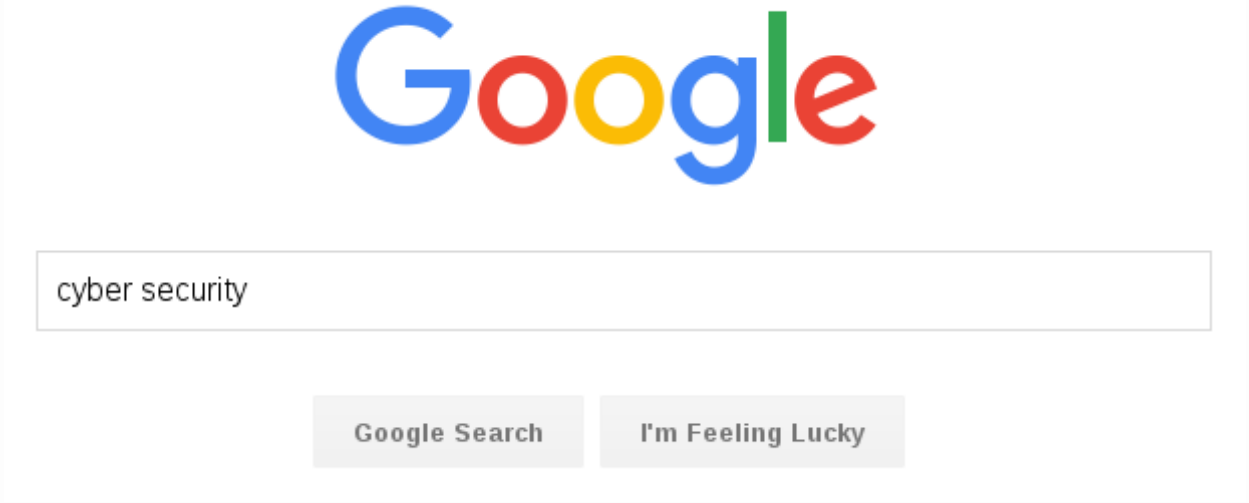
Information Gathering with Google Search Engine

18 July 2016 • 4 mins read

• Information gathering • Google advanced search operators • OSINT

Generally, when we want to search for a particular argument, we open up our favourite browser, navigate to a Web Search Engine and type in some words related to that matter. Depending from how good we set up the research, we obtain more or less pertinent results.

Basically, everytime we launch a search we make a query to the web search engine: there are some particular expressions known to the engine, called Advanced Search Operators, which make a search more effective. Queries built like these are also called “Google dorks”.
For example, we can make a simple search for “cyber security” using Google Search Engine:



By doing this we obtain a good amount of informations about the argument; note that we could achieve the same result by using the following url instead of compiling the search form:

```
https://www.google.com/search?q=cyber+security
```

Google Advanced Search Operators and Special Search Characters

What if we want to do a more specific query, like obtaining all the results for that topic from a particular website? Here come advanced search operators. Suppose we want to analyze results coming from NIST (National Institute of Standards and Technology) for that topic; then the query should be:

```
site:nist.gov "cyber security"
```

The operator “site:” asks to the engine to show all the results associated to the website (which represent the argument of the operator) reported after the colon symbol; beware that there must be no space between the colon symbol and the argument of the operator. In addition to search operators, there are special search characters like the double quotes used before to search phrases or terms composed by more than one word.

A pretty good list of Google advanced search operators is this one from DEFCON 2005 Conference:

Operator	Purpose	Mixes with other operators?	Can be used alone?	Does search work in			
				Web	Images	Groups	News
intitle	Search page title	yes	yes	yes	yes	yes	yes
allintitle	Search page title	no	yes	yes	yes	yes	yes
inurl	Search URL	yes	yes	yes	yes	not really	like intitle
allinurl	Search URL	no	yes	yes	yes	yes	like intitle
filetype	Search specific files	yes	no	yes	yes	no	not really
allintext	Search text of page only	not really	yes	yes	yes	yes	yes
site	Search specific site	yes	yes	yes	yes	no	not really
link	Search for links to pages	no	yes	yes	no	no	not really
inanchor	Search link anchor text	yes	yes	yes	yes	not really	yes
numrange	Locate number	yes	yes	yes	no	no	not really
daterange	Search in date range	yes	no	yes	not really	not really	not really
author	Group author search	yes	yes	no	no	yes	not really
group	Group name search	not really	yes	no	no	yes	not really
insubject	Group subject search	yes	yes	like intitle	like intitle	yes	like intitle
msgid	Group msgid search	no	yes	not really	not really	yes	not really

List for special search characters:

- (+) force inclusion of something common;
- (-) exclude a search term;
- (") use quotes around search phrases;
- (.) a single-character wildcard;
- (*) any word;
- (&) boolean 'AND';
- (|) boolean 'OR';

The special char “-“ is really interesting for listing subdomains. Considering the next query we obtain a list of the subdomains of nist.gov:

site:nist.gov -site:www.nist.gov

All Images News Shopping Maps More Search tools

About 1,630,000 results (0.34 seconds)

The United States Government Configuration Baseline (USGCB) - NIST

https://usgcb.nist.gov/ National Institute of Standards and Technology

Feb 19, 2010 - The purpose of the United States Government Configuration Baseline (USGCB) initiative is to create security configuration baselines for ...

Thermodynamics Research Center - Thermophysical Properties ...

trc.nist.gov/ National Institute of Standards and Technology

Thermodynamic properties tables from NIST's Thermodynamic Research Center offer rigorous chemical and thermophysical properties data over the web.

NIST Computer Security Resource Center

csrc.nist.gov/ National Institute of Standards and Technology

Jan 28, 1996 - CSD publications, events, cryptographic standards and applications. Information on security testing, security management, and research ...

Math, Statistics, and Computational Science - National Institute of ...

math.nist.gov/ National Institute of Standards and Technology

Jun 15, 1994 - Gateway to organizations and services related to applied mathematics, statistics, and computational science at the National Institute of ...

NIST Stone Test Wall

stonewall.nist.gov/ National Institute of Standards and Technology

Dec 19, 2012 - The stone test wall was constructed to study the performance of stone subjected to weathering. It contains 2352 individual samples of stone, ...

Adding more “-“ condition allows to crawl even deeper: keep in mind that Google ranks the results, so the most searched stuff lies at the top, i.e. in the first pages, while the most interesting informations (from an attacker point of view) can be found at the bottom.

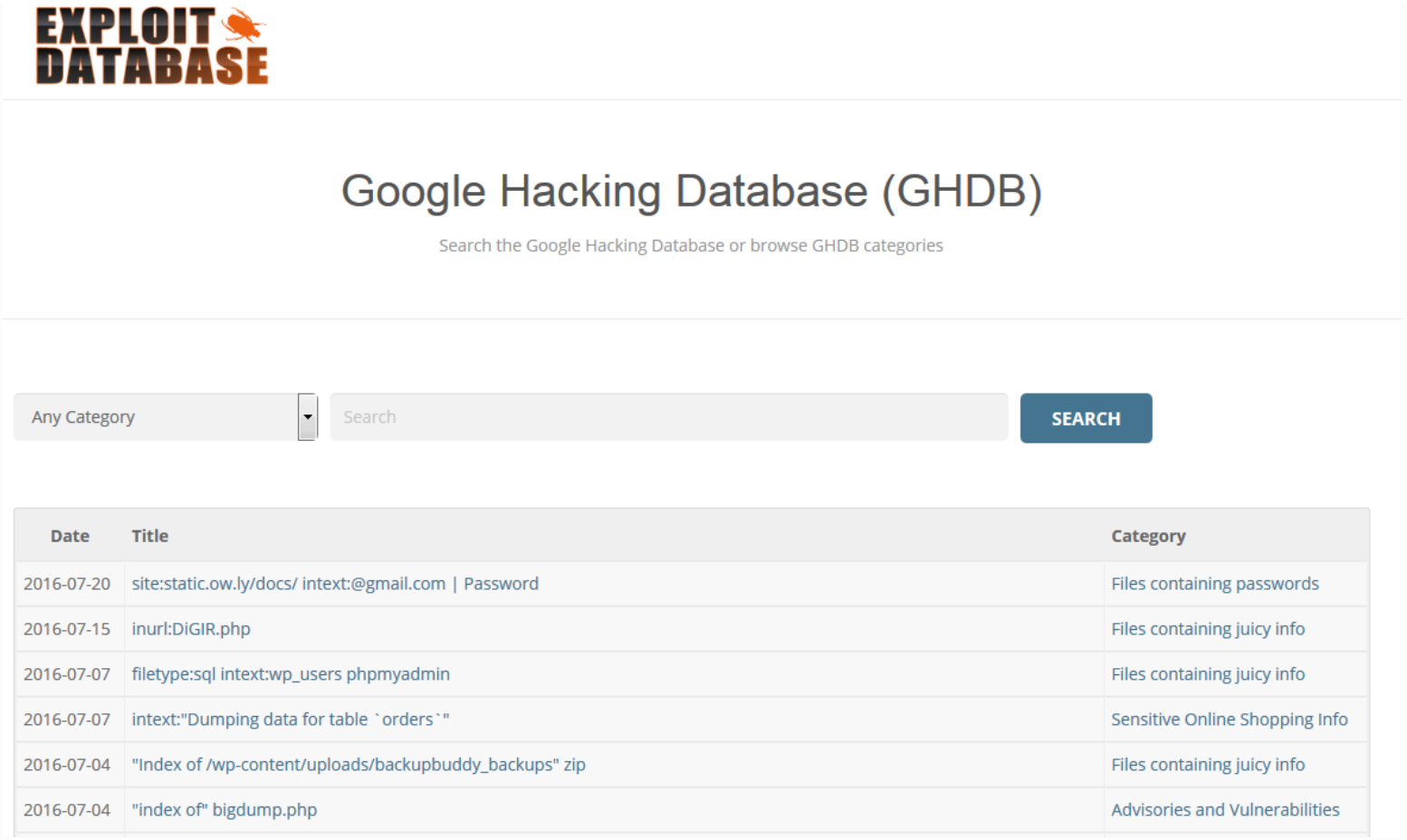
Now we will look for something more interesting in terms of confidential informations; suppose we want to know if in a certain website there is a PDF document containing the word “password”. For this purpose, we query Google with the following request:

```
site:www.nameoftargetsite.com filetype:pdf intext:password
```

The meaning of the previous line is explained next: the first operator is well known; the “filetype:” directive tells Google we want to search all the files with a particular extension, in this case PDF. Lastly, we add a third condition which makes the engine search for a specific word in the text. So, interpreting the whole query: we are asking to search through the site www.nameoftargetsite.com all PDF documents containing the word “password”.

You can easily see how powerful this syntax is and you will be surprised by discovering the amount of confidential informations left to public access. This is why it is important to configure properly the access to informations hosted on the net, for example in a website. However, if not really needed, it is a good practice to avoid hosting of confidential documents on websites, since they are the first contact point for attackers.

If you want to take a look at a really good list of Google dorks I suggest you to visit the following website: https://www.exploit-db.com/google-hacking-database



Here you can search through the Google Hacking DataBase choosing the query category you are interested in and then simply use it to search through Google. Keep in mind that this type of information gathering through search engine is a passive reconnaissance activity since we are not interacting directly with the systems; this helps to maintain a low profile making hard to be detected.

All these search activities can be executed by using automatic tools, which leverage not only Google, but a huge variety of search engines and social network platforms; we will take a look at some of them in the next posts.