Google Hacking 101

Edited by Matt Payne, CISSP

15 June 2005

http://MattPayne.org/talks/gh

Outline

- Google Bombing
- Schneier in **Secrets and Lies**
  - Attack at a distance
  - Emergent behavior
  - Automation
- Google as a mirror
- “Interesting Searches”
  - Software versions
  - Passwords, credit card numbers, ISOs
- CGI Scanning
  - Vulnerable software
- Defense against Google Hacking

Google Bombing
!=
Google Hacking

- http://en.wikipedia.org/wiki/Google_bomb
- A **Google bomb** or **Google wash** is an attempt to influence the ranking of a given site in results returned by the Google search engine. Due to the way that Google's Page Rank algorithm works, a website will be ranked higher if the sites that link to that page all use consistent anchor text.

So What Determines Page Relevance and Rating?

- Exact Phrase: are your keywords found as an exact phrase in any pages?
- Adjacency: how close are your keywords to each other?
- Weighting: how many times do the keywords appear in the page?
- PageRank/Links: How many links point to the page? How many links are actually in the page?

Equation: (Exact Phrase Hit)+(AdjacencyFactor)+(Weight) * (PageRank/Links)

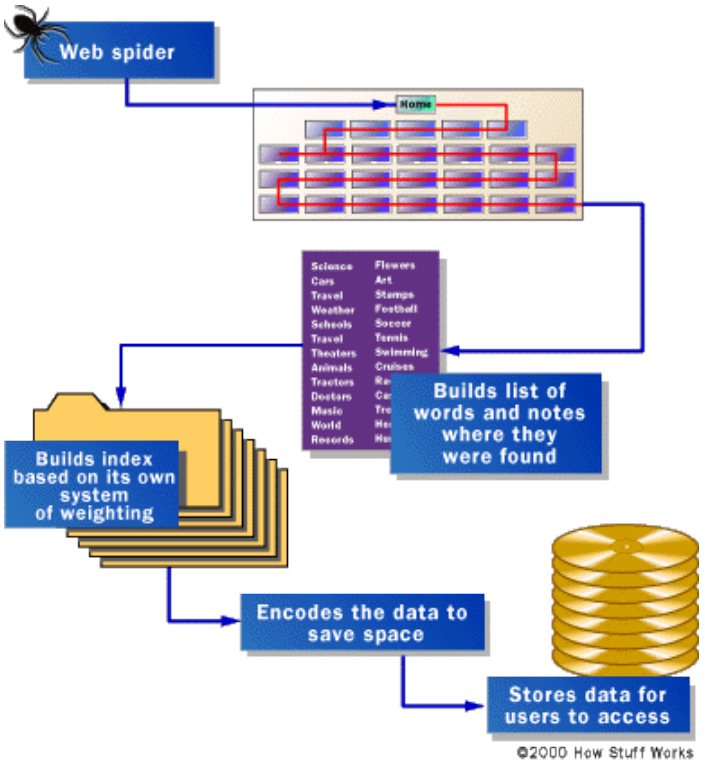
From: Google 201, Advanced Googology - Patrick Crispen, CSU

Simply Put

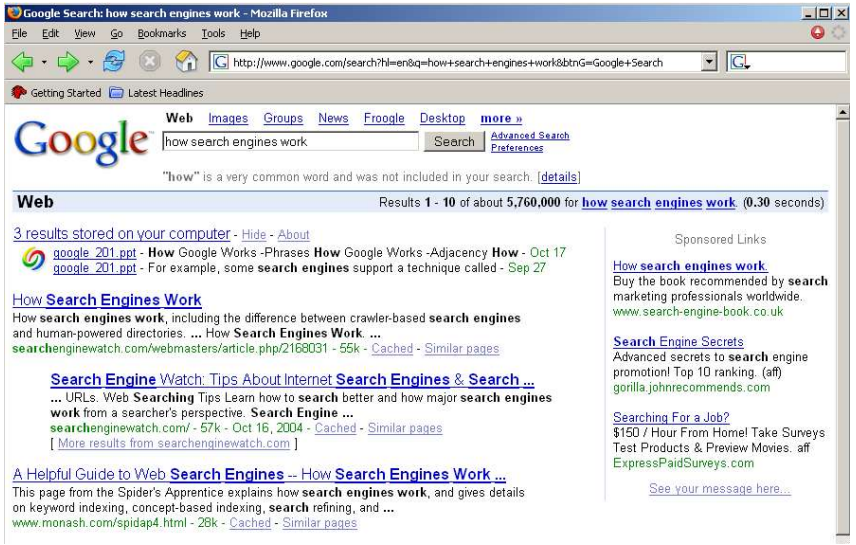
- “Google allows for a great deal of target reconnaissance that results in little or no exposure for the attacker.” – Johnny Long
- Using Google as a “mirror” searches find:
  - Google searches for Credit Card and SS #s
  - Google searches for passwords
  - CGI (active content) scanning

Anatomy of a Search

Server Side



Client Side



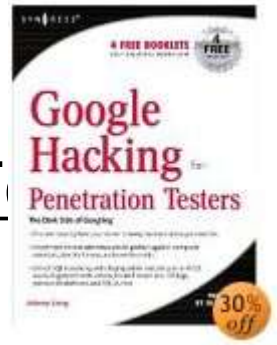
http://computer.howstuffworks.com/search-engine1.htm

How Google Finds Pages

- Are only connected web pages indexed?
- NO!
  - Opera submits every URL viewed to Google for later indexing….

Johnny.ihackstuff.com

- Johnny Long
  - Wrote Google Hacking for Penetration T
    ISBN 1931836361
  - Many free online articles.
    - Two PDFs cached at MattPayne.org/talks/gh
    - See the references slide
    - Or just use google



Google and Zero Day Attacks

- Slashdot Headline: **Net Worm Uses Google to Spread**:
  - **Posted by michael on Tue Dec 21, '04 06:15 PM from the web-service-takes-on-new-meaning dept.** troop23 writes *"A web worm that identifies potential victims by searching Google is spreading among online bulletin boards using a vulnerable version of the program phpBB, security professionals said on Tuesday. Almost 40,000 sites may have already been infected. In an odd twist if you use Microsoft's Search engine to scan for the phrase 'NeverEverNoSanity'-- part of the defacement text that the Santy worm uses to replace files on infected Web sites--returns nearly 39,000 hits."* Reader pmf sent in a few more information links: F-Secure weblog and Bugtraq posting. **Update: 12/22 03:34 GMT** by **T**: ZephyrXero links to this news.com article that says Google is now squashing requests generated by the worm.

Local Example

- Monday 14 February, 2005
  @10:11am
  Update: Now it sounds like everyone was hit with an exploit on awstats which took out quite a few bloggers and other sites. ==> Actually, phorum got hit with it too!

  After running my server something.net for quite awhile on 'borrowed time', it eventually got hacked into - just this weekend. The "Simiens Crew" took credit to a webpage defacement, and by doing some googling... they've hit quite a few websites even just this last weekend! My best guess so far was an attack on one of my many 3rd-party PHP-run services that I have not taken the time to watch and patch for security announcements. Could have been gallery, phorum, webcalendar, icalendar, etc... I'll do some investigating and hopefully find out. I may have been lucky though, it sounds like these were just defacements and not all-out attacks, other victims have not reported any data loss at least. I can respect that. What I *can't* respect though is the many defacements they've put up with "FrontPage" as the HTML generator!

Enough BS, How Do I Get Results?

- Pick your keywords carefully & be specific
- Do NOT exceed 10 keywords
- Use Boolean modifiers
- Use advanced operators
- Google ignores some words*:

a, about, an, and, are, as, at, be, by, from, how, i, in, is, it, of, on, or, that, the, this, to, we, what, when, where, which, with

*From: Google 201, Advanced Googology - Patrick Crispen, CSU

Wall Street, NYC

Google's Boolean Modifiers

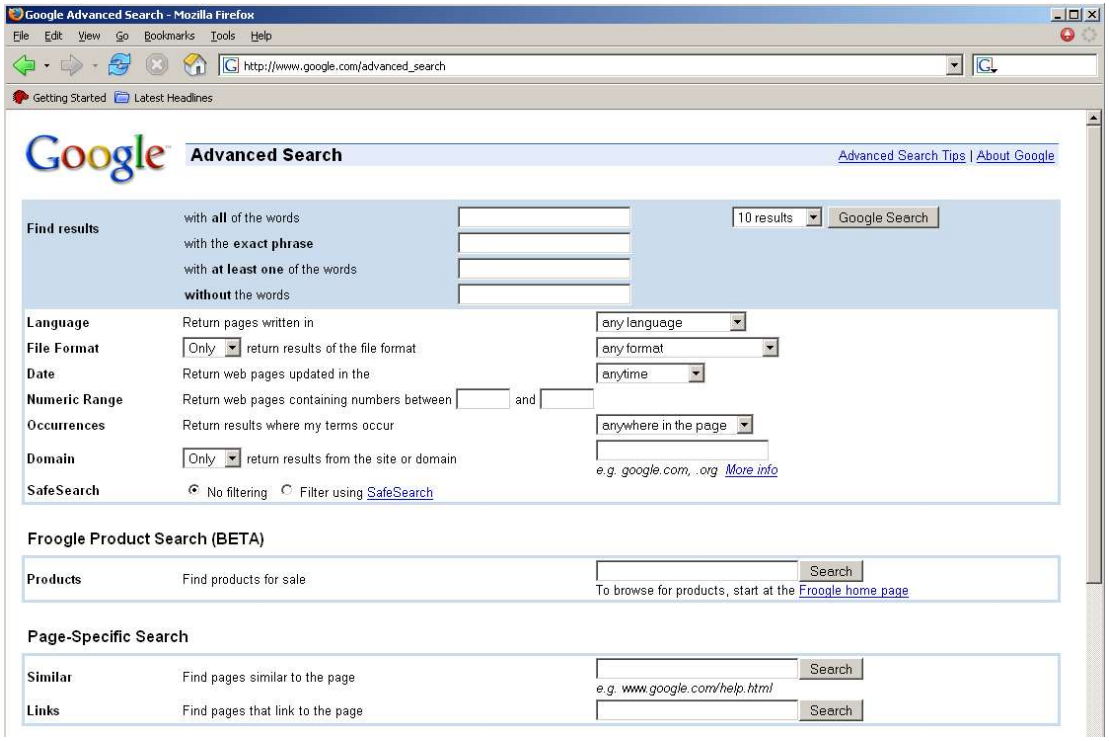
- AND is always implied.
- OR: Escobar (Narcotics OR Cocaine)
- "-" = NOT: Escobar -Pablo
- "+" = MUST: Escobar +Roberto
- Use quotes for exact phrase matching:
  - "nobody puts baby in a corner"

Wildcards

- Google supports word wildcards but NOT stemming.
  - "It's the end of the * as we know it" works.
  - but "American Psycho*" won't get you decent results on American Psychology or American Psychophysics.

Advanced Searching

Advanced Search Page:
http://www.google.com/advanced_search



Advanced Operators

- cache:
- define:
- info:
- intext:
- intitle:
- inurl:
- link:
- related:
- stocks:
- filetype:
- numrange 1973..2005
- source:
- phonebook:

DEMO:

on-2-13-1973..2004

visa 4356000000000000..4356999999999999

http://www.googleguide.com/advanced_operators.html and **http://tinyurl.com/5yjnx**

Review: Basic Search

- Use the plus sign (+) to force a search for an overly common word. Use the minus sign (-) to exclude a term from a search. No space follows these signs.
- To search for a phrase, supply the phrase surrounded by double quotes (" ").
- A period (.) serves as a single-character wildcard.
- An asterisk (*) represents any word—not the completion of a word, as is traditionally used.
- Source: **http://tinyurl.com/dnhc3**

Advanced Operators

- Google advanced operators help refine searches. Advanced operators use a syntax such as the following:
- *operator*:*search_term*
  - Notice that there's no space between the operator, the colon, and the search term.
- The **site**: operator instructs Google to restrict a search to a specific web site or domain. The web site to search must be supplied after the colon.
- The **link**: operator instructs Google to search within hyperlinks for a search term.
- The **cache**: operator displays the version of a web page as it appeared when Google crawled the site. The URL of the site must be supplied after the colon.
  - Turn off images and you can look at pages without being logged on the server! Google as a mirror.

Other parts

- Google searches not only the content of a page, but the title and URL as well.
- The **intitle**: operator instructs Google to search for a term within the title of a document.
- The **inurl**: operator instructs Google to search only within the URL (web address) of a document. The search term must follow the colon.
- To find *every* web page Google has crawled for a specific site, use the **site**: operator.
    - Source: **http://tinyurl.com/dnhc3**

What Can Google Search?

- The **filetype**: operator instructs Google to search only within the text of a particular type of file. The file type to search must be supplied after the colon. Don't include a period before the file extension.
  - Everything listed at http://filext.com/ claims Johnny. Can also ,e.g., say filetype:phps to only search .phps files.
    - filetype:phps mysql_connect
- Adobe Portable Document Format (pdf)
- Adobe PostScript (ps)
- Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks, wku)
- MacWrite (mw)
- Microsoft Excel (xls)
- Microsoft PowerPoint (ppt)
- Microsoft Word (doc)
- Microsoft Works (wks, wps, wdb)
- Microsoft Write (wri)
- Rich Text Format (rtf)
- Shockwave Flash (swf)
- Text (ans, txt)
- And many more….

Directory Listings

- **Directory Listings**
  - **Show server version information**
    - **Useful for an attacker**
  - intitle:index.of server.at
  - intitle:index.of server.at site:aol.com
- **Finding Directory Listings**
  - intitle:index.of "parent directory"
  - intitle:index.of name size
- Displaying variables
  - “Standard” demo and debugging program
  - “HTTP_USER_AGENT=Googlebot”
  - Frequently an avenue for remote code execution
    - http://somebox.someU.edu/~user/demo.cgi?cmd=`cat /etc/passwd`

Default Pages

- Default Pages are another way to find specific versions of server software….

Apache Server Version	Query
Apache 1.3.0–1.3.9	Intitle:Test.Page.for.Apache It.worked! this.web.site!
Apache1.3.11–1.3.26	Intitle:Test.Page.for.Apache seeing.this.instead
Apache 2.0	Intitle:Simple.page.for.Apache Apache.Hook.Functions
Apache SSL/TLS	Intitle:test.page "Hey, it worked !" "SSL/TLS-aware"
Many IIS servers	intitle:welcome.to intitle:internet IIS
Unknown IIS server	intitle:"Under construction" "does not currently have"
IIS 4.0	intitle:welcome.to.IIS.4.0
IIS 4.0	allintitle:Welcome to Windows NT 4.0 Option Pack
IIS 4.0	allintitle:Welcome to Internet Information Server
IIS 5.0	allintitle:Welcome to Windows 2000 Internet Services
IIS 6.0	allintitle:Welcome to Windows XP Server Internet Services
Many Netscape servers	allintitle:Netscape Enterprise Server Home Page
Unknown Netscape server	allintitle:Netscape FastTrack Server Home Page

CGI Scanner

- Google can be used as a CGI scanner. The index.of or inurl searchs are good tools to find vulnerable targets. For example, a Google search for this:
- allinurl:/random_banner/index.cgi
  - Hurray! There are only three…
- the broken random_banner program to cough up any file on that web server, including the password file…

CGI & Other Server Side Programs

- Database errors
- Login portals
  - Coldfusion
  - Remote desktop
  - Dotproject
  - Citrix Metaframe
  - MS Outlook web access

Johnny’s Disclaimer

- “Note that actual exploitation of a found vulnerability crosses the ethical line, and is not considered mere web searching.”

Security Advisory + Source = Google Hack

- Security Advisories and application patches for web application explain the newly discovered vulnerability
- Analysis of the source code of the vulnerable application yields a search for un-patched applications
- Sometimes this can be very simple; e.g.:
  - “Powered by CuteNews v1.3.1”

Automation!

- CGIs and other active content can be located in several places on a server.
- Many queries need to be used to find a vulnerability.
- There are two ways to automate Google searches:
  - Plain old web robots
  - The Google API: http://www.google.com/apis/

Terms of Service

- http://www.google.com/terms_of_service.html
- "You may not send automated queries of any sort to Google's system without express permission in advance from Google. Note that 'sending automated queries' includes, among other things:
- using any software which sends queries to Google to determine how a web site or web page 'ranks' on Google for various queries;
- 'meta-searching' Google; and
- performing 'offline' searches on Google."

Google API

- The Google API is the blessed way of automating Google interaction.
- When you use the Google API you include your license string

Gooscan

- “The gooscan tool, written by j0hnny, automates CGI scanning with Google, and many other functions.
- Gooscan is a UNIX (Linux/BSD/Mac OS X) tool that automates queries against Google search appliances (which are not governed by the same automation restrictions as their web-based brethren). For the security professional, gooscan serves as a front end for an external server assessment and aids in the information-gathering phase of a vulnerability assessment. For the web server administrator, gooscan helps discover what the web community may already know about a site thanks to Google's search appliance.
- For more information about this tool, including the ethical implications of its use, see http://johnny.ihackstuff.com.”

Google Search Appliance?

- It sounds like a good idea to put a search appliance in the enterprise.
- Then someone has their source code searched.
  - /* TODO: Fix the major security hole here */

Googledorks?

- http://johnny.ihackstuff.com/googledorks
- The term "googledork" was coined by the author [Johnny Long] and originally meant "An inept or foolish person as revealed by Google."
- After a great deal of media attention, the term came to describe those who "troll the Internet for confidential goods."
- Either description is fine, really.
- What matters is that the term *googledork* conveys the concept that sensitive stuff is on the web, and Google can help you find it. The official googledorks page lists many different examples of unbelievable things that have been dug up through Google by the maintainer of the page, Johnny Long.
  - **http://tinyurl.com/2ywye**
- Each listing shows the Google search required to find the information, along with a description of why the data found on each page is so interesting.

GooPot

- According to http://www.techtarget.com, "A honey pot is a computer system on the Internet that is expressly set up to attract and 'trap' people who attempt to penetrate other people's computer systems."
- For example, build a page that matches the query:
  - inurl:admin inurl:userlist
- Then examine the referrer variable to figure out how the person found the page. This information can help protected normal sites.
- http://ghh.sourceforge.net/

Protecting Yourself from Google Hackers

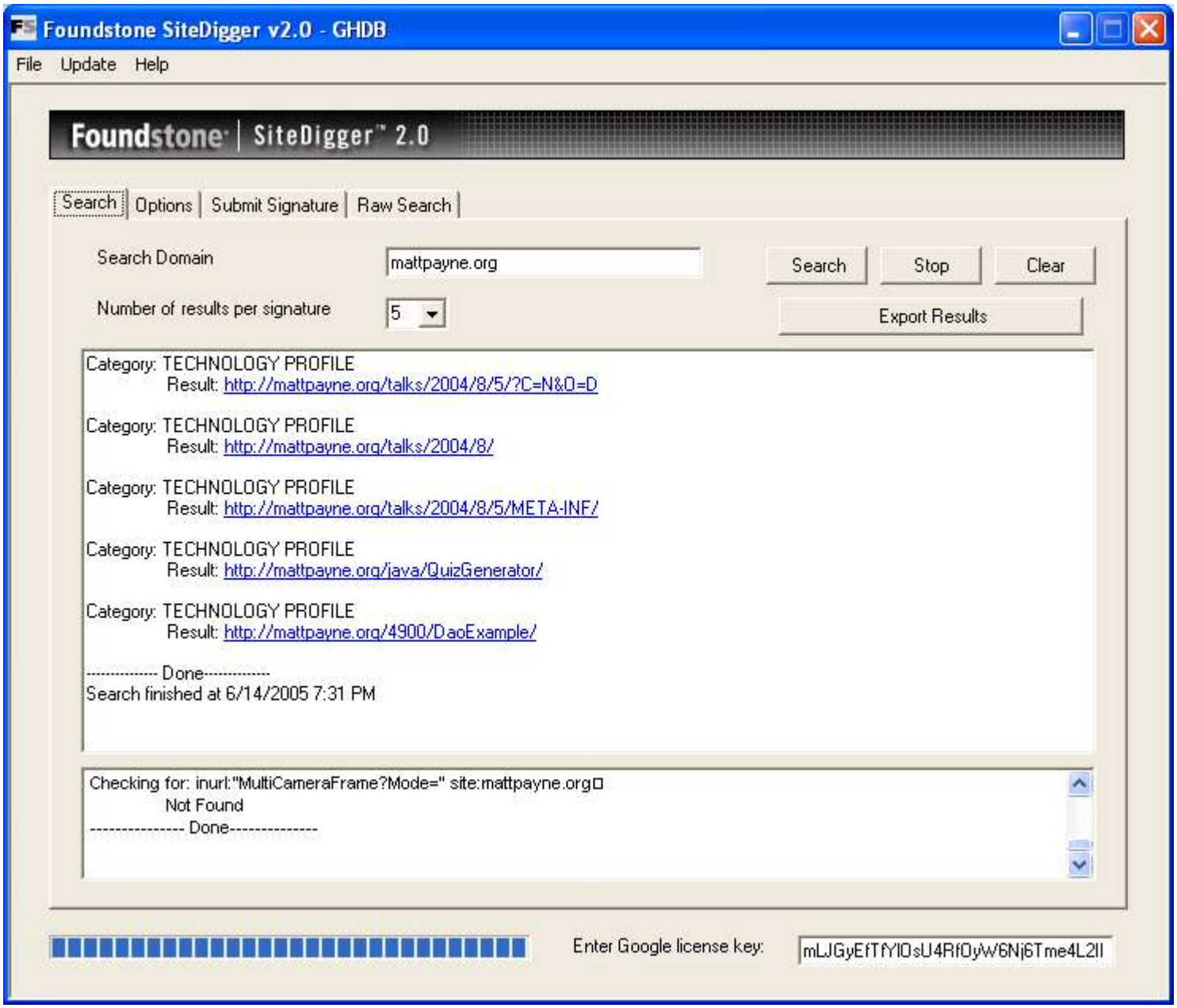
- **Keep your sensitive data off the web!** Even if you think you're only putting your data on a web site temporarily, there's a good chance that you'll either forget about it, or that a web crawler might find it. Consider more secure ways of sharing sensitive data, such as SSH/SCP or encrypted email.

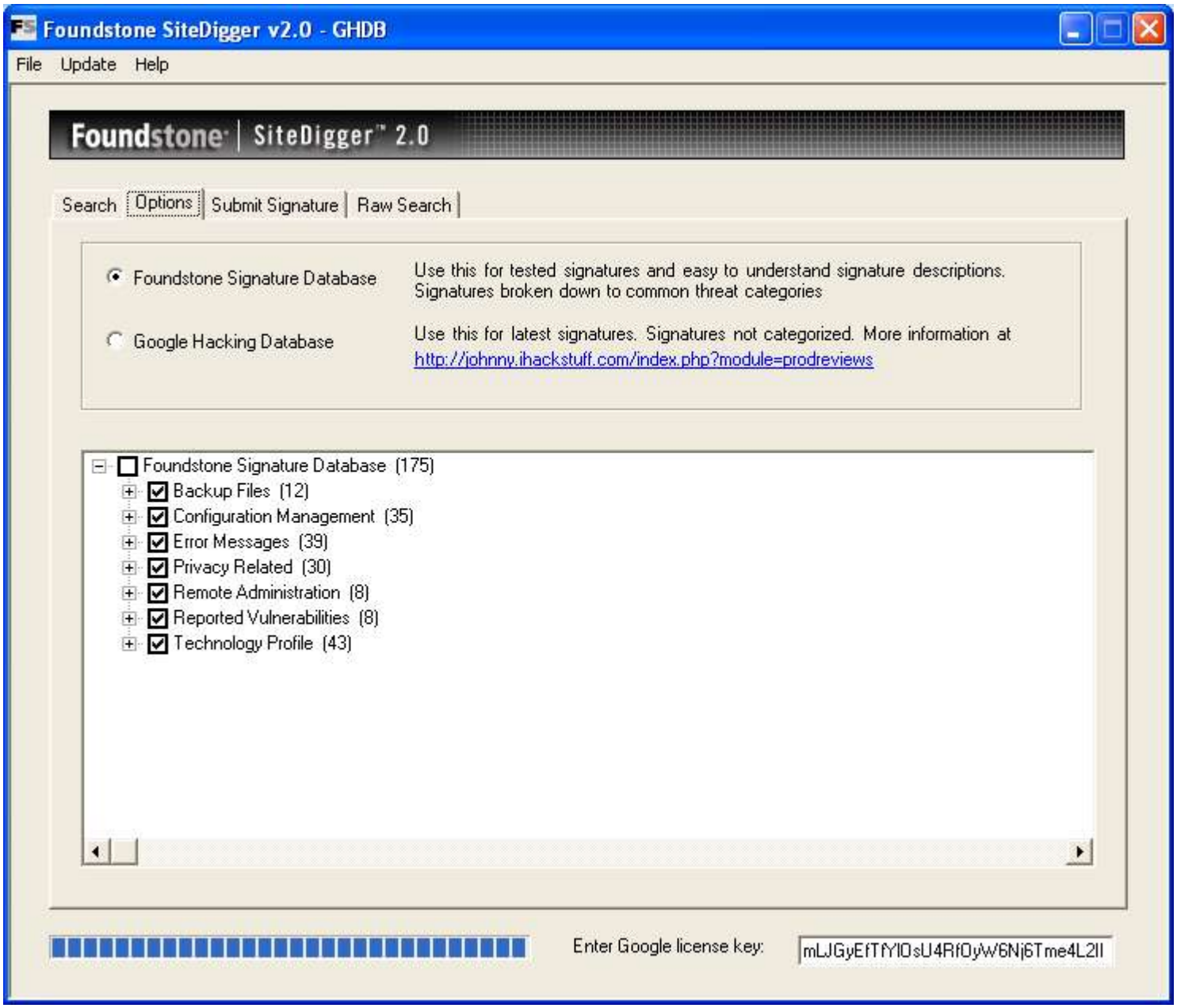
Protecting Yourself…

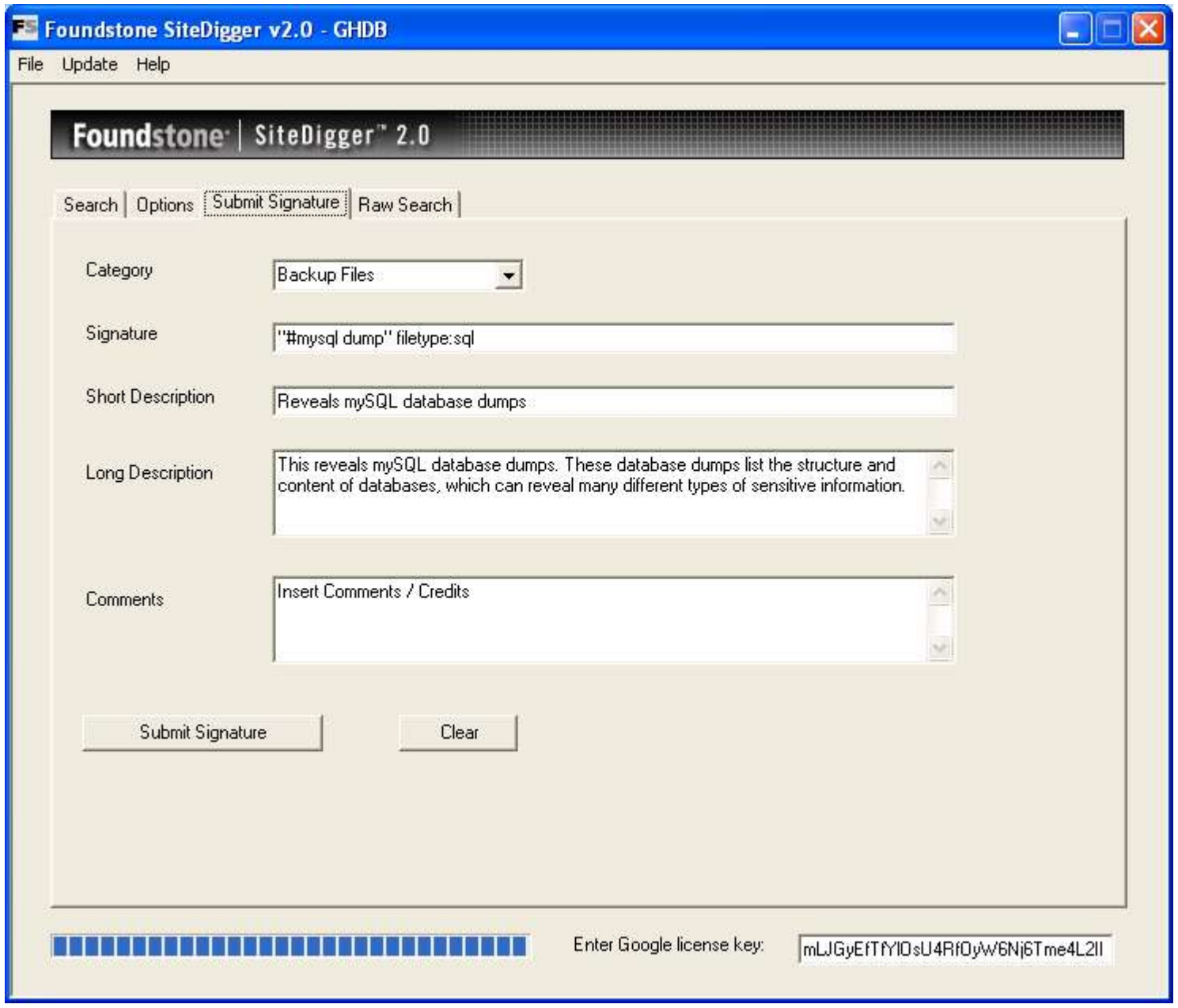
- **Googledork!** Use the techniques outlined in this article (and the full Google Hacker's Guide) to check your site for sensitive information or vulnerable files.
- SiteDigger from FoundStone automates this.
  - Uses the Google API so…
    - Only 1000 searches on Google per day
  - Free beer!

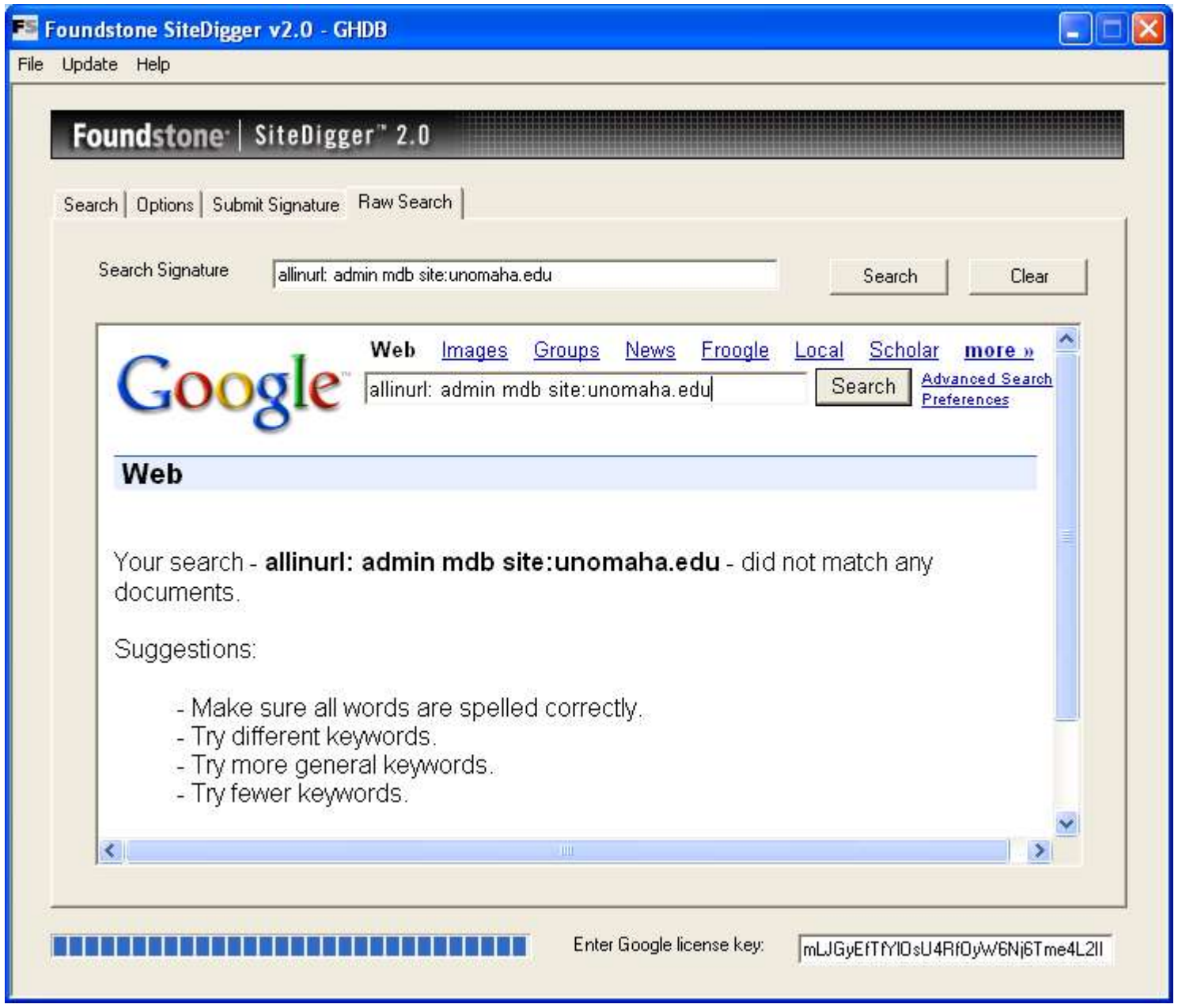
SiteDigger 2.0

- **http://tinyurl.com/28aeh**
- The tool requires Google web services API license key.
  - Your license key provides you access to the Google Web APIs service and entitles you to 1,000 queries per day.
- **System Requirements**
  Windows .NET Framework (can be installed using Windows Update)


Foundstone SiteDigger v2.0 - GHDB
File Update Help
Foundstone | SiteDigger™ 2.0
Search Options Submit Signature Raw Search
Search Domain
mattpayne.org
Search Stop Clear
Number of results per signature
5
Export Results
Category: TECHNOLOGY PROFILE
Result: http://mattpayne.org/talks/2004/8/5/?C=N&O=D
Category: TECHNOLOGY PROFILE
Result: http://mattpayne.org/talks/2004/8/
Category: TECHNOLOGY PROFILE
Result: http://mattpayne.org/talks/2004/8/5/META-INF/
Category: TECHNOLOGY PROFILE
Result: http://mattpayne.org/java/QuizGenerator/
Category: TECHNOLOGY PROFILE
Result: http://mattpayne.org/4900/DaoExample/
---------- Done----------
Search finished at 6/14/2005 7:31 PM
Checking for: inurl:"MultiCameraFrame?Mode=" site:mattpayne.org
Not Found
---------- Done----------
Enter Google license key:
mLJGyEfTfYlOsU4RfOyW6Nj6Tme4L2II


Foundstone SiteDigger v2.0 - GHDB
File Update Help
Foundstone | SiteDigger™ 2.0
Search Options Submit Signature Raw Search
Foundstone Signature Database
Use this for tested signatures and easy to understand signature descriptions. Signatures broken down to common threat categories
Google Hacking Database
Use this for latest signatures. Signatures not categorized. More information at http://johnny.ihackstuff.com/index.php?module=prodreviews
Foundstone Signature Database (175)
Backup Files (12)
Configuration Management (35)
Error Messages (39)
Privacy Related (30)
Remote Administration (8)
Reported Vulnerabilities (8)
Technology Profile (43)
Enter Google license key:
mLJGyEfTfYlOsU4RfOyW6Nj6Tme4L2II


Foundstone SiteDigger v2.0 - GHDB
File Update Help
Foundstone | SiteDigger™ 2.0
Search Options Submit Signature Raw Search
Category
Backup Files
Signature
"#mysql dump" filetype:sql
Short Description
Reveals mySQL database dumps
Long Description
This reveals mySQL database dumps. These database dumps list the structure and content of databases, which can reveal many different types of sensitive information.
Comments
Insert Comments / Credits
Submit Signature
Clear
Enter Google license key:
mLJGyEfTfYlOsU4RfOyW6Nj6Tme4L2II


Foundstone SiteDigger v2.0 - GHDB
File Update Help
Foundstone | SiteDigger™ 2.0
Search Options Submit Signature Raw Search
Search Signature
allinurl: admin mdb site:unomaha.edu
Search
Clear
Google
Web Images Groups News Froogle Local Scholar more »
allinurl: admin mdb site:unomaha.edu
Search
Advanced Search
Preferences
Web
Your search - allinurl: admin mdb site:unomaha.edu - did not match any documents.
Suggestions:
- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.
Enter Google license key:
mLJGyEfTfYlOsU4RfOyW6Nj6Tme4L2II

Protecting yourself…

- **Consider removing your site from Google's index.**
  http://www.google.com/remove.html

Robots.txt

- **Use a robots.txt file.** Web crawlers are supposed to follow the robots exclusion standard. This standard outlines the procedure for "politely requesting" that web crawlers ignore all or part of your web site. This file is only a suggestion. The major search engine's crawlers honor this file and its contents. For examples and suggestions for using a robots.txt file, see http://www.robotstxt.org.

Example Robots.txt

- User-agent: *
- Disallow: /images/
- Disallow: /stats/
- Disallow: /logs/
- Disallow: /admin/
- Disallow: /comment/
- User-agent: Googlebot
- Allow:
- User-agent: BecomeBot
- Disallow:
- Disallow: /
- Disallow: *
- User-agent: MSNBot
- Disallow:
- Disallow: /
- Disallow: *

- By default tells others to not scan specific paths
- Allows Google to scan
- Tells BecomeBot and MSNBot to go away entirely.
- Please the robots.txt in the root of your HTML documents directory.

- See also
- Removing Your Materials from Google **How to remove your content from Google's various web properties.**
- http://hacks.oreilly.com/pub/h/220

- Robots.txt generator **http://tinyurl.com/7pc4k**

CAPTCHA

- Completely Automated Public Turing Test to Tell Computers and Humans Apart

- http://www.captcha.net/
- http://en.wikipedia.org/wiki/Captcha

Google Extras...

- Translation and Language options - over 100 to choose from: http://www.google.com/language_tools
- Stock Quotes - enter stocks:, example: stocks:GOOG
- Newsgroups - http://groups.google.com
- Calculator - "1024 minus 768" or "12 to the 10 power"
- Froogle - http://froogle.google.com
- Images - http://images.google.com
- Spell Checking - just type it in: "convienence"
- Blogger - http://www.blogger.com/start

Extras can be found at http://www.google.com/help/features.html

Sets from Google Labs

- http://labs.google.com/sets
- Automatically create sets of items from a few examples.
- When you’re tired of relating keywords yourself, let Google do it for you….

References

http://bss.sfsu.edu/bsscomputing/training/onth

http://www.googleguide.com/advanced_operators.html

Google Hacking Mini Guide by Johnny Long

http://www.informit.com/articles/article.asp?p=170

Search Engine Watch:

http://searchenginewatch.com

References

1. Google Hacks: 100 Industrial-Strength Tips & Tools
2. by Tara Calishain, Rael Domfest
3. Protect yourself from Google hacking: **http://tinyurl.com/8q3fg**
4. Johnny I Hack Stuff: http://johnny.ihackstuff.com
5. Google:http://www.google.com
6. http://www.i-hacked.com/content/view/23/42/
7. HowStuffWorks:
8. http://computer.howstuffworks.com/search-engine1.htm

Interesting Searches…

- Source http://www.i-hacked.com/content/view/23/42/

- intitle:"Index of" passwords modified
- allinurl:auth_user_file.txt
- "access denied for user" "using password“
- "A syntax error has occurred" filetype:ihtml
- allinurl: admin mdb
- "ORA-00921: unexpected end of SQL command“
- inurl:passlist.txt
- "Index of /backup“
- "Chatologica MetaSearch" "stack tracking:"

Credit Cards

- Number Ranges to find Credit Card Numbers
  - Amex Numbers: 300000000000000..399999999999999
  - MC Numbers: 5178000000000000..5178999999999999
  - visa 4356000000000000..4356999999999999

Listings of what you want

- change the word after the parent directory to what you want
- "parent directory " **DVDRip** -xxx -html -htm -php -shtml -opendivx -md5 -md5sums
- "parent directory "**Xvid** -xxx -html -htm -php -shtml -opendivx -md5 -md5sums
- "parent directory " **Gamez** -xxx -html -htm -php -shtml -opendivx -md5 -md5sums
- "parent directory " **MP3** -xxx -html -htm -php -shtml -opendivx -md5 -md5sums
- "parent directory **" Name of Singer or album”** -xxx -html -htm -php -shtml -opendivx -md5 -md5sums

Music

- You only need add the name of the song/artist/singer.
- Example: intitle:index.of mp3 jackson

CD Images

- inurl:microsoft filetype:iso
- You can change the string to whatever you want, ex. Microsoft to Adobe, .iso to .zip etc…

Passwords

- "# -FrontPage-" inurl:service.pwd
  FrontPage passwords.. very nice clean search results listing !!

  "AutoCreate=TRUE password=*"
  This searches the password for "Website Access Analyzer", a Japanese software that creates webstatistics. For those who can read Japanese, check out the author's site at:
  http://www.coara.or.jp/~passy/

Passwords in the URL

- "http://*:*@www" domainname
  This is a query to get inline passwords from search engines (not just Google), you must type in the query followed with the domain name without the .com or .net

  "http://*:*@www" gamespy or http://*:*@www”gamespy

  Another way is by just typing
  "http://bob:bob@www"

IRC Passwords

- "sets mode: +k"
  This search reveals channel keys (passwords) on IRC as revealed from IRC chat logs.

- eggdrop filetype:user user
  These are eggdrop config files. Avoiding a full-blown discussion about eggdrops and IRC bots, suffice it to say that this file contains usernames and passwords for IRC users.

Access Database Passwords

- allinurl: admin mdb
  Not all of these pages are administrator's access databases containing usernames, passwords and other sensitive information, but many are!

DCForum Passwords

- allinurl:auth_user_file.txt
  DCForum's password file. This file gives a list of (crackable) passwords, usernames and email addresses for DCForum and for DCShop (a shopping cart program(!!!). Some lists are bigger than others, all are fun, and all belong to googledorks. =)

MySQL Passwords

- intitle:"Index of" config.php
- This search brings up sites with "config.php" files. To skip the technical discussion, this configuration file contains both a username and a password for an SQL database. Most sites with forums run a PHP message base. This file gives you the keys to that forum, including FULL ADMIN access to the database.

The ETC Directory

- intitle:index.of.etc
  This search gets you access to the etc directory, where many, many, many types of password files can be found. This link is not as reliable, but crawling etc directories can be really fun!

Passwords in backup files

- filetype:bak inurl:"htaccess|passwd|shadow|htusers"
  This will search for backup files (*.bak) created by some editors or even by the administrator himself (before activating a new version).
  Every attacker knows that changing the extension of a file on a web server can have ugly consequences.

Serial Numbers

- Let's pretend you need a serial number for Windows XP Pro.
- In the Google search bar type in just like this - "Windows XP Professional" 94FBR
- the key is the 94FBR code.. it was included with many MS Office registration codes so this will help you dramatically reduce the amount of 'fake' sites (usually pornography) that trick you.
- or if you want to find the serial for WinZip 8.1 - "WinZip 8.1" 94FBR