



Reversing our searching habits

"Power searching without google"

Fravia's contribution for the Recon 2006 conference (Montreal, Canada - 16 June 2006)

This file dwells @ http://www.searchlores.org/reversing_our_searching_habits.htm

Introduction

Today's quarry ~ Look ma, no google!
Sliders, SEOs beasts, and Sinking google ~ Google weaknesses (and strengths)
The golden rules of searching ~ Conclusions
(Assignments ~ Forms)

<h2>Introduction</h2>	top
-----------------------	---------------------

Behold reversers' foresight, tremble before their insider knowledge...

Thank-you for the invitation. Please excuse three shortcomings of mine:

- 1) English is not even my first foreign language, so bear with me for my language shortcomings
- 2) I'm not and I'm not going to be politically correct, so bear with me for my attitude shortcomings

3) I won't use any boring powerpoint style presentation, so bear with me for my formal shortcomings

I am happy to speak -once again- to a bunch of reversers. I like reversers. Today I intend to discuss with you some 'alternative' searching techniques. As a proof of concept we will search some assembly books, given that assembly knowledge is a most powerful weapon, and in doing so we will, maybe, demonstrate a) that everything is on the web; b) that you can search effectively even without google; c) that some "alternative searching paths" can be quite useful for seekers.

Of course we will find anything: whatever. Any book you fancy is there for the take. In fact **anything that may have been digitized**: books, music, films, software, is there for the take. And not only "tangible" targets. **Solutions** are there for the take. Ideas. Methods. Techniques. Line of attacks. Tactics. Approaches. Strategies...

But we won't overdo it today. We'll just search some assembly-related books. And we will find our targets, of course: we will find our books.

Note that somebody, some time ago, told me -politely- that we should not download books that are patented, or have not yet been released in the public domain. OK. I can understand that, as silly and strange as it sounds: in fact patents and copyrights are just ways to negate all freedom of speech. Remember Freenet: **You cannot guarantee freedom of speech if you enforce copyright law...**

But "today let's obey", for the sake of our experiment, and therefore we will just have an innocent look at our targets on line. Don't download forbidden fruits, please, this whole talk is just an exercise, aimed at empowering reversers, giving them some sprinkles of **web-seeking cosmic power**. You should become seekers, not leechers.

(In general) reversers are a bunch of 'matter of fact' fellows: their most peculiar endowment is the capacity to reverse any "coded reality" around them.

In a world of codes no one is supposed to understand, reversers are among the very few that have the decoding knowledge or -maybe even better- the capacity **to find out any necessary decoding knowledge...**

As a consequence of the useless 'commercialisation' and hypermarketing of the web, some important searching tools, like the **main** search engines, are less and less capable of delivering useful results. This is -alas- true for google as well, whose results have kept getting worse over the last months, mainly due to the growing spamming and link-farming activities of the **SEO beasts**.

Since google is the most used among the main search engines, today's talk will concentrate on google's limits, but I surely do not intend to blame only google for problems that in most cases all other search engines have as well.

Note that google-marketing hypes, like google being 'your friend', or google being 'the mother of all modern search engines', with a 'nice brand' face and a 'commitment to the community of searchers', do not mean anything to us. Frills and hypes never work with reversers.

Google is just another commercial bastard without soul... exactly like microsoft. The only difference is that google is (still) way more useful than microsoft for knowledge gathering purposes.

We use search engines as starting tools, in order to gather knowledge, and we use tools only as long as they are useful and 'deliver' results. As soon as they become useless, we ditch them without regrets and switch to different ones.

This happened five years ago with **altavista**, once upon a time the favourite seekers' search engine (note that altavista still has a powerful & useful Boolean **NEAR** operator, btw), and it may happen anew and again with google. *Panta rei*.

The real point is that the **main search engines** are just ONE among many possible ways to search the web. As we will see together today. And the main search engines are not the friendliest tools for searching purposes either... just remember the very REASONS they exist... 🍷

Privacy concerns are only a small part of the "non friendliness" problem of all main search engines: search engines data provide a look into people's privacy, but privacy awareness & consumer protection have not kept (and will never keep) pace with this: if you allow private companies to collect such reams of data, any bogus attorney with half an excuse will easily get hold of them.

Just imagine the volume of personal information that a search engine can provide: every search query you've ever made on a given computer with a given browser.

(That's btw a good reason by itself to use not only different searching approaches, but also different browsers and different internet access providers (in general you are better served wardriving and avoiding **your own provider** as much as you can :-)) ... 🍷

Searching with google, using a GSM or buying stuff with your credit card is always the same sad story: you just smear your own data around for everybody and his dog to see, gather, collate and use against you.

But seekers know how to maintain some (relative) anonymity... 🍷.

And seekers also know how to maintain some (relative) speed wen searching. Your browser (**Opera**) is of tantamount importance for this, and can do wonders, if correctly trimmed... 🍷.

The real problem -when searching- is not anonymity and it is not speed: it is the **relevance**, **coherence** and **reliability** of our searching results.

Unfortunately we are now in a phase where even clueless bystanders notice how google is getting spammed much too much. It's usefulness as the "best and only" search engine has been therefore severely reduced.

This is sad, but it does not matter very much: we began long ago to prepare alternative paths to web-knowledge. Let's present some of them -today- to this community of distinguished reversers.

Today's quarry (the simple, direct way to shoot a quarry down)	top
--	-----

O gosh, I didn't know all these books were already in the public domain!

Ah! Assembly! The most dangerous weapon a reverser can master!

The following books about assembly could represent an appropriate "quarry" for today's example queries:

1. *Assembly Language Step-by-Step: Programming with DOS and Linux*, by Jeff Duntemann. (John Wiley, ISBN:0471375233), 2000
A fairly old text to begin with; should be available all over the web
2. *Linux Assembly Language Programming* by Bob Neveln. (Prentice Hall, ISBN: 0130879401), 2000

Another fairly old text, hence relatively easy to find.

3. *The Art of Assembly Language* by Randall Hyde (a MASM expert). (No starch press, ISBN 1886411972), 2003
Our target is the 'published edition' of this book, which is also [freely available on the web in html, pdf and chm format](#)
(This book should not be confounded with "*The Art of Disassembly*", by my friend Zero and alia)
 4. *Professional Assembly Language* by Richard Blum. (Wrox, ISBN: 0764579010), 2005
 5. *Disassembling Code IDA Pro and SoftICE* by Vlad Pirogov (A-List, ISBN: 1931769516), 2005
- the last two examples will show how even relatively recent books can easily be found.

Now, of course we all know that 'printed' books on such matters are often both obsolete and superficial, and we also know that the best knowledge (and material) about assembly is to be found instead in some 'grey areas' of the web, for instance visiting [Iczelion's wondrous site](#) or my friend Woodman's [incredible repositories](#).

This query is just a proof of concept, the specific quarries don't matter that much: you'll be able to adapt the following approaches to OTHER, different, targets of yours... replace for instance "assembly" with "python" and you'll obtain a library of python books instead. Or whatever. The approaches and the techniques we will see together are important, the targets themselves are irrelevant.

As you will see another interesting side effect of a correct web-seeking approach is that often, when searching for something, you will find on the same servers [many other targets related to your topic that you did not even know existed](#). Imagine you are retrieving a book inside a library, so that you can have a look at the books -related to the same topic- that are [physically located on the same shelf, next to your target](#) when you arrive to pick it up... 🍷

Searching for **books** on the web is -most of the time- extremely simple.

Please note that -in general- you can find most targets just inputting a title "tel quel" in any search engine: for instance a simple query "[Reversing: Secrets of Reverse Engineering](#)" will almost immediately give us [Eilam's brick \(complete\)](#), quite useful -if you bother to print it- to fix the height of your screen.

Therefore, let's begin with our **first** target simply inputting [Assembly Language Step-by-Step: Programming with DOS and Linux](#) in google [tel quel](#)... and whoop! We receive SERPs full of clowns that want to "sell" us this target :-)

Now -in theory- I should say 'see how useless is google', and show you how to find this book without using google.

Unfortunately (for the purposes of this talk) adding a simple [&as_filetype=pdf](#) to this 'tel quel' searchstring we can fetch through google the complete pdf version of this target **at once**, much too easy! Go figure!

<http://www.coltech.vnu.edu.vn/ttmt/ebooks/John.Wiley.And.Sons.Assembly.Language.Step-by-Step.Programming.with.DOS.and.Linu>

Well what can we do? This was too easy. Since we landed here, let's have a look at the folder where we found this target. Let's 'peel back the URL-onion', something you should routinely do, and -see- we'll land here:

<http://www.coltech.vnu.edu.vn/ttmt/ebooks/>

Woah! A whole collection. Quod erat demonstrandi. And now we have already found our **second** book: [Linux Assembly Language Programming](#), Prentice Hall. Two bingos with just one arrow.

For our **third** target *The Art of Assembly Language* by Randall Hyde, a simple google **tel quel** query won't apparently deliver any useful results.

Good. Let's try an alternative, non-google, (classical) approach to this kind of quarry using MSN Search and its very useful sliders and looking for one of the many possible **file repository** target, for instance 'rapidshare':

<http://search.msn.com/results.aspx?q=%7Bfrsh%3D94%7D+%7Bmtch%3D69%7D+%7Bpopl%3D33%7D+rapidshare+%22Art+of+As>

Please notice the **{frsh=94} {mtch=69} {popl=33}** in the querystring...

The rapidshare query worked: here we already have *The Art of Assembly Language*, our third target (beta draft: do not distribute). It is worth noting that this result was retrievable through google as well, just massaging the querystring a little: **ebook "The Art Of Assembly Language" "No Starch Press"**.

Another possible approach is to use a phrase from the text itself: a text-snippet. In google: **"Hello, World of Assembly Language"**, or in yahoo: **"Hello, World of Assembly Language"**. This is the preferred method to fetch on-line html (or pdf) editions (also non *.zip, *.rar or *.chm files).

The "complete snippet" trick is most useful for fiction books. You want Harry Potter? Have it: **"There was a definite end-of-the-holidays gloom in the air when Harry awoke next morning. Heavy rain was still splattering against the window as he got dressed in jeans and a sweatshirt; they would change into their school robes on the Hogwarts Express."**

and it does not have to be so long, of course:

"There was a definite end-of-the-holidays gloom" will suffice, albeit including some beastly spammers in the SERPs.

Let's now try our **fourth** example: *Professional Assembly Language* by Richard Blum, searching in google **tel quel** we fetch a lot of commercial noise, but with the simplest trick (adding a geographical limitation) if we search **tel quel in russia**, we fetch this target at once. It may be worth noting in this context that when perusing your search results it is ALWAYS a good idea **to open the cached version** of a target first, and only then consider using the original link. This saves bandwidth (most crap goes away), saves time (interesting servers are usually slow or overloaded), and does not unnecessarily increase the popularity of the target URL... So, in this case, a possible link (among many) would be

http://66.249.93.104/search?q=cache:45kcLLqeyd0J:avaxhome.ru/ebooks/professional_assembly_la

Finally, for our **fifth** title, we just try a "chinese" search: **"Disassembling Code IDA Pro and SoftICE"**, note the **&lr=lang_zh-CN** snippet.

And **bingo!**

A possible alternative for our fifth title is the **adding "htm" or "html"** searching trick. Here an example with yahoo: **disassembling-code-ida-pro-and-softice.html**

Note that we could also search **all our titles together** with a nice "potpourri" approach.

Potpourri searches

At times simply guessing that interesting places **MUST** have all your targets on the same page can be useful... in order to find interesting places **and** your targets :-)

Here a "potpourri" example: **"Disassembling Code : IDA Pro and SoftICE" "Professional Assembly Language"**

Anyway, as you have seen, any and every book is at seekers' disposal.

(**Caveat:** Please download these books only if you are positively sure they have been released on the public domain. Real seekers do not need to waste hddisk space downloading dubious stuff. Besides it could even be constructed as 'illegal' by the beastly patents holders and their political lackeys. **Downloading is not necessary!** Seekers will always find again and again on the fly -and consult on line- whatever they fancy).

Ok, ok. It was too easy. Much too easy. In order to continue let's imagine for a moment that it would NOT have been so excessively easy to find these books through such simple searches. Imagine we could not find them. Imagine we will not find them again on those specific URLs. After all [the web is a quicksand](#) and the specific locations where we found these targets could soon disappear (and in fact probably will, after today's talk :-)

Look Ma: no google!

[top](#)

Sprinkles of cosmic searching power

Let's find again our 'assembly' targets WITHOUT using the simple querystrings above

- We could go for the format.

For obvious reasons, given that our quarry are books, the most probable formats will be [.chm](#), [.pdf](#), [.rar](#) (or [.zip](#) or [.ace](#)).

Of course we could take advantage of this through google as well: `-inurl:htm -inurl:html intitle:"index of" +("/ebooks"|"book") +(chm|pdf|zip) +"assembly"`.

Here a similar query for yahoo: `+("/ebooks"|"book") +"assembly" intitle:"index of"` (note the [&vf=pdf](#) inside the querystring).

Let's examine these two querystrings step by step... 🍷

- We could go for the name.

A banal yahoo/rapidshare search: **"Professional Assembly Language" rapidshare** will give us not only some locations, but also some (much more valuable) **names**:

- [Wrox.Professional.Assembly.Language.Jan.2005.eBook-DDU.zip.html](#) (DDU for "Day Day Up")

(<http://rapidshare.de/files/993649/Wrox.Professional.Assembly.Language.Jan.2005.eBook-DDU.zip.html>)

- [WPALJ2005-DDU.pdf.html](#)

(<http://rapidshare.de/files/7456769/WPALJ2005-DDU.pdf.html>)

- [W.P.A.L.rar](#)

(<http://rapidshare.de/files/3119399/W.P.A.L.rar.html>)

- [WPAL.rar](#)

(<http://rapidshare.de/files/2595619/WPAL.rar.html>)

Now, even if these specific repositories' URLs disappear, and they will, you already have the golden **names**, and hence the keys to find these files again and again whenever you want. Note also that some **guessing** can be useful when searching a target on the web.

WPAL is clearly the acronym of Wrox.Professional.Assembly.Language, but the author name, Richard Blum, could also be used, hence [RBlum.rar](#), [RBlum.zip](#), [RBlum.pdf](#) are all possible valid options.

And indeed, again and for the last time with google: ([RBlum.rar](#) OR [RBlum.zip](#) OR [RBlum.pdf](#))

- We could do it like lamers.

The lamers' way: using **P2P** and torrents

Using **P2P specific search engines** is a somehow banal, but often useful way to find your

targets.

Let's *exempli gratia* give [filedonkey](#) a chance: let's just input [assembly](#) and see what we fish. See?

Another "lame but valid" possibility is of course to check what [torrents](#) can deliver when you search -say- [assembly](#).

Anyhow real seekers [deprecate](#) P2P and Torrents "searches". It is always [much quicker](#) and more elegant to fetch your targets from URL locations nobody is visiting, than to slowly download them from overwhelmed servers together with a zillion other zombies doing the same thing.

- We could search elsewhere

1. [FTP](#)

Using [ftp specific search engines](#) is in many cases a good way to quickly find your targets. Let's give [our lithuanian friends](#) a chance: let's just input [professional assembly](#).

Here we have two results: [217.16.23.33 root/books/edocs/Wrox](#) (in rar format) and [85.30.196.165](#)

[root/pub/Info/Books/_ebuki.powernews.ru/_downloads_/downloads.ebuki.apvs.ru/Wrox](#) (in pdf format).

We could also try Dalian's ftp: [assembly](#)

2. Going [local](#) ([Homepages](#), [Usenet](#), [topic-related messageboards](#) & [Webrings](#))

At times it can be useful to "throw a narrow net" in the searchscape, going "specific". Searching [webrings](#) you'll discover that there's a [x86 Assembly Language Webring](#) and the [Win32Asm ring](#).

On [usenet](#) you have [comp.lang.asm.x86](#), [alt.lang.asm](#) and [alt.os.assembly](#) (this last -alas- spammed to death).

You could also [Search private homepages](#) or [AOL](#) for "assembly language".

A further very good idea is to individuate relevant and authoritative topic-related messageboards. There you may even find topic-related specific search engines, like for instance [community.reverse-engineering.net](#) and [woodman's](#).

It is worth underlining that on the web much knowledge can be found in "grey areas" that are completely outside the "academic circuits". In fact if you are interested in assembly you'll soon realise that the four books we have chosen as targets today are obsolete and quite irrelevant if compared with the knowledge you can gather visiting the messageboards pointed out above.

3. Going [regional](#)

Going "regional" is ALWAYS a very good idea when searching. We have already seen how adding a simple [.ru](#) to our queries can help. But WHERE should we search? Which are the, how should I say? the "less copyright-obsessed" countries? Here on the right you can see a interesting "piracy subdivision" published a week ago by [The Economist](#)...

Ok, ok, I know: obviously this kind of bogus "home made" researches, promoted by US-lobbyist Robert Holleyman's "Business Software alliance", are clearly just intended to scare the pants off some corporate clown in order to scrap some money, yet since we are reversers, we may as well [reverse such data for our own purposes](#)... :-)

And look! As you can see, Vietnam, Zimbabwe, Indonesia, China, Pakistan, Kazakistan, Ukraine, Cameroon, Russia, Bolivia, Paraguay and Algeria seem indeed to have a more relaxed attitude towards the "patents mafiosi". Good to know :-)

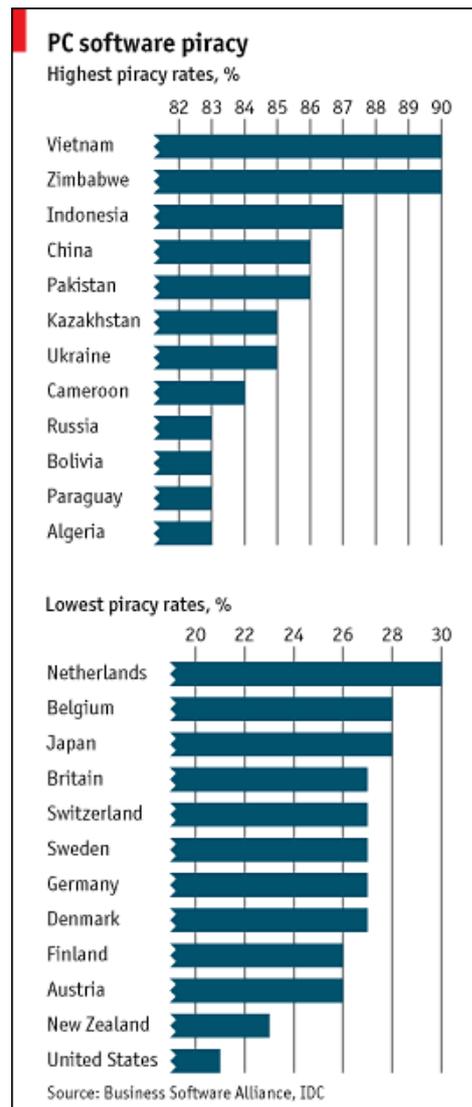
Here the relevant [country codes](#): [.vn](#), [.zw](#), [.id](#), [.cn](#), [.pk](#), [.kz](#), [.ua](#), [.cm](#), [.ru](#), [.bo](#), [.py](#) and [.dz](#), that we could use to [restrict searches](#) to such relaxed places.

Of course some of these countries are just "local niches" with next to inexistent activity and extremely weak signals, and can be ignored.

Let's say that -in general- [.vn](#)(Vietnam), [.id](#) (Indonesia), [.cn](#) (China), [.pk](#) (Pakistan), [.ua](#) (Ukraine) and [.ru](#) (Russia) look promising enough. We may add to these countries -out of experience- Iran, Korea, Bulgaria and India ([.ir](#), [.kr](#), [.bg](#) and [.in](#)).

So let's go local: let's visit [China](#), where we can find [following this link](#), some other interesting assembly books.

Of course we should also have a look in [Vietnam](#), in [Russia/Ukraine](#) (where



4. IRC channels and blogs

Searching through IRC channels and blogs can be -for some targets- quite useful. However the ratio noise/signal is quite bad on these channels, and therefore IRC-searching and blog-searching is -mostly- a waste of time if compared with more effective searching techniques.

After all, and behind the hype, blogs are just messageboards where only the Author can start a thread.

I'll just point you to some [blogs search engines](#) and to some [IRC search engines](#) like [this one](#). Nuff said.

5. Various "uncommon" search engines

At times simply switching to less known (but interesting) search engines can cut mustard.

Here a assembly-related search with [kartoo](#) and here another search using [gigablast](#).

Sliders, SEO beasts, and Sinking google

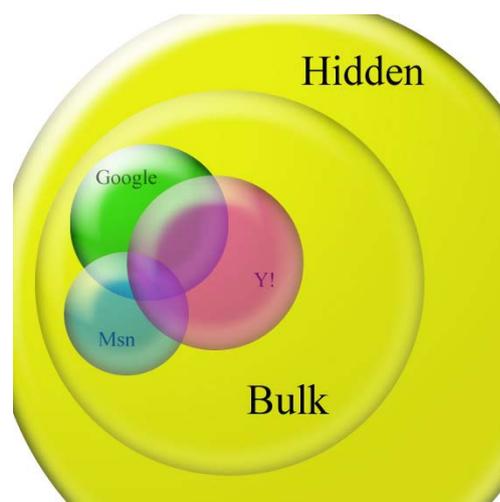
[top](#)

Google alone and you'r never done

Many clueless zombies consider "searching the web" tantamount to digit one term inside google and then clicking enter.

In fact such a simplistic approach is as wrong as it may get. And not only for the "one-termness" of it. The real problem is that google [covers only a tiny part](#) of the web.

The power of its servers and the beautiful simplicity of its interface notwithstanding, [google](#) is only one of the many main search engines, and its database, while currently already past the 20 billions sites mark, covers at most one third of the visible web (and less than 1/50th of the "invisible" one). In fact Yahoo's database is bigger (albeit saddled with a lot of useless commercial crap). The web is just too big for a single search index alone, and still growing quickly. Moreover the "invisible" web content "the real bulk of the web" is hidden behind firewalls or commercial services (that will try to restrict access asking for "subscriptions" or



"money" or a "valid id").

In order to access (part of) it you will need to use techniques that go from **stalking** to **social engineering**, through **trolling** and **passwords breaking**.

Note also that **links** (the food used by all search engines, and especially by google, in their algos) does not convey any real **meaning**.

In fact a link, *per se*, does not mean nothing, you can just count the number of links, as most search engines in fact do, and then try to decide what those links really mean using a bunch of "best-guess" algos. A rather crude approach if you ask me: this gives all advantages to the beastly spammers, none to the users.

Of course some correct developments are already under way: users -rather than spammers- should be able to influence the ranking of search results and some search engines (MSNsearch's **sliders** and Yahoo's **philtron**) already provide users with such a possibility of influencing -at least rudimentary- their own ranking algorithms.

A classical case is the infamous 'popularity' ranking criterion. This you should by all means slide to insignificance, since it is *eo ipso* tantamount to crappiness.

Contrary to what search engines' algos designer still seem to believe, we immediately **TURN POPULARITY DOWN** if given half a chance, since what -say- some idiots in Idaho are massively looking for has no relevance whatsoever for humans with brains. "Sites popular for zombies" are exactly those you never need. Remember the beautiful old trick of adding a **-.com** specification to all your searchstrings per default: all the com sites will disappear: good riddance.

For search engines that do not allow any algo fine-tuning, a possible approach is the **"yo-yo"** approach: jumping from the start onto lower SERPs.

The old "best-guess" and link related algos are what makes life so easy for the beastly spammers: google's results, after many years of legendary quality, are nowadays being spammed more and more by the beastly commercial clowns that call themselves "Search engines optimizers".

Using a plethora of methods (cloaking, doorway pages, hidden text, blog-farms, you name it) these criminals deny everybody the possibility to gather real knowledge in order to push up their crap commercial sites into the first positions of the SERPs.

In fact, searchers are directly affected by these criminal deeds. As even SEO-spammers managed to admit: "searchers have something to gain if they obtain the search results that best match their queries and, consequently, something to lose if they cannot do this".

SEOs, these "Judas of the web" sell -for money- their knowledge and insights of search algos' weaknesses in order to purposely deliver dubious and crap results to our queries... *Quelle vulgarité!*

Acquiring a working knowledge of the many alternative searching paths is *eo ipso* useful and may already now allow even beginners to find valuable results more quickly and reliably.

Such alternatives can soon prove even more crucial for Internet searching purposes: while google may not be yet a sinking boat, anyone can see how much water is already entering through its many holes.

So we have **to reverse our own searching habits**.

Instead of just using google à la "va banque" every time we begin a search, we should carefully

consider [how and where](#) we start our searches, delve a little more inside our searches own specific requirements, else we will waste too much time on irrelevant side paths... 🍷

And since we are all reversers, an ancient and savvy race, incredibly "apt to adapt", we'll be able to reverse first and foremost ourselves and our own working habits.

Google has its weaknesses and its strength, let's analyze them.

Google weaknesses (and strengths)

[top](#)

Le plus fort est celui qui n'oublie pas sa faiblesse

Google is/was/became the best search engine because of its clean interface in a frantically commercial on screen world, and BECAUSE it didn't pollute its results with advertisements and BECAUSE it didn't practice any censorship whatsoever on the results.

It is -alas- now losing ground on both last terms.

Try any search for mp3s, for instance, and you'll see at once both advertisements and censorship at work... 🍷

The third 'raison d'être', its effective and simple interface, still survives somehow, its more and more frequent "cartoonish" cracks on the querymask logo notwithstanding. More annoying is the fact that today up to half of each SERP screen is dedicated to paid ads, compared to the ad-free original "Old-Google".

Google's (relative) cleanliness was so powerfully convincing that many rivals went "back" to a similar clean approach, ditching their useless heavy-commercial portals (compare on [alexa](#) the evolution of [Yahoo's](#) portal... 🍷)

The biggest weakness of google, is that it's 'patented ranking algos' are now pretty well known. Their 'secret combination' of 'thousand of algos' was all just hype from the very beginning, and their ranking approach -never really hidden- is now well known by countless commercial spammers, thus making it a liability rather than an asset.

In the main search engines panorama there are at the moment hundreds of different prototypes and companies that all utilize more or less the same algos. Yet even slight variations can make the difference: their results overlap only for a small part (around 1/4 of the SERPs do overlap). This is where the depth and freshness of the supporting database plays a bigger role than the cleverness of the ranking algos.

In fact (as per spring 2006) only four contenders: Google, Yahoo, Microsoft and Teoma/AskJeeves have enough muscles to guarantee a relatively useful and regular indexing of the web.

But -as we have seen- these four cover together, at best, just one half of the visible web and a tiny part of the invisible one.

This makes it extremely important to use alternative approaches when searching.

Since google is still a useful and powerful quick search engine, and since it owns the whole archive of newsgroup postings, we will never be able to ditch google completely anyway... 🍷

There's also a google bias towards "established sites", due to its links algos: if you are searching content that is likely to have been on the Internet for a LONG TIME, google is a good choice. On the other hand, if you are looking for "fresh" content, you better use MSNsearch (or even good

ole altavista).

Google's real strength is its "quality database" of useful sites. It is not a matter of the quantity of sites listed, it is a matter of quality. Yahoo's database, while bigger than google's, hosts an abominable amount of ".com" sites (five times the amount of google) which heavily skew all results towards irrelevance.

But how do you judge results? How do you prepare your search? Knowledge of some basic searching rules can help.

The golden rules of searching

[top](#)

Quaerite et invenietis

There are some basic rules for seekers. Of course things are different depending from the KIND of search you are performing. There are rules for [long term web searching](#) and rules for [short term web searching](#) ...

But almost every query can be subdivided into the following steps: [think](#), [find](#), [refine](#), [evaluate](#), [collate](#).

(The [Finder](#) [Reverses](#) [Every](#) [Corner](#))

1. **THINK about your query**

Seekers do not "plunge" into a search out of the blue. Like artists, they visualize the correct result before they begin. The 'perfect' answer is driving their queries. The perfect answer creates the correct question(s)

What kind of results do you want? Books? Doctoral thesis? Images? News? Biographies? How many results do you want? Three hundred pages of material? One single authoritative book? A dozen pdf-articles? A short and concise essay?

Obviously you cannot be an expert in all single field of any and every query you will launch. But you must be an expert in the field of finding the right resources for each and every kind of query.

A seeker needs TWO skills: [to formulate a question correctly](#) and [to know where to look](#). And this means knowing which resources you should use for your searches. And this means you must first of all know how to search those very resources you should use for your searches.

In fact each 'part' of the web requires a different approach. For instance, searches on [usenet](#), on [blogs](#) or on [ftp servers](#) are not ruled by the same [lore](#). Also each [kind of target](#), each quarry, requires a different approach: for instance when searching [news](#), [images](#) or [books](#). You must also decide if for a given query you will have to use [combing](#) techniques like [stalking](#), [luring](#) or [trolling](#).

Before even beginning, think about your query: [prepare your question\(s\)](#) for the perfect result and [decide which resources you will use](#).

2. **FIND what you are looking for**

Easier said than done, I know, I know. In fact this very complex step is at the same time the whole point of the exercise, duh. However, depending on the previous "thinking about your query" step, you will at least already know where you should be looking for and what kind

of techniques you'll have to use.

A general advice is to **comb** as much as you can, i.e. use the knowledge that others have gathered, search those that have already searched, and do not 'reinvent the wheel' at every query.

A second generally useful advice is to go '**regional**' as much as you can, that is to use information and resources that are located on the same plane (geographically, temporally, academically, conceptually) as your quarry.

Anyway, if your **question has been formulated correctly** and if you already **know where to look**, the 'finding' part will not be too hard.

7. **REFINE while searching**

Your queries are usually either too wide or too narrow. Usually -in fact- they are too wide. If a subject is too wide, as it is most of the time, you have to limit and narrow your search. Using boolean operators (**AND** and **NOT** or **+** and **-**) will narrow the search adding and/or eliminating terms. You can also limit your query temporally (for instance only **2005/2006**), geographically (for instance only **.ru**) or formally (for instance only **.pdf** files).

These limits allow you to restrict results to items meeting specific criteria. I.e.: a particular type (newspaper articles, journal articles, complete books, small snippets of text); a particular language (English, German, Spanish, Russian, Italian, French, etc.); a target published or produced within a particular time frame (2000-2004)

8. **EVALUATE your results**

This is easier said than done, again. The **evaluation phase** is of paramount importance, but -alas- far from being simple.

Whatever you are looking for, you are bound to find very good quality results, good quality results, average quality results and poor quality (or no quality at all) results. This is not only due to the spam, but to the simple fact that the web allows anyone to publish anything he fancies.

A possible approach to evaluation is to use as a rough evaluation guide the seven old classical questions: **quis**, **quid**, **ubi**, **quibus auxiliis**, **cur**, **quomodo**, **quando**: Who, What, Where, Helped by whom, Why, How and When.

Well, first of all you should maybe ask yourself **why the heck** you need to search for something at all :-)

Continue only if you have an answer to this fundamental question.

If you manage to answer the fundamental question and continue, whenever you find a result, it is useful to ask yourself for evaluation purposes the whole bag of classic questions. Does not take long and helps a lot.

Let's begin: **quis** WHO is the Author (and therefore, given his biography, what qualifies him to write about the matter at hand); **quid** WHAT is in fact the result you found (a complete explanation, a proof of concept, a small addition, an hypothetical solution...); check **ubi** WHERE you did find your result (look at the URL, look at the server, look at the links pointing to it...); **quibus auxiliis**, WHO helped the Author? (look at who OWNS the server hosting the result, look at eventual references, links, etc...); and ask yourself **cur**: WHY the result has been produced and put on the web; **quomodo** HOW the result has been produced (again, similar to quid/what: years of research or one half-afternoon sudden jerk?); and finally **quando**: WHEN was the result produced/published/updated, when was the web site created/updated. (**Archive** org may prove **invaluable for such dating purposes**. Note that you can also retrieve a site for **specific date** of the past).

Simply answering the seven classical questions will already allow you to proceed towards a proper evaluation of a set of results.

Finally a word about those "ready-made" evaluations you can find on the web. Should you use them? Yep, *Cum grano salis*.

First of all there's a tendency to ignore "grey areas" of the web when evaluating targets. Some seem to believe that a pdf file should automagically be *eo ipso* more worthy than an html file, independently from its actual content. No way. A text is not worth anything just because it has been printed and published in a book. Its worthiness is always and only **intrinsic**. Many ready-made evaluations on the web are blinded by frills' bells and excessive 'formal-bowing' and utterly incapable of judging content at face value.

It may also be worth noting that -in general- east european places (.ru, .bg, .cz etc.) are (still) "culturally" less commercially oriented and therefore offer more "sound" valid evaluations of books/software/targets, instead of the bogus fanbois "evaluations" that are purposely planted on -say- amazon or ebay. In fact you can hardly find a non-paid -sorry- non-biased review or comparison of certain products on the west-side of the web.

11. COLLATE your results

Ok, you have performed a long search. Gathered tons of results. Painstakingly weeded out bogus and crap sites, understood which are the most important authoritative results... and now you stop your search and go to sleep satisfied.

This is a serious mistake. A query is not finished when you have found your results. Most will be lost if you don't COLLATE your results, squeezing the most authoritative results into a coherent and valid interpretation. A 'conclusion' of sort.

Systematic record keeping is OF PARAMOUNT IMPORTANCE when searching. A classical mistake is to 'forget' to keep records during complex searches.

For this purpose I suggest you simply use the NOTE function in **Opera**: just highlight the target text you are interested in, rightclick, and then choose **copy to note** (or use the keyboard shortcuts, either CTRL+SHIFT+C or CTRL+ALT+E depending on the version of Opera you'r using): *the URL* of the page you'r viewing at that moment *and the date* will be automatically stored in your note *together with the highlighted text*.

You may want to create ad hoc **note folders** (for instance "**research_assembly_books_29MAY2006**") and, at the end of your search, before switching the computer off and go to sleep, just move all your related notes inside some correctly named folders.

Opera's Notes are just text format, very easy to edit, cat, search or prune.

Alternatively use **something** to take notes, even a pen and a sheet of paper will do. DO NOT rely on your memory alone (or on your extraordinary seeking capabilities to re-find at once what you may have lost :-)

If you do, you will regret it. Sooner than you believe.

Once you create some crumbs-paths of well kept records, collating the results will be a quick and easy process.

12. A final note about your "searching environment"

Listening to music while searching is NOT a good idea, chatting while searching is NOT a good idea, being interrupted while refining a query is verry, verry bad: Always search & seek **in a quiet and relaxed environment**, with as few disturbances as possible. No music, no telephone, no skype, no email distractions, no IRC, no chat (and of course no TV, duh).

A serene, calm atmosphere, will allow you to take full advantage of your seeking efforts in

an optimal way.

Serenity CREATES serendipity.

This does not have to mean soberness, austerity or ascetics, though. If you fancy something to drink, have it ready before starting, and always chose excellent products: wines like [pomerol](#) or [refosco](#) or the most finest teas ([Darjeeling second flush](#), for instance). Seekers can often fetch such seemingly expensive items for next to nothing using the old usual [barcode tricks](#).

Let's cut it short: Consider yourself [a monk of the early middle ages](#), sitting in his peaceful cell, seeking old forgotten knowledge, sipping good wines, while barbarians and zombies are burning everything in sight and torturing each other not far from the abbey's walls... consider yourself a monk of the early middle ages among barbarians because this is exactly what you are and that is exactly what is happening nowadays :-)

Conclusions

[top](#)

Quaeras ut possis, quando non quis ut velis

The various techniques described above can and should be used together with the [main](#) search engines: on the ever moving web-quicksands it does not make much sense to give a "list of links" to places where you can "alternatively search".

Of course there are various important non commercial databases, like Infomine (<http://infomine.ucr.edu>), Librarians Internet Index (<http://lii.org>), The Internet Public Library (<http://www.ipl.org/>) Resource Discovery Network (<http://rdn.ac.uk>), Academic Info (<http://www.academicinfo.net/>), The Front (for journals: <http://www.arxiv.org/multi?group=math&%2Ffind=Search>) and finally the best one of all: The Open Directory Project (<http://dmoz.org>).

These are of course all [possible](#) alternatives to the main search engines approaches.

Yet lists of links are and remain just lists of links. Bound to decay into obsolescence. It is much better to describe the different APPROACHES, that will remain valid for many many years even on our extremely 'quicksandish' web-environment.

We have seen some of these approaches during our search for assembly books, let's quickly summarize them again:

1. [All the various main](#) search engines (a "treasure chest" for seekers)
2. [Regional searching](#): the paths beneath the horizon
3. [Combing the web](#): dos and do-nots (usenet, messageboards, blogs, irc, ftp...)
4. [Unorthodox searching](#): guessing, stalking, social engineering, luring, trolling, klebing & more

[aap](#)

Regional searching: paths

[top](#)

<h2>beneath an anglo-centric horizon</h2>

our vietnamese, chinese, persian and russian friends... 🍷

aap	<h2>Combing the web: dos and do nots</h2>	top
-----	---	-----

"potpourri" searching (not only for books, but also for **music**, or many **other things**. And note that with both searches we are "yo-yoing" to page 10 or 8 of the SERPs to avoid spammers nested in the first pages :-)

snippet searching,

"html adding", name-guessing and all the other tricks.

Messageboards, usenet, ftp, blogs, rapidshare alike repositories, torrents... 🍷

aap	<h2>Unorthodox searching: guessing, stalking, social engineering, luring, trolling, klebing & more</h2>	top
-----	---	-----

... 🍷

Tools for seekers

Using **wget** for fun and pleasure

wget, which **exists for windows as well**, is a file retrieval tool that can be used via FTP or via HTTP (the two most widely used Internet protocols). Used mostly in order to mirror websites, it can also be used to find files across the web. Wget supports proxy servers, and most of the features are configurable. An invaluable tool for searchers.

<h2>Assignments</h2>	top
----------------------	-----

Nil perpetuum, pauca diuturna sunt

Two assignments, one easy, one not.

1) For the "lazy occasional searcher" a simple and easy assignment: (just in order to practices the

various techniques explained above): find [Python 2.1 Bible](#), by Dave Brueck & Stephen Tanner, (Wiley, ISBN: 0764548077, 2001). Be careful: around the web there are few "unencrypted" pdf editions and many "encrypted" pdf editions. Obviously you want to copy, use, extract whatever you fancy. So find the "unencrypted" edition of this target. This search should take you at most 10 minutes.

2) For the "serious seeker warrior" a more complex assignment: Find data to stalk and highlight the wrongdoings of [these dangerous clowns...](#)

You can see how naïvely direct they can -and do- propose to sell disinformation services for dictators & dictatorships to be used against democratic media and against citizens at large. As usual there's no shortage of scumbags, ready to sell their informatics skills to governments, military establishments or private buyers for whatever dirty and undemocratic use or purpose. There's instead, alas, a shortage of reversers that would care countering and denouncing this. Using google AND especially using some of the alternative approaches you have seen today, this search/stalking exercise can be accomplished in less than one week.

And now I'm finished.

Thank-you for your patience. Any questions?

FORMS

[Top](#)

SEARCHING THE PAST (DISAPPEARED SITES)

<http://webdev.archive.org/> ~ The 'Wayback' machine at Alexa: explore the Net as it was!

Visit [The 'Wayback' machine](#) at Alexa, or try your luck with the form below.

Alternatively, learn how to navigate through [\[Google's cache\]](#)!

Alternatively a new US-centric "preservation" project [Webcapture](#) is coming along.

aap	All the various main search engines.	top
-----	---	-----

A quick tour of the [main](#) search engines... 

Quick forms			
√	<input type="text"/>	GOOGLE	T Always 100 rez, safe off

√		MSN	T	Use the sliders! e.g. {frsh=86} {popl=13} {mtch=80}
√		YAHOO	T	Da biggest index?
√		A9	T	You can personalize it
√		ASK	T	ex-teoma
√		GIGABLAST	T	Quite good
√		WISENUT	T	You can set your preferences
√		ALTAVISTA	T	With its wondrous NEAR operator!
√		FAST	T	fastsearching for: the exact phrase <input type="text" value=""/> 100 rez
√		SEEKZ		Da über compound
√		EXALEAD		Exalead!
√		BBC.CO.UK		inktomi type
√		DMOZ		the ALPHA & OMEGA of searching

√ = advanced; T = tools; visit the **main** search engines page.
 Learning the advanced specs and operators of a few major search engines (not only **google's**), knowing their common commands and their specific features, will improve your chances to fetch relevant information when **seeking** the web.

[back to portal](#) [back to top](#)

(c) 1952-2032: [[fravia+](#)], all rights reserved, all wrongs reversed