Web-searching Session @ BERLIN, 29/12/2004

(CCC 21C3, Berlin ~ 29 December 2004)

*How to find *anything* on the web*
Advanced internet searching strategies & "wizard seeking" tips

by fravia+

19 December 2004, Version 0.013

This file dwells @ http://www.searchlores.org/berlin2004.htm

Abstract

This document is listing some points to be discussed is my own *in fieri* contribution to the 21C3 ccc's event (December 2004, Berlin). The aim of this workshop is to give European "hackers" cosmic searching power, because they will need it badly when (and if) they will wage battle against the powers that be.

The ccc-friends in Berlin have insisted on a "paper" to be presented before the workshop, which isn't easy, since a lot of the content may depend on the kind of audience I find: you never know, before, how much web-savvy (or clueless) the participants will be.
Hopefully, a European hacker congress will allow some more complex searching techniques to be discussed. Anyway, as usual, the real workshop will differ a lot from this list of points, techniques and aspects of web-searching that need to be explained again and again if we want people to understand that seeking encompasses MUCH MORE than just using the main search engines à la google, fast or inktomi with one-word simple queries.
I have kept this document, on purpose, on a rather schematic plane, but you will at least be able to read THIS file before the workshop itself, and - as you'll see - there are various things to digest even during this short session.
The aim is anyway to point readers towards solutions, and, above all, to enable them to find more material by themselves. If you learn to search the web well, you won't need nobody's workshops anymore :-)
Keep an eye on this URL, especially if you do not manage to come to Berlin... It may even get updated :-)

Introduction

I'll try my best, today, to give you some cosmic power. And I mean this "im ernst".
In fact everything (that can be digitized) is on the web, albeit often buried under tons of commercial crap. And if you are able to find, for free, whatever you're looking for, you have considerable power.
The amount of information you can now gather on the web is truly staggering.
Let's see... how many fairy tales do you think human beings have ever written since the dawn of human culture?
How many songs has our race sung?
How many pictures have humans drawn?
How many books in how many languages have been drafted and published on our planet?

The Web is deep! "While I am counting to five" hundredthousands of new images, books, musics and software programs will be

uploaded on the web (...and millions will be downloaded :-) ONE, TWO, THREE, FOUR, FIVE
The mind shudders, eh?

The knowledge of the human race is at your disposal!
It is there for the take! Every book, picture, film, document, newspaper that has been written, painted, created by the human race is out there somewhere in extenso, with some exceptions that only confirm this rule.
But there are even more important goodies than "media products" out there.
On the web there are SOLUTIONS! WORKING solutions! Imagine you are confronted with some task, imagine you have to solve a software or configuration problem in your laptop, for instance, or you have to defend yourself from some authority's wrongdoing, say you want to stop those noisy planes flying over your town... simply imagine you are seeking a solution, doesn't matter a solution to what, ça c'est égale.

Well, you can bet: the solution to your task or problem is there on the web, somewhere.
Actually you'll probably find MORE THAN ONE solution to your current problem, and maybe you'll be later able to build on what you'll have found, collate the different solutions and even develop another, different, approach, that will afterwards be on the web as well. For ever.
The web was made FOR SHARING knowledge, not for selling nor for hoarding it, and despite the heavy commercialisation of the web, its very STRUCTURE is -still- a structure for sharing. That's the reason seekers can always find whatever they want, wherever it may have been placed or hidden. Incidentally that is also the reason why no database on earth will ever be able to deny us entry :-)

Once you learn how to search the web, how to find quickly what you are looking for and -quite important- how to evaluate the results of your queries, you'll de facto grow as different from the rest of the human beings as cro-magnon and neanderthal were once upon a time.
"The rest of the human beings"... you know... those guys that happily use microsoft explorer as a browser, enjoy useless flash presentations, browse drowning among pop up windows, surf without any proxies whatsoever and hence smear all their personal data around gathering all possible nasty spywares and trojans on the way.

I am confident that many among you will gain, at the end of this lecture, either a good understanding of some effective web-searching techniques or, at least (and that amounts to the same in my eyes), the capacity to FIND quickly on the web all available sound knowledge related to said effective web-searching techniques :-)

If you learn how to search, the global knowledge of the human race is at your disposal

Do not forget it for a minute. Never in the history of our race have humans had, before, such mighty knowledge chances.
You may sit in your Webcafé in Berlin or in your university of Timbuctou, you may work in your home in Lissabon or study in a small school of the Faröer islands... you'll de facto be able to zap FOR FREE the SAME (and HUGE) amount of resources as -say- a student in Oxford or Cambridge... as far as you are able to find and evaluate your targets.
Very recently Google has announced its 'library' project: the libraries involved include those of the universities of Harvard, Oxford, Michigan and Stanford and the New York Public Library.
Harvard alone has some 15 million books, collected over four centuries. Oxford's Bodleian Library has 5 million books, selected over five centuries. The proposed system will form a new "Alexandria", holding what must be close to the sum of all human knowledge.

Today we'll investigate together various different aspects of "the noble art of searching": inter alia how to search for anything, from "frivolous" mp3s, games, pictures or complete newspapers collections, to more serious targets like software, laws, books or hidden documents... we'll also briefly see how to bypass censorship, how to enter closed databases... but in just one hour you'll probably only fathom the existence of the abyss, not its real depth and width.

Should you find our searching techniques interesting be warned: a high mountain of knowledge awaits you, its peak well beyond the clouds.
I'll just show you, now, where the different paths begin.

Scaletta of this Session

Examples of "web-multidepth"

"I've heard legends about information that's supposedly "not online", but have never managed to locate any myself. I've concluded that this is merely a rationalization for inadequate search skills. Poor searchers can't find some piece of information and so they conclude it's 'not online"

The depth and quantity of information available on the web, once you peel off the stale and useless commercial crusts, is truly staggering. Here just some examples, that I could multiply "ad abundantiam", intended to give you "a taste" of the deep depths and currenrts of the web of knowledge...

A database and a search engine for advertisements, completely free, you may enlarge (and copy) any image, watch any ad-spot.
Advertisements from Brasil to Zimbabwe. Very useful for anti-advertisement debunking activities, for avertisement reversing and for the various "casseurs de pub" and anti-advertisement movements that are -Gott sei dank- now aquiring more and more strength, at least in the European Union.

And what about a place like this?
http://www.britishpathe.com/: "Welcome to Version 3.2 of the world's first digital news archive. You can preview items from the entire British "Pathe Film Archive" which covers news, sport, social history and entertainment from 1896 to 1970"...3500 hours of movies! and 12,000,000 (12 MILLIONS) still images for free!

Or what about a place like this?
Anno: Austrian newspapers on line. 1807-1935: COMPLETE copies of many Austrian newspapers from Napoleon to Hitler... for instance Innsbrucker Nachrichten, 1868, 5 Juni, page 1... you can easily imagine how anybody, say in Tanzania, armed with such a site, can prepare university-level assignements about European history of the late XIX century "ziemlich gründlich", if he so wishes.

And the other way round? If you'r -say- in Vienna and want access to -say- Tanzanian resources?
Well, no problem! UDSM virtual library (The University of Dar es Salaam Virtual Library), for instance, and many other resources that you'll be able to find easily. And this is just a tiny example of a world where, I'll repeat it again for the zillionth time, EVERYTHING (that can be digitized) is on the Web.

Let's have a closer look at books, using the Gutenberg and the University of Pensylvania engines.

Project Gutenberg

Project Gutenberg at http://www.gutenberg.org/, or Project Gutenberg at http://promo.net/pg/Home Pages: One of the first full-text Internet collections. We'll see if Google manages to do better with its new Library project. Project Gutenberg should be accessed by its alphabetic or specific search masks for author/title. Note also that there are various "current" Project Gutenberg sites. So link correctly. Many links provided on the web, alas, point to earlier addresses which are no longer being maintained.

Gutenberg's online catalogue: http://www.gutenberg.org/catalog/
Gutenberg's advanced search engine:http://www.gutenberg.org/catalog/world/search

Gutenberg's Database search
Search by Author or Title. For more guidance, see the Advanced Search page, where you can specify language, topic and more.

Author: doyle Title Word(s): EText-No.: Go!

Note that often enough you have links to computer generated audio books in mp3 format as well...

offline catalogues
recent books

University of Pensylvania

(University of Pensylvania's Digital Library)
Au thor: doyle

○ Words in last or first name
○ Exact start of name (last name first)

Title:

○ Words in title
○ Exact start of title ("The", "A", and "An" can be omitted)

Search! Clear the Form **Examples:**

- Entering **austen, jane** in the Author field finds books by Jane Austen.
- Entering **Baum** in the Author field and and **oz** in the Title field finds L. Frank Baum's Oz books.
- Entering **dosto** in the Author field, choosing the Exact start of name option, and entering **underground** in the Title field finds Fyodor Dostoevsky's *Notes from the Underground*, even if you don't remember how to spell more than the start of the author's name!

http://onlinebooks.li brary.upenn.edu/ Upenn's online books.
http://onl inebooks.library.upenn.edu/search.html Upenn's online books, search mask, the same reproduced above.
For instance: doyle.

These are but examples. Remember that whole national libreries & complete government archives, are going on line *in this very moment* in some god-forgotten country in Africa or Asia... a world of knowledge at your finger tips, as I said... provided you learn how to search...

The Web and the main search engines
Structure of the Web, Playing with Google

Growth, Spam, Seos, noise, signal

The searching landscape has changed abruptly during the last months of 2004. New, powerful search engines have appeared, all trying to snatch google from its (until now still deserved) top position. Yahoo has bought fast/alltheweb... and promptly degraded it, almost destroying what many considered the best search engine of the web (way better than google thank to its boolean operators).

A9 is amazon's amazing search engine, that will allow any simple bot to fetch COMPLETE books, snippet by snippet, using the context search functions.
Another New contendent: MSN new beta "super" search, while still in its infancy, has introduced three sliders that put google to shame... Alas, MSbeta's own algos deliver queryresults that are not as pertinent as google, and its SERPs are -as a consequence- next to useless. But we should never underestimate the enemy, and we should never underestimate Microsoft.

A9 and MSSearch Beta are just two examples. Of course they now compete not only with google, but also with teoma and fast (alltheweb), now powering yahoo.
There are MANY other main search engines though, and some of these deserve attention, for instance inktomi, maybe the most underestimated big search engine in the world, which has one of the richest search syntaxes, with lots of unique features and a ranking algo which works often quite well.

Also, Kartoo is a very interesting "meta-engine" for seekers, because of its useful graphic "semantic connections". Using it you'll often find new angles for your queries.

You would be well advised to note that there are also -now- more and more engines with their own CACHE, a complete copy of the web they have indexed, copied pages that you can access even if the original ones have disappeared, a fact that turns out to be EXTREMELY important in our quicksand web, where sites disappear at an alarming rate. At the moment, apart google, we have A9, MSNbetasearch, Baidu & Gigablast, all of them with their very useful caches.

Of course there is always -also- good ole Webarchive to take care of all those disappeared sites.

So we have many main search engines (and you would be wrong in using only google, because they overlap only in part), and yet you should understand that all main search engines together cover but a small part of the web.

Google, the biggest of them all, covers -allegedly- around 8 billion pages. Altogether, when you count the overlappings, all main search engines cover at most one half of the web (if ever).
Let's have a more detailed look at google's depth using, for instance the "Rimbaudian" wovels approach:

a (like apple)	7,440,000,000
i (like i-tools)	2,750,000,000
e (like e-online)	1,840,000,000
o (like O'Reilly)	923,000,000
u (like whatuseek)	457,000,000

If you want to get the whole 8 billions ka-bazoo, you simply query using the english article the :-)

REDUNDANCE SEARCHING

Let's play a little Epanalepsis, just to show you some "angles":
the the : 72,800,000
Yumm, just doubling the article reduces from 8,000,000,000 to 72,800,000 (more pertinent) sites :-)
the the the : 73,500,000
the the the the : 71,900,000
Usually the redundance trick gets 'stuck' when you try to repeat the searchterm too much. Just repeating a search term twice, however, cuts a lot of noise.

So, to make an example that some of you will enjoy, the moronical one-word search string ddos gives you 899,000 results, while the slightly less moronical query ddos ddos gives you around half that number (474,000) and these results have also less noise.
Can we do better? Yes of course... let's kill all those useless "com" sites: ddos ddos -".com" : 203,000 results, mucho more cleano.

Some of you would probably think: great, then this is the way to go... just double the queryterm and eliminate the ".com" sites, how simple and elegant...
Maybe, for a broad search, but for a serious work on ddos attacks you may also find relevant signal with a specific SCHOLAR search engine (limiting it to the most recent months):
ddos ddos "june | july | august | september | october 2004" This is a MUCH more useful ddos query
However seeking, once more, is NOT (or only in part) made using the main search engines.

In order to understand searching strategies, you have first to understand how the web looks like.
First of all the web is at the same time extremely static AND a quicksand, an oximoron? No, just an apparent contradiction.

See: Only less than one half of the pages available today will be available next year.
Hence, after a year, about 50% of the content on the Web will be new. The Quicksand.
Yet, out of all pages that are still available after one year (one half of the web), half of them (one quarter of the web), have not changed at all during the year. The static aspect

Those are the "STICKY" pages.
Henceforth the creation of new pages is a much more significant source of change on the Web than the changes in the existing pages. Coz relatively FEW pages are changed: Most Webpages are either taken off the web, or replaced with new ones, or added *ex novo*.
Given this low rate of web pages' "survival", historical archiving, as performed by the Internet Archive, is of critical importance for enabling long-term access to historical Web content. In fact a significant fraction of pages accessible today will be QUITE difficult to access next year.
But "difficult to access" means only that: difficult to access. In fact those pages will in the mean time have been copied in MANY private mirroring servers. One of the basic laws of the web is that
EVERYTHING THAT HAS BEEN PUT ON THE WEB ONCE WILL LIVE ON COPYCATTED ELECTRONS FOREVER
How to find it, is another matter :-)

Some simple rules:
1. always use more than one search engine! "*Google alone and you'll never be done!*"
2. Always use lowercase queries! "*Lowercase just in case*"
3. Always use MORE searchterms, not only one "*one-two-three-four, and if possible even more!*" (5 words searching);
This is EXTREMELY important. Note that -lacking better ideas- even a simple REPETITION of the same term -as we have seen- can give you more accurate results:

Playing with google

yo-yo;
images;

scholar;
timeslice;
long phrase arrow: "'who is that?' Frodo asked, when he got a chance to whisper to Mr. Butterbur")

Structure of the web. Explain tie model and diameter 19-21: do not dispair, never

---

Regional searching.
The importance of languages and of on line translation services and tools

One of the main reasons why the main search engines together cover (at best) just something less than 1/2 of the web is a LINGUISTIC one. The main search engines are, in fact, "englishocentric" if I may use this term, and in many cases - which is even worse - are subject to a heavy "americanocentric bias".

The web is truly international, to an extent that even those that did travel a lot tend to underestimate. Some of the pages you'll find may point to problems, ideals and aims so 'alien' from your point of view that -even if you knew the language or if they happen to be in english- you cannot even hope to understand them. On the other hand this multicultural and truly international cooperation may bring some fresh air in a world of cloned Euro-American zombies who drink the same coke with the same bottles, wear the same shirts, the same shoes (and the same pants), and sit ritually in the same McDonalds in order to perform their compulsory, collective and quick "reverse shitting".

But seekers need to understand this Babel if they want to add depth to their queries.
Therefore they need linguistic aids.

There are MANY linguistic aids out there on the web, and many systems that allow you to translate a page, or a snippet of text from say, Spanish, into English or viceversa.

As an example of how powerful such services can be in order to understand, for example, a Japanese site, have a look at the following trick:

Japanese dictionary
Just input "search" into the search mask.
Then copy the japanese characters for search.
And paste them back again in the same search form.
See?
You can use this tool to "guess" the meaning of many a japanese page or -and especially- japanese search engine options, even if you do not know Japanese :-)
You can easily understand how, in this way, you can -with the proper tools- explore the wealth of results that the japanese, chinese, korean, you name them, search engines may (and probably will) give you.

Let's search for "spanish search engines"... see?
Let's now search for "buscadores hispanos"... see?

Combing
Stalking & Klebing

The first -simple- combing approach (remember, COMBING: searching those that have already searched) is to use old glorious USENET!
Usenet

Messageboards
Homepages
Webrings
getting at the target "from behind": netcraft, synecdochical searching, guessing.

More "webs" in the web: the rings: USENET, IRC, P2P

How many webs out there? A lot!
It is always worth THINKING about your target's habitat before starting your long term searching trip. If you are looking for assembly knowledge, for instance, you should know that there are DIFFERENT and MANY groups that deal with that:
1) Virus writers (that of course must know assembly cold)
2) and their corollary: virus-protection software programmers (maybe the same guys, who knows? :-)
3) crackers (that break software protection schemes, often enough changing a single byte of a kilometer long 'high language' protection :-)
4) on-line gamers, those that would sell their aunts to have a character with more lifes or magic swords when playing on their on-line game servers. By the way: on-line gamers are often also -for the same reason- quite good IP-protocol and server-client experts :-)

Similarly, if you were looking for password breaking and database entering (without authorization,la va sans dire), you would also have to consider different communities:
1) seekers (as I'll explain below, we need to be able to go everywhere on the web, otherwise we cannot seek effectively :-)
2) porn-afecionados (that have devised incredible methods to enter their beloved filth-databases)
3) people that need consulting scholarly (often medical) magazines (that, alas, often enough require registration and money and/or a university account to be read... something which is rather annoying :-)

Longterm searching and short term searching
Our Bots and scrolls

"One shot" queries and careful "weeks-long" combing and klebing preparation and social engineering practices.
The "15 minutes" rule. If in 15 minutes you don't hear the signal, your search strategy is wrong. Do not insist and change approach.

Databases, hidden databases, passwords
Politically correct Borland & lists
Nomen est omen & Guessing

---

password searches:
Searching entries 'around the web', no specific target, using 'common' passwords:
For instance: bob:bob

For instance: 12345:54321
james:james ~

Searching entries to a specific site (not necessarily pr0n :-):
For instance: "http://*:*@www" supermodeltits

Fishing info out of the web:
password3.htm
The above is not 'politically correct' is it? But it works. And speaking of "politically correctness", some of you will love the Borland hardcoded password faux pas... Databases are inherently weak little beasts, duh, *quod erat demonstrandum*.

Also some lists?

WEBBITS
powerful arrows fo everyday use

What we call webbits are specific "ready made" queries that will allow you to bypass most of the crap and of the censorship, most of the noise that covers your signal.

RABBITS (out of the hat)

Examples of absolute password stupidity: http://www.smcvt.edu/access/ejournal_passwords.htm

The "index of" approach using MSN new beta search: http://search.msn.com/results.asp?f=any&q=%2B%22Index+of%22+%2BName+%2B%22Last+modified%22+%2B%22Parent+Directory%22%0A&FORM=SMCA&cfg=SMCINK&v=1&ba=0&rgr
Inktomi: http://169.207.238.189/search.cfm?query=%2B%26quot%3BIndex%20of%26quot%3B%20%2BName%20%2B%26quot%3BLast%20modified%26quot%3B%20%2B%26quot%3BParent%20Direc

More webbits for you to try out at the bottom of this paper.

Homepages and Email one-shot providers and 'light' anonymity

It's simply amazing how many possibilities you have to create "one-shot" email addresses with huge repositories where you can upload, share and download QUICKLY whatever you fancy. For these very reasons, on these places you can also, often enough, find some very juicy targets -)

Of course, for "survival" reasons, you should use some tricks for your files. Especially in copyright obsessed countries, or if you personally are not politically correct - or obedient - vis-à-vis your local national copyright dictatorship.
A simple trick is what I like to call the "zar" compression method (zip+ace+rar): You zip (and password protect), then ace (and password protect), then rar (and password protect) your file. Or choose any other sequence of the various packers, for instance first zip (then stegano into a -say- wav file), then ace the result (then stegano into a -say- rm file), then, again, rar the result (then stegano again into -say- a (huge) pdf file)...

you get the hang of it... You decide the sequences.

(You'll of course automate this process using a simple batch file)

Then, once done, you change the name of your resulting file to -say- a tiff format (even if it is no tiff file, who cares? :-) and up it goes! Better if split into many (at least two) parts. Noone/nobot will really try to see/reconstruct the picture, they will think, at worse, it is some kind of corrupted file, especially if you call it in your email subject something like "tiff with x rays of my last shoulder fracture": they won't be so easily able to sniff your real data either, unless they have really *a lot of time* and/or are really after you :-).

Once your friends download the file, they will go all the steps you have chosen in reverse (with a batch file), and that's it.

Phone them the password(s) for some added (light) anonimity.

Here is a short list of 1 (or more) giga providers, and then also an *ad hoc* webbit to find more...

● Yahoo: (USA, very quick, 1 GB free storage space + homepage)
● Yahoo china: (USA/China, very quick, 1 GB free storage space + homepage, you'll have to compile your data using the real yahoo as a muster, coz everything is in chinese here)
● Walla: (Israel, quick, 1 GB free storage space ~ 10 MB mail attachments)
● Rediff: (India, quick, 1 GB free storage space ~ 10 MB mail attachments)
● gmx.de: (Germany, quick, 1-3 GB free storage space)
unitedemailsystem: (Singapore, slow, 3 GB free storage space)
interways: (USA, quick, 1 GB free storage space)
mailbavaria: (USA, part of interways, quick, 1 GB free storage space)
omnilect: (Texas, quick, 2 GB free storage space)
maktoob: (Arabic, slow, 1 GB free storage space)

"Light anonymity" must know
Of course when you sign up for these services you should **NEVER** give them your real data.
Lie to them a-plenty and shamelessly, like there were no tomorrow... coz there isn't one :-)
But in order to "build" a credible lie, you need some real data. And there are plenty of personal data around if you use the right webbit
A very simple method: just take a book from your library... here for instance, Bill Rosenblatt, Learning the Korn Shell, O'Reilly, 103 Morris Street, Suite A, Sebastopol, California 95472. Such data are more than enough to "get signed" anywhere as, for instance, "Rose Billenblatt", 103 Morris Street, Suite B (if there's a "suite A", chances are there'll be a "suite B", duh), Sebastopol, CA 95472, USA. A very credible, solid address. Should you have to answer any further question when signing up for a "free" email address (your occupation, your level of income...) just choose either the FIRST option of the proposed alternatives ("account", "over 10 million bucks per month") or select "other" so that they will add even more crap to their lists :-)
Point to remember: on the web you NEVER give away your true identity.

Do not feel bad while feeding them only lies: the very reason they HAVE such "free" email addresses sites is -of course- to READ everything you write and to have a copy of everything you upload or create. Of course no human being will ever read what you write, but their bots and grepping algos will do it for the owners of the "free" email services (or of the "free" search engines), presenting them nice tables built on your private data as a result.

Imagine you yourself are controlling, say, yahoo, and you notice (through your greps) that 2 thousand (or hundredthousand) bozos are suddendly going right now to advise their friend to sell tomorrow all shares of, say, pepsi cola... Youd ig it? Insider trading is NOTHING in comparison with the insider data you can have sniffing emails or search queries on the main search engines... or why did you think you have "free" search engines in the first place? Coz yahoo and google owners are nice people that want to help you finding stuff on the web? Nope. The real reason is obvious: in order to know what people are searching for, duh. That's the reason you should always strive to give them as few data as you can. It's like the supermarket "fidelity cards": they just want to stuff their databases for next to free in order to know how much you cry/shit/sleep and/or make love. To spend less money and gain more money from their customers, not the other way round for sure, despite the lilliputian "discounts" they advertise for the zombies that use fidelity cards.

A last point: the list of free email accounts above is of course NOT exhaustive. To fetch MANY more you just build a simple webbit ad hoc:
walla rediff unitedemailsystems

Of course, for email providers, as for everything else, there are ad hoc communities and specific messageboards

There are also MANY providers that will give you limited accounts (as many as you want, but they'll die after -say- 30 days, for instance runbox...)
Such accounts are IDEAL for quick transfer of files between friends (Runbox: 1 GB email storage ~ 100 MB file storage ~ 30 MB message size limit, 30 days limit).
Accounts that are nuked after a given number of days are even better, in "twilight" cases :-) vis-à-vis accounts that remain for ever, even when you have forgotten having used them :-)

Yet please note that, in order to offer 1 gigabyte throw-away email addresses for free, you need to be able to offer a SIGNIFICANT server configuration, which is rarer as you may think, and -hence- that most of the "small" 1-giga email repositories are -mostly- just scam sites that do not work, so, if you want to be "sure and pampered" stick with the known big working ones: walla, yahoo (& yahoo china) gmx, and rediff.

Where to learn

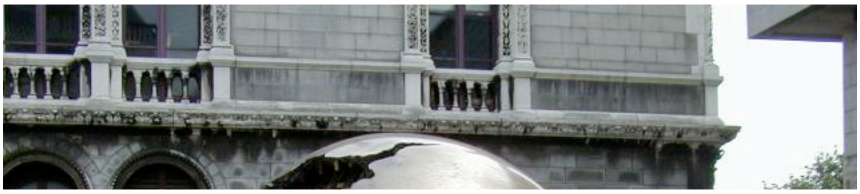
Libraries
scholar

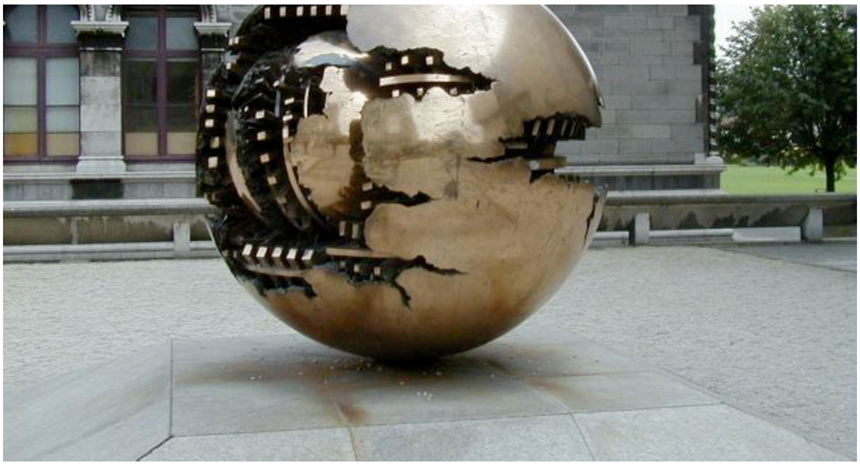
What browser to use, what tools

Opera (the browser is your sword and you fight for time)
Ethereal
Proxomitron
Ipticker

A glimpse of the web?

How does the web look like?
Probably like some "sfera con sfera" sculptures of the artist A. Pomodoro (the "pomodoro" model):





here another example   and another one

The outer sphere would represent the web "at large" with its ~ 23 billions sites.
The inner sphere is the INDEXED web, that you could never reach through the main search engine alones, with its ~ 11 billions indexed sites.
The holes in the structure are all the "disappeared" pages

Another theory is the well-known "tie model", with linked, linkers and its reassuring always tiny "click diameter". Yet another one is the "rings model", with IRC,P2P and USENET as separate rings from the bulk.

Let's find a book

(this is legal for all those among you that are students of the university of Moldova, one of the many countries without copyright enforcement laws, the others should buy their books in the bookshps, la va sans dire)

O'reilly Google hacking
Lord * of the ring (or also msn search)
Historia langobardorum (just to show that this feature is useful for studying purposes, and not only for stealing books :-)

Let's find a song

mp3 wm4 webbits

MP3

So, I imagine you want to know HOW to find mp3 on the web? Say some music by Dylan (I am old, ya know?)
Well the answer is of course NOT to use arrows like mp3 or, say, dylan mp3 music, which will only sink you into the most awful commercial morasses.
Even the old good arrow +"index of" +mP3 +dylan has been recently broken by the commercial pests, and even the -".com" suffix wont help in these cases.

But we have MORE arrows :-)

"index +of" "last Modified" "size" dylan mp3 let's try it :-)

Wanna try another one? "Apache/1.3.29 Server at " mp3 lavigne

See?
Quod erat demonstrandum: The web was made to SHARE, not to hoard :-)

Of course we have more "musical" webbits: here for instance an ogg related one
Ogg as also the advantage of being not PROPRIETARY like mp3...

But if you insist in searching for mp3 another good idea would be to use search engines situated in less "copyright obsessed" countries, like Baidu...

Let's find a program

It's very easy once you have the exact name, for instance TorrentSpy-0.2.4.26-win32.zip. Else you can simply try the serial path.
Examples of the serial path.

See? The point being, as usual, that you should never give money away when searching for books, music, software or scholarly material. Chances are you do not need to give money away for your STUDIES any more very soon: some universities have begun to put all their courses, for free, on the web. An approach still *in fieri*, that will probably be more common in a few years time.
Example: .

Gran Finale

Let's see if we can fish something interesting out of the web for our friends in Moldovia (where there are Moldova">no laws defending copyright, unfortunately). La va sans dire that you are allowed to use programs found in this way only in Moldova...

Rosen's page
Pocket PC: http://www.google.com/search?hl=en&lr=&as_qdr=all&q=booklib+textmaker&btnG=Search
Palm: http://beta.search.msn.com/results.aspx?q=Chromacast++++Converter+++CplxcalPro+&FORM=QBHP

SEARCHING FOR DISAPPEARED SITES

http://web.archive.org/collections/web/advanced.html ~ The 'Wayback' machine, explore the Net as it was!

Visit The 'Wayback' machine at Alexa, or try your luck with the form below.

http:// [Take Me Back!]

Alternatively learn how to navigate through [Google's cache]!

Search the Web of the past

Weird stuff... you can search for pages that no longer exist! VERY useful to find those '404-missing' docs that you may badly need...

□1996 □1997 □1998 □1999 □2000 □2001

[Submit]

Max. Results 50 ▾

NETCRAFT SITE SEARCH

(http://www.netcraft.com/ ~ Explore 15,049,382 web sites)

VERY useful: you find a lot of sites based on their own name and then, as an added commodity, you also discover immediately what are they running on...

Search Tips

site contains ▾ [Netcraft!]

Example: site contains [searchengi] (a thousand sites eh!)

Google timeslice daterange

SLIDES for BERLIN (21C3: 29/12/2004)

The web is still growing

Nobody knows how big the web is, but we may make some good guesses watching the growth (or reduction) of some small specific portions of the web. You'll find in our library more material about the different methods that have been used to measure the width of the web.

The data below are extrapolations I have made since January 2000, using the frequency, visiting patterns and referrals data gathered from my three main sites (www.searchlores.org - Oz, www.searchlore.org - States, www.fravia.com - Europe)
The algo I used is a development of previous extrapolations, made since 1995 on some -now obsolete- old reverse engineering sites of mine, and has proved over many years to be rather correct, given or taken a ~

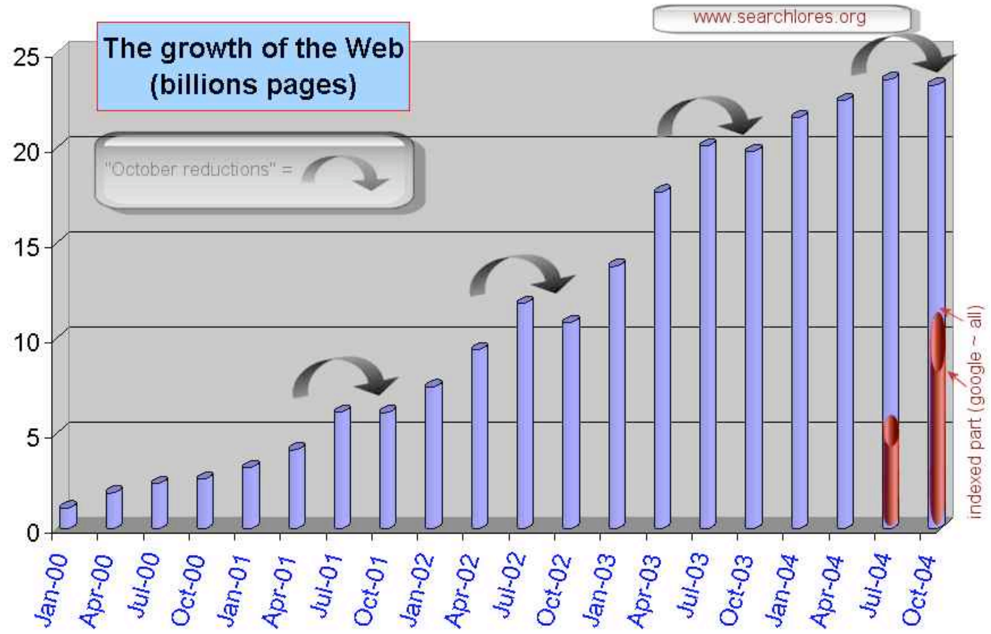
15% margin of error, so I -personally- trust it.

However I am not going to explain nor justify my parameter choices here, suffice to say that the data are not just taken off thin air (in fact you'll easily find out, searching the web, that most scholar authors and many -mostly self-called- experts DO indeed confirm these data).

The growth of the Web in billions pages (January 2000 - October 2004)

Coupla things worth noticing in this slide:
1) The web grows much more slowly since mid-2003, see also the next "pace of growth" slide.
2) Every October of the last 5 years there has been a remarkable REDUCTION of the web width. Why this happens, frankly, still beats me.
3) Google (says they have) expanded its index to 8 billions sites in early october 2004 (as an answer to the arrival on the searchscape of the new MSbetasearch, the new Amazon's A9 and the new, alltheweb powered, yahoo) doubling its indexes from the previous 4 million sites total (one wonders where google kept those extra 4 billions indexed pages before such enlargement, btw :-)
This brings the indexed part of the web to a little less than the half of it: around 11 billions indexed pages against the 23,3 billions existing pages (as per October 2004).

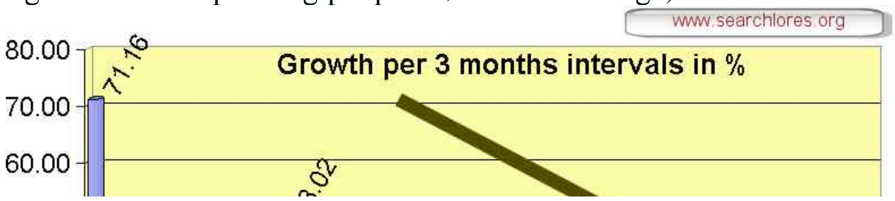
(Image resized for printing purposes, click to enlarge)

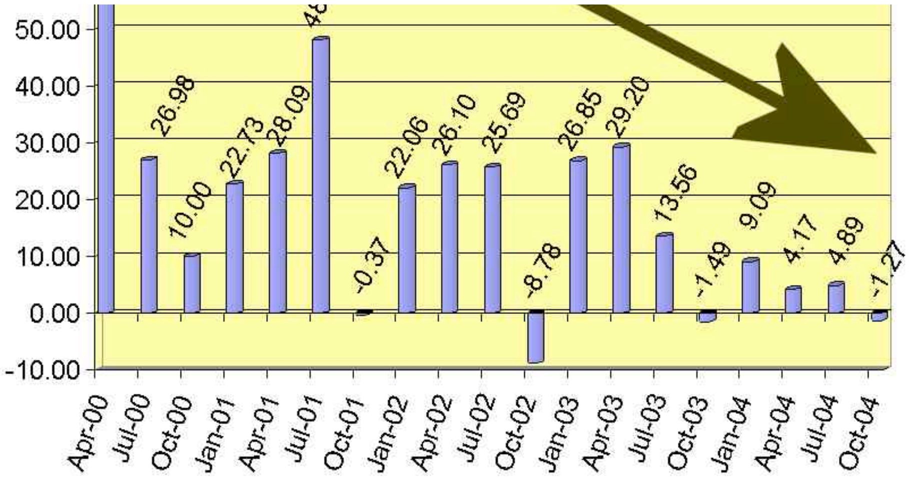


The PACE of growth is slowing down

Coupla things worth noticing in the next slide:
1) The web grows much more slowly since mid-2003.
2) Note the "october reductions" as negative growth. In fact these reductions often begin already in September. Usually there's a new, positive, growth towards November, that usually lasts until the following autumn fogs). Why there should be such a contraction towards the end of every sommer is awaiting explanation... go search, find out, and I'll gladly publish YOUR results if you do :-)
3) Data may be skewed as much as 15% (and probably much more for the first bar on April 2000)

(Image resized for printing purposes, click to enlarge)





Older slides (may come handy)

Structure

Kosher - non kosher

Web coverage, short term, long term

Magic

Sourceror2

```
javascript: z0x=document.createElement('form'); f0z=document.documentElement; z0x.innerHTML = '<textarea rows=10 cols=80>' + f0z.innerHTML + '</textarea><br>'; f0z.insertBefore(z0x, f0z.firstChild); void(0);
```

http://www.google.com/complete/search?hl=en&js=tru%20e&qu=fravia

---

Some reading material (delving further inside the seeking lore)

As I said, a lot of work on database accessing / passwords gathering is still *in fieri*.
In the meantime you may quite enjoy reading the following older essays:

- The Art of Guessing by .sozni
- The weird ways of searching and the weird findings by SvD
- Cat burglers in the museum after dark (How to eneter a Museum database "from behind") by Humphrey P.
- The Zen of porn-images searching by Giglio
- Feed the search engines with synonyms by sonofsamiam
- A small research into SIRS researcher database accesses by Humphrey P.
- A re-ranking trilogy by fravia+
- [eurosearch.htm]: Das grosse européenne bellissimo search
  by fravia+ (Taking advantage of free polylinguistic tools when searching)
  Part of the searching essays and of the Seekers' Linguistic Station sections.
- [inktomi.htm]: Inktomi's search syntax
  by Nemo, part of the essays.htm section.
  Inktomi is one of the best search engines out there. Unfortunately its search syntax is not well documented, which is a pity, because Inktomi offers one of the richest search syntaxes, with lots of unique features and a ranking algo which works often quite well. "Our Tools" Lore
- Quite an addition, by ronin and another addition by Nemo, additions to the [ar1essay.htm]: *The "index +of"*

seekers' trick
"parent directory" +"*.nfo" +"*.rar" +"*.r05"..." power searching

- [rabbits.htm]: *Catching web-rabbits* (Catching the rabbit's ears & pulling files out of the hat) by VVAA, part of the [Essays]. Advanced Web searching tricks ~ Updated!
- A PHP-LAB production:
  [http://fravia.2113.ch/phplab/wopen.htm]: *The Wand of Opening* ~ an accompanying essay by ~S~ loki, and ~S~ Mordred
  Demonstrate the weakness of a large part (not to say the majority) of website's protections built on 'client side' gates ~ Create a script that'll break through most of these protections.
  part of the [PHP-oslse] section. "Hidden database" opening!

more webbits (& klebing) arrows

```
#mysql dump filetype:sql
AIM buddy lists
allinurl:/examples/jsp/snp/snoop.jsp
allinurl:servlet/SnoopServlet
cgiirc.conf
cgiirc.conf
filetype:conf inurl:firewall -intitle:cvs
filetype:eml eml +intext:"Subject" +intext:"From" +intext:"To"
filetype:lic lic intext:key
filetype:mbx mbx intext:Subject
filetype:wab wab
Financial spreadsheets: finance.xls
Financial spreadsheets: finances.xls
Ganglia Cluster Reports
generated by wwwstat
haccess.ctl
haccess.ctl
Host Vulnerability Summary Report
HTTP_FROM=googlebot googlebot.com "Server_Software="
ICQ chat logs, please...
Index of / "chat/logs"
intext:"Tobias Oetiker" "traffic analysis"
intitle:"index of" mysql.conf OR mysql_config
intitle:"statistics of" "advanced web statistics"
intitle:"Usage Statistics for" "Generated by Webalizer"
intitle:"wbem" compaq login
intitle:admin intitle:login
intitle:index.of "Apache" "server at"
intitle:index.of cleanup.log
intitle:index.of dead.letter
intitle:index.of inbox
intitle:index.of inbox dbx
intitle:index.of ws_ftp.ini
inurl:"newsletter/admin/"
inurl:"newsletter/admin/" intitle:"newsletter admin"
```

```
inurl:"smb.conf" intext:"workgroup" filetype:conf conf
inurl:admin filetype:xls
inurl:admin intitle:login
inurl:cgi-bin/printenv
inurl:changepassword.asp
inurl:fcgi-bin/echo
inurl:main.php phpMyAdmin
inurl:main.php Welcome to phpMyAdmin
inurl:perl/printenv
inurl:server-info "Apache Server Information"
inurl:server-status "apache"
inurl:tdbin
inurl:vbstats.php "page generated"
ipsec.conf
ipsec.secrets
ipsec.secrets
Most Submitted Forms and Scripts "this section"
mt-db-pass.cgi files
mystuff.xml - Trillian data files
Network Vulnerability Assessment Report
not for distribution confidential
phpinfo.php
phpMyAdmin "running on" inurl:"main.php"
phpMyAdmin dumps
phpMyAdmin dumps
produced by getstats
Request Details "Control Tree" "Server Variables"
robots.txt
robots.txt "Disallow:" filetype:txt
robots.txt "Disallow:" filetype:txt
Running in Child mode
site:edu admin grades
SQL data dumps
Squid cache server reports
Thank you for your order +receipt
This is a Shareaza Node
This report was generated by WebLog
```

Mmmm... I sense you are using Netscape 5.Mmmm... I sense you are using Netscape 5.