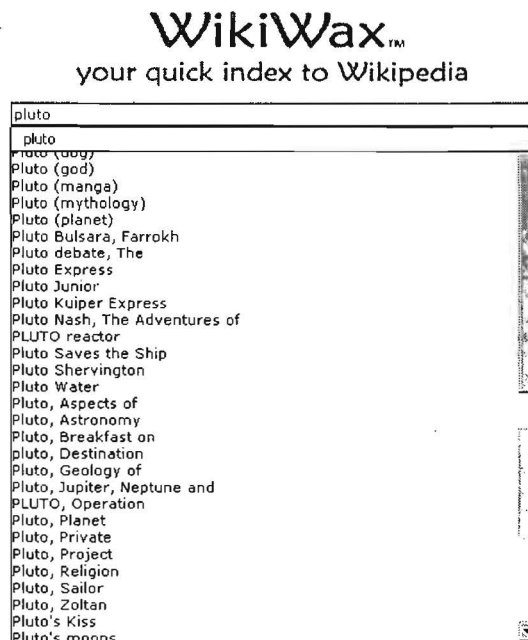
UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

but a search for [viagra] will quickly prove it is no better (in fact worse) than the major search engines in filtering spam. There are no preferences to change the number of results, for example, or to limit the search only to Wikipedia or only to links, but since Wikiseek is still in Beta, these features may appear later. Wikiseek also offers a Firefox plug-in to add Wikiseek to the Wikipedia search form on all Wikipedia pages.

WikiWax

<http://www.wikiwax.com/>

WikiWax also uses "Look Ahead" AJAX technology to show very extensive lists of dynamically generated related terms. However, WikiWax cannot parse non-Latin search terms, e.g., Σμύρνη.



**Using Search Engines to Search Wikipedia**

Yahoo now includes Quick Links for any Wikipedia results. For example, a Yahoo search for [internet] will return Wikipedia as result number eight and will include "Quick Links" to specific Wikipedia articles on this topic:

- 8. [Internet - Wikipedia, the free encyclopedia](#)  
 The **Internet** is the worldwide, publicly accessible network of interconnected computer networks that transmit data by packet-switching using the standard **Internet Protocol** ...  
 Quick Links: [Creation of the Internet](#) - [Today's Internet](#) - [Internet protocols](#)  
[en.wikipedia.org/wiki/Internet](#) - 97k - Cached - More from this site - Save

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

You can further restrict the search to Wikipedia by clicking on "More from this site," which is an excellent way to search Wikipedia using Yahoo:

Web | Images | Video | Audio | Directory | Local | News | Shopping | More »

**YAHOO! SEARCH** internet

Answers | My Web | Search Services | Advanced Search | Preferences

Search Results 1 - 10 of about 232,000 for internet - 0.05 sec. (About this page)

Also try: [internet explorer](#), [free internet](#), [internet radio](#), [internet tv](#) More...

Your search was restricted to "en.wikipedia.org". For more matches, try searching [all sources](#) instead.

- Internet - Wikipedia, the free encyclopedia**  
The **Internet** (also known simply as the Net) can be briefly understood as "a network of networks". Specifically, it is the worldwide, publicly accessible network of ...  
Quick Links: [Creation of the Internet](#) - [Today's Internet](#) - [Internet protocols](#)  
[en.wikipedia.org/wiki/Internet](#) - 37k - [Cached](#) - [Save](#)
- Internet Explorer - Wikipedia**  
Encyclopedia article about **Internet Explorer**, Microsoft's web browser. Covers its history, features, component architecture, common criticisms, and market share.  
Category: [History of Internet Explorer \(IE\)](#)  
[en.wikipedia.org/wiki/Internet\\_Explorer](#) - 37k - [Cached](#) - [Save](#)
- Internet troll - Wikipedia, the free encyclopedia**  
In **Internet** terminology, a troll is someone who comes into an established community such as an online discussion forum, and posts inflammatory, rude, repetitive or ...  
Quick Links: [Etymology](#) - [Vicious circles](#) - [Troll culture](#)  
[en.wikipedia.org/wiki/Internet\\_troll](#) - 37k - [Cached](#) - [Save](#)
- History of the Internet - Wikipedia, the free encyclopedia**  
The history of the **Internet** dates back to the early development of communication networks. The idea of a computer network intended to allow general communication among ...  
Quick Links: [Before the Internet](#) - [A lack of inter-networking](#) - [Three terminals and an ARPA](#)  
[en.wikipedia.org/wiki/history\\_of\\_the\\_Internet](#) - 56k - [Cached](#) - [Save](#)

You can also use the **site:** syntax to search just the Wikipedia (or Wikipedias, if you like) in:

- Yahoo <http://search.yahoo.com/>
- Google <http://www.google.com/>
- Ask <http://www.ask.com/>
- Windows Live Search <http://www.live.com/?searchonly=true>
- A9 <http://a9.com/>
- Gigablast <http://www.gigablast.com/>
- Exalead <http://www.exalead.com/search>
- Clusty (site: and host: are interchangeable, but Clusty has a special Wikipedia search option) <http://clusty.com/>

This is an especially useful option for non-Latin searches, such as [site:wikipedia.org Çeşme], which returns results not only from the English and Turkish Wikipedias but from the German and Serbian Wikipedias as well:

UNCLASSIFIED//~~FOR OFFICIAL USE ONLY~~

UNCLASSIFIED//FOR OFFICIAL USE ONLY

[Web](#) | [Images](#) | [Video](#) | [Audio](#) | [Directory](#) | [Local](#) | [News](#) | [Shopping](#) | [More »](#)  
**YAHOO! SEARCH**    
[Answers](#) | [Search Services](#) | [Advanced Set](#)

Search Results

1 - 100 of about 311 from wikipedia.org for **Çeşme** - 1.63 sec

- Çeşme** - [Wikipedia, the free encyclopedia](#)  
Çeşme is a town on the west coast of Turkey and one of the districts of Izmir Province. It is a prominent center of international tourism in Turkey and is famous for ...  
Quick Links: [See also](#) - [External links](#)  
en.wikipedia.org/wiki/Çeşme - 18k - [Cached](#) - [More from this site](#)
- İlica, Çeşme** - [Wikipedia, the free encyclopedia](#)  
İlica is a small village near Çeşme (pronounced Tcheshme), which is a district of Izmir Province in Çeşme Peninsula in the extreme western tip of Turkey.  
Quick Links: [Aegean region of Turkey geography stubs](#) - [Districts of Izmir](#) - [Izmir](#)  
en.wikipedia.org/wiki/İlica,\_Çeşme - 13k - [Cached](#) - [More from this site](#)
- Çeşme** - [Wikipedia](#) - [Translate this page](#)  
Çeşme [] ist ein Ferienort etwa 100 Kilometer westlich von Izmir. Der Name "Çeşme", zu Deutsch "Brunnen", leitet sich von der großen Zahl dieser ab.  
Quick Links: [Ort in der Türkei](#)  
de.wikipedia.org/wiki/Çeşme - 12k - [Cached](#) - [More from this site](#)
- Cezayirli Gazi Hasan Pasha** - [Wikipedia, the free encyclopedia](#)  
Cezayirli Gazi Hasan Pasha (1713-1790), (Hasan Pasha of Algiers) was an Ottoman grand vizier and a navy and army commander of the late 18th century.  
Quick Links: [References](#)  
en.wikipedia.org/wiki/Cezayirli\_Gazi\_Hasan\_Pasha - 16k - [Cached](#) - [More from this site](#)
- Çeşme** - [Wikipedi](#)  
... anlam ayrım sayfası, Çeşme kavramının farklı kullanımlarını ... Retrieved from "http://tr.wikipedia.org/wiki/Çeşme" Sayfa kategorisi: Anlam ayrım ...  
tr.wikipedia.org/wiki/Çeşme - 10k - [Cached](#) - [More from this site](#)
- Mehmet Culum** - [Wikipedia, the free encyclopedia](#)  
Mehmet Culum is a contemporary Turkish novelist who was born in a western town of Turkey called Çeşme in 1948. He studied political sciences at the University of Ankara.  
Quick Links: [European writer stubs](#) - [Turkish people stubs](#)  
en.wikipedia.org/wiki/Mehmet\_Culum - 13k - [Cached](#) - [More from this site](#)
- Marinas in Turkey** - [Wikipedia, the free encyclopedia](#)  
www.northdram

## Search Tip:

To search all Wikipedias:

**[site:wikipedia.org]**

To search language-specific Wikipedias:

site:DIGRAPH.wikipedia.org, e.g.,  
**[site:de.wikipedia.org nordafrika]**