Daniel Mider
Jan Garlicki
Wojciech Mincewicz

The Internet Data Collection with the Google Hacking Tool – White, Grey or Black Open-Source Intelligence?

Google search engine is as much common (and having almost no alternative), as it is unknown at the same time. Its potentials in the so-called sensitive data collection regarding individuals and institutions are underestimated. The well-constructed query, submitted to the Google, makes it possible to find deleted and archival sites, restore the structure of a webpage or the structure of an intranet, access server configuration parameters, obtain information secured intentionally from unauthorized access (paywall, passwords), obtain user names and passwords, their identification numbers (for example the numbers of payment cards, IDs, etc.) and access device configuration parameters (servers, web cameras, routers and others), to take control of them. Such activities are described in the literature as Google Hacking (GH), Google Dorks (GD) or (rarely) – Google Scanning (GS) or Engine Hacking (EH). These terms mean such queries to the Google search engine so that it would make such data available to the users, who are unauthorized in ethical or legal sense, or both.[1]

The notion of Google Hacking was introduced by the authority figure in the area, Johnny Long.[2] The notion of Google Dork means a person who is inept in securing online content, mainly web sites. The inability can be revealed by Google very easily. As the author points out, the meaning of this lexeme has changed over the time and currently it signifies a person who obtains confidential information from the Google.[3]

The history of GH/GD started with the defining of the phenomenon in December 2002 by Johnny Long, the method's pioneer, although it existed already in 2000

[1] Cf. J. Long, *Google Hacking for Penetration Testers*, Rockland 2007, p. 534. Collins dictionary contains also a similar definition: https://www.collinsdictionary.com/submission/9695/google+dorks [access: 26 I 2018]. Apart from such definitions there are other using IT sociolect, although meaning more or less the same, for example the understanding of GH as "consistent search exploits" or "obtaining a sui generis virtual notebook". Cf. *Google Hacking – w służbie hakerom*, "Haker.edu.pl", 10 VII 2015, in: https://haker.edu.pl/2015/07/10/google-hacking-google-dorks/ [access: 26 I 2018]. There are two most popular notions used interchangeably in the text, i.e. Google Hacking and Google Dorks. Cf. *Roll Call Release. Intelligence for Police, Fire, EMS, and Security Personnel*, 7 VII 2014, https://info.publicintelligence.net/DHS-FBI-NCTC-GoogleDorking.pdf [access: 26 I 2018].

[2] He is a famous author of the already non-existent web page http://johnny.ihackstuff.com; at present its content was moved to http://www.hackersforcharity.org/ghdb/, i.e. Google Hacking Database. He is also known as "j0hnny" and "j0hnnyhax".

[3] J. Long, *The Google Hacker's Guide. Understanding and Defending Against the Google Hacker*, http://pdf.textfiles.com/security/googlehackers.pdf [access: 26 I 2018].

as a technique used by hackers.[4] In 2004 the first IT self-defensive tool against GH was constructed, i.e. SiteDigger (1.0). One year later Google Hack Honeypot and MSNPawn (as a matter of fact Search.msn) programmes were made available. The fundamental work *Google Hacking for Penetration Testers*[5] was published in 2005 for the first time and reprinted many times from that moment on. The notion of GH/GD can be a bit misleading. The same data can be collected also from other search engines like Bing, Yahoo, Yandex and DuckDuckGo.[6] Differences in the structure of the engines (i.e. commands made by the user, utilizing pre-defined elements, called operators) are usually slim and involve their names, the way some of them are written (George Boole operators) and some operators existing exclusively in one given search engine.[7]

The research literature comprises different proposals of GH queries classification and systematics. For example J. Long names as many as 14 types of GH queries, while Flavio Toffalini and his team suggest the following four general categories of queries: 1. – localizing servers (software versions), 2. – localizing sensitive folders, 3. – localizing files with passwords and 4. – localizing files (logs) with system errors.[8] Another idea of organizing the area is based on the following triad, regarding the qualities of the queries, not their function: 1. – regarding URL (Uniform Resource Locator) structure, 2. – regarding file extension, 3. – regarding file contents.[9] The spectrum of acceptance for data collection ranges from authorized – as they are legal and ethical, to unauthorized – as they are collected illegally and unethically. Three classes of data collection have been adopted: white, grey and black open-source intelligence-collection.[10] White open-source intelligence is the way of data collection which does not arouse any ethical, or legal doubts. It is estimated that 80% of data

[4] *Smart searching with Google Dorking*, „Exposing the Invisible", https: //exposingtheinvisible.org/guides/google-dorking/#dorking-operators-across-google-duckduckgo-yahoo-and-bing [access: 26 I 2018].

[5] J. Long, B. Gardner, J. Brown, *Google Hacking for Penetration Testers*, Amsterdam 2005.

[6] It is worth pointing out the existence of search engines like Shodan, Censys and Zoomeye, which are particularly interesting, innovative and not typical projects in the categories of such search engines. They are Internet of Things search engines, i.e. computer and online devices (in fact, from the IT and technical point of view they are scanners of online ports). The search engines are to be found on: https://www.shodan.io/; https://censys.io/; https://zoomeye.org.

[7] There is, for example, the "language:" operator in Yahoo and Bing used with a certain language code, enabling the search of a particular term in this particular language. Google has "book:" and "maps:" operators enabling the search of files, defined as books or maps. The "cache:" operator exists in Google only, while "site:", "intitle:" and "filetype:" operators are universal. The authors have a comparative set of operators (which is constantly supplemented). They share them upon request.

[8] F. Toffalini and others, *Google Dorks: Analysis, Creation, and new Defenses*, in: *Detection of Intrusions and Malware, and Vulnerability 2016*, J. Caballero and others (ed.), San Sebastian 2016, pp. 255–275.

[9] Ibidem.

[10] Cf. A.W. Dorn, *United Nations Peacekeeping Intelligence*, in: *National Security Intelligence*, L.K. Johnson (ed.), Oxford 2010, p. 280.

are collected at present from the exploration of open and overt sources, for example state-owned, press or private open sources. Black open-source intelligence, in turn, is unequivocally unethical and illegal. This group comprises, *inter alia*, tapping and bugging phones and rooms, burglary, identity and biometric data thefts, breaking cryptographic security tools and obtaining information through blackmail and corruption. According to Kazimierz Turaliński, this type of open-source intelligence provides 5% of all data. The grey open-source intelligence is situated between the two above mentioned types; it regards the activities which cannot be ranked unequivocally. The activities are legal but unethical. They comprise surveillance (infiltration and monitoring), as well as surveillance combined with an action of socio-technical nature. This way ca. 15% information is collected.[11] The division of the article's contents into white, grey and black intelligence techniques has a value as a matter itself, as well as it serves the organizational purpose. It can help in recognizing and distinguishing of what is permitted or not, from the perspective of different normative systems (law and ethics). White open-source intelligence comprises Google searching techniques which are not controversial from legal or ethical perspective, i.e. searching for archive and deleted information, as well as obtaining publicly available personal data. Grey open-source intelligence is understood as the obtaining of sensitive personal data, getting access to secured contents, access to online devices and to device configuration parameters (for example printer files), re-configuring the structure of web pages or intranet structures. GH black open-source intelligence comprises the obtaining of lists of users and passwords, sets of highly sensitive data (key personal data, like credit card numbers, ID numbers, social security numbers) and access to online devices, both private, as well as institutional (state institutions and private businesses), including monitoring network/cameras.

It is worth pointing out that using the techniques and tools analyzed in this article may be a subject to criminal liability. The first of the so-called "hacker paragraphs" described in article 267 of the Polish Penal Code is of a particular significance. Using GH in an unauthorised way that is without a written consent or the knowledge of the subject, to whom such kind of activity refers, is a violation of Article 267 of the Polish Penal Code. This article provides a penalty of two years of imprisonment, or the restriction of freedom or it imposes a fine for the obtaining of information without authorization or for obtaining access to the whole or to a part of the information system. The regulations in this respect have been a little liberated recently, on 11 April 2017, the President of Poland Andrzej Duda signed amendments to the Polish Penal Code. The so called "hacker articles", i.e. the Articles from 267 to 269b, have been supplemented with circumstances excluding the lawlessness. The previous legislation provided for unquestioned adjudicating penalties for manufacturing, obtaining, selling or facilitating to other people of devices or programmes adapted for IT-related crimes

[11] K. Turaliński, *Wywiad gospodarczy i polityczny. Podręcznik dla specjalistów ds. bezpieczeństwa, detektywów i doradców gospodarczych*, Warsaw 2015, pp. 31–33.

(Article 269b), including the extended seizure. Such legal arrangement disregarded the activities of professional security testers and researchers developing new security methods. Upon the amendments their activities are entirely lawful.[12]

There were three goals to be reached by this article. The first one is to make the readers aware of the threats and opportunities connected to the matter. The second is to present information regarding practical aspects that would allow them to use the described techniques on their own. There is, however, with one reservation in this respect: the article does not comprise a comprehensive compendium of techniques how to obtain data from the Internet, but it provides for typical techniques that make further self-development easier in the field. Cognitive efforts were focused on the Google search engine because it has a dominant position in the Internet – nearly nine out of ten queries (87,16%) are submitted to it.[13] The analysis of obtaining information via FOCA program is a complement of the article. Google robots do not index the so-called meta-tags of the documents[14] (which contain *inter alia* dates of their production, modification, personal data of their authors, etc.). However, the FOCA program[15] (Fingerprinting Organizations with Collected Archives) enables, *inter alia*, to obtain the data.

1. Google Hacking as white open-source intelligence

Google services described as white open-source intelligence are, as follows: searching for deleted and archival web pages, searching for information on users (e-mail addresses, social networks users, phone numbers – the service limited only to the USA)

[12] The issue is regulated by Article 269a section 1a of the Penal Code: "The crime described in §1 is not committed if the person acts only to secure the IT system, the communication and information system or communication and information net from the crime provided for in this article or to develop a method of such protection". A thorough analysis of the so called "hacker" articles was conducted by Filip Radoniewicz. Cf. F. Radoniewicz, *Odpowiedzialność karna za przestępstwo hackingu,* https://www.iws.org.pl/pliki/files/Filip%20Radoniewicz%2C%20Odpowiedzialno%C5%9B%C4%87%20karna%20za%20przest%C4%99pstwo%20hackingu%20%20121.pdf [access: 5 VI 2018] – The legal status for 2013, before the amendment of April 2017. The Ministry of Digitisation presented the Cybersecurity strategy for 2017-2020 which considers the possibility of regulating the bug-bounty system in 2017. It should therefore be presumed that a quick and cheap method of testing IT security will be introduced in our country. Cf. M. Długosz, *Legalny hacking w Polsce. (Analiza)*, 30 V 2017, http://www.cyberdefence24.pl/legalny-hacking-w-polsce-analiza [access: 5 VI 2018].

[13] C. Glijer, *Ranking światowych wyszukiwarek 2017: Google, Bing, Yahoo, Baidu, Yandex, Seznam*, „K2 Search", 25 VII 2017, http://k2search.pl/ranking-swiatowych-wyszukiwarek-google-bing-yahoo-baidu-yandex-seznam/ [access: 26 I 2018].

[14] It is only Bing search engine that indexes document meta-tags. Bing introduced a "meta" operator for this purpose.

[15] FOCA is an analytical tool created by ElevenPaths to find, download and analyze documents for metadata and other hidden information that may not be easily visible. Detailed information and a free version are available at: https://www.elevenpaths.com/labstools/foca/index.html [access: 29 I 2018].

and facilitating the searching of files of interest (searching by hashtags, searching for similar web pages or connected words, searching for definitions of encyclopaedic and lexicographic notions, searching for pages with references to other pages of interest, searching for certain types of files).[16]

Searching for deleted and archival pages

The most practical and at the same time - interesting Google service, which can be classified as white open-source intelligence, is the option of searching for deleted and archival pages. One can do it with "cache:" operator. The operator works in such a way that it shows a historical (deleted) version of a web page stored by the Google in the cache. A typical syntax is the following:

cache:www.inp.uw.edu.pl

After the above-provided command is written into the Google search engine, we are provided the access to the former version of the web page of the Institute of Political Science, University of Warsaw. The command allows to bring the full version up (HTMP or written in any other xTalk and displaying "what is"), the text version or the source (of the script). Also the exact time is given (date, hour, minute, second), of the indexation made by the Google spider. The page is displayed in the form of graphic file, although searching within its frames is still possible (CTRL+F shortcut). The results of the "cache:" command depend on how often Web pages are indexed by the Google robot. If the author himself sets the indicator with a certain frequency of visits in the HTML document heading, the Google recognizes is as optional and is usually ignored in favour of the PageRank rate factor which is the main factor of the page index frequency. Therefore, if a particular web page was changed between Google robot visits, it would not be indexed and thus it would not be read with the "cache:" command. Object that work particularly well when testing this function are blogs, social network accounts and Internet portals or vortal pages, updated very frequently.

Deleted information or data which had been posted by mistake or during the work on the webpage, or which require deleting at a particular moment, can be restored very easily. These functions are applied both in grey and black open-source intelligence. The negligence of the page administrator can put him at risk for making the unwanted information public.[17]

[16] The basic rules of Google search, comprising, *inter alia*, the George'a Boole logical operators and span operators will be submitted in the text where it is necessary. But the authors gave up on expositing the basics of Google search.

[17] In this context a systematic archiving of web pages in the frame of the Internet Wayback Machine project seems valuable. The project has been operational since 2001. The data base can be found at https://web.archive.org. It contains a significant amount of former versions of web

Searching for information on users

Searching for information on users with some restrictions only can be regarded as GH. In this case it seems a bit excessive. For this purpose advanced operators are used which make search results more detailed and correct.[18] The "@" (at) operator is used for searching users in social networks - Twitter, Facebook, Instagram are indexed this way. For example, the application of this operator is the following:

inurl:twitter @mider

The query in Twitter finds the user "mider".

Let us assume that we know the place of work of the person we search for, (for example the Institute of Political Science, University of Warsaw) and their name. Instead of tedious searching through the institution's web page, one can submit a proper query based on the e-mail address and assuming that at least the name of the searched person should be placed in the name of the address. Example:

site:inp.uw.edu.pl "*mider@uw.edu.pl"

One can also use a less sophisticated method and submit a query just on e-mail addresses in the way shown below, hoping for good luck and lack of administrator's professionalism.

emaile.xlsx
filetype:xls +emaile

One can also try to get the e-mail addresses from the web page with the following query:

site:inp.uw.edu.pl intext:e-mail

The queries would then search for the word "e-mail" on the webpage of the Institute of Political Science, University of Warsaw. Searching for e-mail addresses is of limited use and mostly requires some information on the user

pages, possible to be searched by numerous detailed queries. The "cache:" operator indexes only the previous version of a webpage and the above-mentioned project indexes also earlier versions (although not systematically).

[18] The obvious techniques, direct queries based on names, aliases, known or supposed nicknames and identifications were omitted. The authors also refrained from poresenting the operators modifying the queries, including G. Boole logical operators. However, less known methods as well as the ones with significant restrictions have been presented.

beforehand. And searching for telephone numbers via Google "phonebook:" is limited to subscribers in the USA only. For example:

phonebook:John Doe New York NY

Searching for information on users is also possible via Google "image search". It enables the finding of identical or similar photographs (the set of shapes and colours) in websites indexed by the Google.

Searching for e-mail addresses in Google is rather tedious, comparing to Maltego or The Harvester applications. Nevertheless, it is possible.

Searching for substantial information

Google introduced a couple of useful facilitations, *inter alia* "related:" operator, which displays a list of "similar" websites to the required one. The similarity is based on the functional links, not on the logical or substantial ones.

related:www.inp.uw.edu.pl

In this case the pages of other scientific centres are displayed. This operator works as "Similar pages" button in the Google advanced search. In the same way the "info:" query makes it possible to display information on the particular webpage. This is the load of information facilitated by the authors of a particular webpage, introduced in the website heading (<head>), in the description meta-tags (<meta name="Description"…"). Example:

info:inp.uw.edu.pl

The "define:" query is quite useful, particularly in the scientific work. It makes it possible to obtain the definitions of words from such sources like encyclopaedias and online dictionaries. The example of its application:

define:political participation

A universal operator is tilde ("~"). It allows to search for related words or similar words (synonyms):

~nauki ~polityczne

The above query would display both websites with words "nauki" and "polityczne", as well as websites with the synonym "politologia". The "link:" operator modifying the query limits the search range to links given on a certain

webpage. Using it we can check whether anybody else introduces links to the webpage or the file of our interest. Example:

link:www.inp.uw.edu.pl

However, the operator we are talking about is defective. It does not display all the results and expands search criteria.

Hashtags are a kind of tags that enable grouping information preceded by the "#" sign. At present, they are mainly used on Instagram, but also on Facebook, Google+, Tumblr and Wykop. Google makes it possible to search in many networks simultaneously, or only in recommended networks. Example of a typical query to any search engine:

#polityka

The "around(n)" operator enables the search for two words remaining in a certain distance from each other. Example:

google around(4) hacking

The effect of the above query is displaying the websites which contain those two words ("google" and "hacking"), but they are separated from each other by four other words.

Searching according to file types is extremely useful, because Google indexes materials also according to formats, in which they were written. The "filetype:" operator is used for this purpose. Currently a very wide range of files is administered.[19]

Among all available search engines Google provides the most complex range of operators for white open-source intelligence. Other search engines offer slightly fewer useful operators, and thus - they do not guarantee such a precise search. For the purposes of the white open-source intelligence, however, we can recommend such tools as Maltego[20], Oryon OSINT Browser[21] and FOCA program. They enable automatic data search and they do not require the knowledge of operators. The mechanism of the program is very simple: with the use of a proper query directed to the Google, Bing and Exalead we find documents published by the institution of our interest and it analyses meta-data from these documents. A potential information resource for the program is every file with any extension, for example doc, pdf, ppt, odt, xls or JPG. It is a service delivered by FOCA, the most practical and the easiest service to get at the same time, provided by the FOCA, which can be classified as the white open source-intelligence.

[19] The updated list of file types is accessible at: https://support.google.com/webmasters/answer/35287?hl=en [acces: 26 I 2018].

[20] Maltego, https://www.paterva.com/web7/buy/maltego-clients/maltego-ce.php [access: 26 I 2018].

[21] Oryon OSINT Browser, https://sourceforge.net/projects/oryon-osint-browser/ [access: 26 I 2018].

It is therefore significant to take care of the proper "meta-data cleaning" before making the files available. In some web manuals there are at least several ways provided on how to get rid of unwanted data. It is not possible to show *a priori* the best way because it depends on the users' individual preferences. The authors of this report advise to write the file in the format that does not store metadata first and then to make the file available. So it is desirable – for example, to convert a doc. document[22] into txt. or rtf. formats and pictures written as JPG files facilitate in PNG format.

Internet is also a source of numerous free cleaning programs for metadata, mainly as far as pictures are concerned. ExifCleaner[23] can be treated as proven and desirable. In the case of text files manual performing this activity is highly recommended. The process depends on the package we have at our disposal. Also checking the preferences and settings for applications or a device we use can be a good way to automatically limit the number of the metadata we store.

2. Google Hacking as grey open-source intelligence

The term grey open-source intelligence is used to denote the access to the content left unconsciously, the reconstruction of the website structure or of the intranet structure and the access to www. server configuration parameters. The activity is unethical, although legal.

Information left (unconsciously) by the authors and websites owners

In the case of grey open-source intelligence Google search indicates the access to resources which should not be visible for outside users. Such resources are left unconsciously (for example – the old contents left by the web administrators, internal documents and company materials remaining on a server), or left for convenience and use of the same people who left them, for example music files or film files, private photographs). The search for such content can be made with the Google in many ways. The easiest way is guessing. If there are – for example, 5.jpg, 8.jpg and 9.jpg files in a certain catalogue, you can predict that there are – for example, 1–4 files, 6–7 files and files over 9 as well. Therefore, we can potentially access the materials, which were not supposed to be made available by the person who had placed them there. Another way is searching through websites for certain types of content. We can search for music files, photos, films and books

[22] In case of ms Word files Microsoft prepared a detailed manual on the minimization of meta-data, which gives the users a full feedback on how to get rid of the potential risk manually. Cf. *Jak zminimalizować ilość metadanych w programie Word 2003*, https://support.microsoft.com/pl-pl/help/825576/how-to-minimize-metadata-in-word-2003 [access: 5 VI 2018].

[23] The basic form of metadata are EXIF tags, which allow to read, *inter alia*, the date and time of taking the picture, camera settings and GPS position. Other forms of record are IPCT or XMP. The ExifCleaner is available in several versions and can be found on the distributor's page. It gives the possibility of cleaning photographs from the above data by only one click. Cf. *Zanim wgrasz wakacyjne zdjęcie do sieci…*, https://niebezpiecznik.pl/post/zanim-wgrasz-wakacyjne-zdjecia-do-sieci/ [access: 5 VI 2018].

(e-books, audio books). Quite often these are the files the user made available unconsciously (for example - music on ftp server only for their own use). We can get them in two ways: using "filetype:" operator or "inurl:" operator. For example:

filetype:doc site:edu.pl
site:www.inp.uw.edu.pl filetype:pdf
site: www.inp.uw.edu.pl intitle:index.of.mp3

We can also search for program files using the search query:

filetype:iso

Information on the web page structure

The activity is legal, although it seems unethical – we look at a certain webpage from the insight, disclose its whole construction, which does not comply with its authors' intentions. We can do it easily using only "site:" operator. Let us analyze the following phrase:

site: www.inp.uw.edu.pl inp

We initiate the search for the word "inp" in the domain www.inp.uw.edu.pl. Each site from this domain (Google searches both in text, in titles and in the site header) contains this word. So we obtain the structure of all sites of this particular domain. A more precise result (although it is not always possible to obtain) after the structure of catalogues becomes accessible, we obtain by using the following query:

site:uw.edu.pl intitle:index.of "parent directory"

It reveals the least secured sub-domains of uw.edu.pl, sometimes with the possibility to search the whole catalogue tree and download all files. So, naturally, such query is not applicable to all domains as they are secured or operate under the control of some other www. server.

www. server configuration parameters

Obtaining configuration parameters of the www. servers is on the verge of grey and black open-source intelligence. It can be a prologue to data collection on an attack or only a download of data referring to the kind and quality of services the device provides. In order to obtain the name of the server, its version and other parameters (for example ports) the following query is made:

site:uw.edu.pl intitle:index.of server.at

Each www. server has its unique phrases, for example Microsoft's Internet Information Service (IIS):

intitle:welcome.to intitle:internet IIS

Recognition of the www. server depends only on inventiveness. One can, for example, try to do it through a technical specification query, manual or the so-called help pages. This goal can be reached by the following query:

site:uw.edu.pl inurl:manual apache directives modules (Apache)

The access can be more advanced, for example thanks to a file with SQL errors:

"#mysql dump" filetype:SQL

Errors in the SQL database could - *inter alia*, provide the information on structure and content of the datasbase away. In turn, the whole webpage, its original and (or) its updated versions can be accessed with the following query:

site:uw.edu.pl inurl:backup
site:uw.edu.pl inurl:backup intitle:index.of inurl:admin

Currently using the above phrases quite rarely gives the expected results, because they are blocked by the security issues-aware users. Such activities should be situated somewhere on the verge of grey and black open-source intelligence.

Also, with FOCA program we can find content from this category. One of the first activities the program starts with after the new project begins is the analysis of the domain's structure and of all other sub-domains attached to the servers of the particular institution. Such information can be found in a dialogue box, in the Network tab.

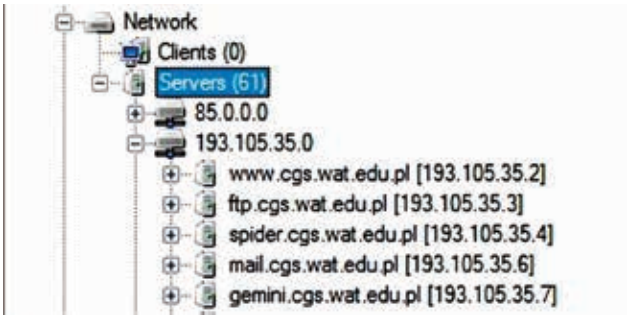

Diagram: Screenshot of the FOCA program.

Source: self-study.

This way the potential "aggressor" can intercept the contents left by the web administrators, internal documents and materials of the company left even on a hidden server.

3. Google Hacking as black open-source intelligence

Black open-source intelligence means illegal activities, regarded also as unethical. These are: obtaining information secured previously from interception, access to personal sensitive data (*inter alia* users names, passwords, identification numbers) and the access to device configuration parameters made to take control of them.

Information intentionally secured from interception

These are mostly the information from web pages charging for allowing to access them (for example – "The Boston Globe" and "The New York Times" subscription, in Poland "Rzeczpospolita" daily subscription). One of the ways to obtain a free access to payable articles has become the "pretending" to be a Google Internet robot (Googlebot, Google Spider, Google net spider) by a user. It is the basic mechanism for indexation of web pages for Google engine. Because of their function, net programs recognized as Google can do more than ordinary users – authors of payable contents want to index them to let the users know about their existence but they do not want to make them available for free. This way the Google robot has (at least up to now) the access to such kind of pages. It is possible to imitate Googlebot by changes in one's own search engine identity tag. It can be done directly – by changes of configuration entries in Windows record editor, or indirectly – by changes of configuration parameters of the search engine, or in the easiest possible way – by downloading proper plugs-in (in the case of Chrome it is User-Agent Switcher).[24] There are also automatic ways of access to this content, *inter alia,* Block Referer and User Agent Switcher, used together. This possibility of access to restricted (payable) content has been noticed and eliminated by major information distributors (in Poland, *inter alia*, Archiwum Rzeczpospolitej and Archiwum Wyborczej – archival resources of the two major daily newspapers in Poland). Nevertheless, the race is constant; following the elimination of this channel the Paywall-Pass plug-in emerged and then - more or less successful attempts of URL manipulation of archival articles (for example in Archiwum Rzeczpospolitej the end of the link had to be changed from "?template=restricted" to "?template=%00" and perform the so called "null byte injection". However, when this solution was blocked, the new way of obtaining access to the contents involved adding to URL – after html, the phrases "#ap-1" and "html?templateOld=00%&templateOld=0%%&". Currently those options are

[24] Whether you can forge your own identity, it can be checked at: https://www.whoishostingthis.com/tools/user-agent/ [access: 26 I 2018]. It is worth noting that the change of our search engine identity into a net robot makes it impossible to use some services, for example, the e-mail.

also blocked although there are certainly other solutions. However, these techniques go beyond the classic understanding of GH. With the Google search engine the protected information can be obtained through the use of simple queries using servers' administrators errors. Examples are the following:

budżet filetype:xls
inurl:gov filetype:xls "restricted"
allintitle:sensitive filetype:doc
allintitle:restricted filetype:mail

Sensitive personal data

User data, their logins, nicknames and passwords make the first group of sensitive data. The obtaining of such data enables the identity theft. This kind of information can be acquired by Google search (and not only) in many ways. This type of activities is obviously unethical and illegal. The success in searching this kind of information depends on the knowledge of structures of operational systems and the structures of individual programmes. Generally speaking, the activities we are talking about rely on making queries with the use of potential phrases and elements co-existing with users' names and passwords. The example of a query which enables to find a file with user names:

allintext:username filetype:log

Here are a few basic examples of queries finding user passwords:

hasla.xlsx
passwords.xls
intitle:password
filetype:log inurl:password
intitle:"index of password"
inurl:passwd filetype:txt

In the above-mentioned examples you can obtain passwords left by irresponsible administrators. Different versions were given to show numerous options of these activities and to show how significant guessing is in the course of such search. However, we should not count on storing passwords by administrators in this way – at present (2018) Google shows only a few or a dozen results upon such queries. The example of a more advanced query:

"index of /" +password.txt

Passwords written with the open text are obtained through the algorithm. Some more advanced queries, from which the first one regards the administrator's passwords and the second searches any passwords and login information, are given below:

http://admin:*@www
filetype:bak inurl:"htaccess|passwd|shadow|htusers"

On the other hand, obtaining files of passwords in an indirect way is, as follows:

inurl:config.txt

After the above-mentioned phrase is entered, one is given the access to config. txt files, which potentially contain information on the server's configuration, such as encrypted administrators' passwords or data enabling the access to databases. The address of the login page can be obtained in the easier way. In the first case, this is the login page of the administrator, and in the second case - the webpage of ordinary users:

inurl:adminlogin
intitle:login

Not just guessing, but also the knowledge of configuration parameters of the programmes is important in the process of obtaining passwords. Let us look closer into an example of FTP (*File Transfer Protocol*) client. At present, this kind of transfer when data id updated is used less frequently, and that is the reason why we provide such an example. Passwords in one of the most popular programme of this kind (Total Commander) are encrypted with an open text and are stored in the .ini file. The query sent to the Google to obtain the passwords is the following:

inurl:wcx_ftp filetype:ini

For obtaining passwords and other data necessary for authorisation the analysis of the technical specification of the programme and the method and place of storing the data in the structure of a particular application or the system structure are of key significance. GH is meant for files saved with open text as well as encrypted files. The secured files can be decrypted with one of the numerous decrypting programmes (popular algorithm DES, i.e. symmetric block cipher, is broken by the John the Ripper software).

The next sensitive data type are different kinds of identification numbers, for example credit cards numbers, insurance numbers, or ID numbers (in Poland the social security number called PESEL). The same way of searching serial numbers of programs of games is also possible. The simplest structure serving this goal is the following:

index of/credit-card

The key element for mass web searching is ".." scope solution operator and (optionally) "numrange:" operator:

PESEL+74010100000..76123199999
numrange:74010100000..76123199999

The above-mentioned query allows to search for ID numbers from the time period between 1 January 1974 and 31 December 1976. Searching for credit cards numbers is the same – the knowledge of the pattern of the numbering, typical for a particular provider is enough.

+MasterCard 5500000000000004..5599999999999999
numrange:370000000000002..379999999999999
numrange:586824160825533338..89999999999999999[25]

Searching for information on other financial services, for example on accounts number, is the same. At present, because of multiple violations, this service is blocked by Google but some other search engines (Yandex) facilitate this kind of search.

Advanced GH makes it also possible to obtain other sensitive data, for example from online stores. Nowadays they are quite well secured, though. Most search ideas are several years old and only one comes from 2016.[26] They are based on the knowledge of the store's software. Example:

intext:"Dumping data for table `orders`"

The above phrase facilitates searching files of SQL drop, which contain potentially personal data.

Configuration parameters of programmes and devices

Recognition of configuration parameters of programmes and devices means searching for security loopholes with the aim of using them or taking a direct control over the programmes (devices) using original (default) configuration parameters. Also, the software of servers of different services in the web is prone to the GD attacks. The target of the attack can be operational systems, and the basic query can be constructed in the following way:

ip:212.85.108.185 index of /admin

[25] As first the so-called test numbers of MasterCard, American Express and Maestro International were used.

[26] J. Long, *Google Hacking Database*, https://www.exploit-db.com/google-hacking-database/?action=search&ghdb_search_cat_id=10&ghdb_search_text= [access: 26 I 2018].

ip:212.85.108.185 index of/root
ip:212.85.108.185 allinurl:winnt/system32/

"Inurl:" or "allinurl:" orders enable the search for machines with certain (known) security loopholes, especially configuration elements, while the "ip:" operator indicates a particular server as the target of an attack. In the last example given above we check whether there are system catalogues under the given address accessible in the web (the administrator's error). If this happens, then the cmd.exe file enables the taking of control over the server. Searching for program configuration parameters can also be executed as follows:

inurl:config.txt
inurl:admin
filetype:cfg
inurl:server.cfg rcon password
allinurl:/provmsg.php

There is a possibility to search for old versions of scripts administering such services as internet fora, blog platforms, online stores and other services, including classic web pages. Finding a new loophole is a starting point to search for old versions of scripts. The example could be the following:

intext:"Powered by: vBulletin Version 3.7.4"

It is also possible to obtain serial numbers of programmes, including operational systems. The example of GH query aimed at finding the Windows XP Pro registration key:

"Windows XP Professional" 94FBR

GH attacks do not have to focus on the program configuration parameters. The subject of an attack can also be the disclosed security loopholes. The query on the server/parser errors is typical. They are usually stored in text files and they are sometimes indexed by the Google search engine. The examples of typical phrases:

filetype:txt intext:"Access denied for user"
filetype:txt intext:"Error Message"

GH enables also the access to different specialized Web devices. The most frequent and described activity in this area is getting access to web cameras. A significant number of cameras is installed without configuration – without any passwords, or with logins and original passwords. This way the potential attack is possible. The construction

of the query to use is dependent on the manufacturer and on the type of camera. Paths to cameras login pages and default greeting information can be searched via Google. Knowing the information and entering it into the search, we can find numerous cameras indexed with these data. Some part of devices enables the intruder not only preview but also the manipulation option. Using operators like "ip:", "site:" or "inurl:" facilitates the narrowing of the search range to a certain subject or geographical area. We obtain the access to private monitoring cameras around the house, in children's rooms, animals facilities, to monitoring in work places – offices building and their neighbourhood, CCTV cameras, etc. Sets of logins and original passwords to cameras are commonly published.[27] There are also generators of passwords to cameras.[28] Access to cameras is connected to a sui generis illegal services' market; there are web fora for those who seek and get access to such visual content (inter alia anon-ib, 4chan, to a lesser extent also overchan and torchan[29]). The examples of queries enabling access to web cameras:

"/home/homeJ.html"
inurl:"CgiStart?page="
intitle:"Biromsoft WebCam" -4.0 -serial -ask -crack -software -a -the -build -download -v4 -3.01 -numrange:1-10000

Other devices, like printers, can be the subjects of attack as well:

inurl:hp/device/this.LCDispatcher
intitle:"Dell Laser Printer" ews

switches:

inurl:"level/15/exec/-/show"

routers and other devices:

intitle:"Welcome to ZyXEL" -zyxel.com

If we want to obtain the name of the user who constructed a particular file left on the server of a particular organization, we can use the FOCA programme. After

[27] *Default login and password for DVR NVR IP*, in: kizewski.eu/it/hasla-domyslne-default-login-and-password-for-dvr-nvr-ip/ [access: 26 I 2018].

[28] Cf. for example in: http://www.cctvforum.com/viewtopic.php?f=19&t=39846&sid=42bdd50a426bea9296f1a2e78f09c226 [access: 26 I 2018].

[29] *Oto jak wykrada się nagie zdjęcia gwiazd. I to nie tylko z telefonów*, https://niebezpiecznik.pl/post/oto-jak-wykrada-sie-nagie-zdjecia-gwiazd-i-to-nie-tylko-z-telefonow/, 3 IX 2014 [access: 26 I 2018].

the domain search and choosing "Download All" option there is a possibility to download all the found files and using "Extract/Analyze All Meta-data" we can check whether we managed to find anything interesting. Searching "Meta-data Summary" we obtained information on the user names and, if we are lucky enough, on the folders' paths, printers, software, e-mail addresses, operational system, servers, maybe even on the passwords left by unaware users.

* * *

Threats from GH are generated because of ignorance and negligence of the owners and users of various programmes, servers and other web devices, so the rules of self-defence and protecting data cause no difficulties at all. Information-related threats are connected to the Google robot activities and the robots of other search engines, so the basic rule of addressing risks makes a reference to limiting or blocking the robot access to data in the net. A total blockade of the www server against search by net spiders is done by a simple text file "robots.txt" in the main catalogue of the web page. This file should contain two commands:

User-agent: *
Disallow: /

Placing the "*" sign eliminates all search engines, although only one engine can be indicated, if we enter its name. On the other hand, the "Disallow:" parameter determines which elements of the web structure are to be eliminated.[30] Barriers for net spiders can be introduced also on individual pages – both typical web sites, blogs as well as configuration pages. In the HTML header filed they should be accompanied by one of the following phrases:

<meta name="Robots" content=""none" />
<meta name="Robots" content="noindex, nofollow" />

If we enter such a record on the main page, neither the second-rate pages nor the main page would be indexed by the Google robot. This way they will be safe from the GH. The phrase can also be entered on the pages that we want to be omitted by the Googlebot. However, it is a solution the security of which is based on the gentlemen's agreement. Although Google and other net spiders respect the above-mentioned restrictions, there are net robots "hunting" for such phrases to obtain the data, which are supposed not to be indexed. From the group of more advanced security systems the CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*) system is worth suggesting. This is the security system

[30] Detailed instructions on the Google web page: http://www.google.com/remove.html [access: 26 I 2018].

that allows access to the page content only to humans, not virtual formations, which download contents automatically. This solution hat, however, some disadvantages. It is awkward to the users. The potentials of self-defence tend to increase along with the growing competence and awareness on GH techniques of the administrator. A simple defensive method to limit the Google Dorks can be - for example, sign encoding in administrative files by ASCII codes, making it harder (although not impossible) to use GH.

The IT tycoon – McAfee – recommends the following six protocols for administrators to follow in order to avoid GH threats:

1) systematic update of operational systems, services and applications;
2) implementation and maintenance of anti-hacking systems;
3) awareness of robots and search engines routines, knowledge on GH potentially compromising contents and ways of verification of such processes;
4) removing sensitive contents from public locations;
5) consistent division on publicly accessible contents and private non-accessible contents and eventually - blocking access contents for outside users;
6) frequent penetration tests (pentests).[31]

Practice number 6 – penetration tests, seems to be particularly important, because such tests determine univocally the vulnerability level of the web page or the server, including GH. There are special tools for pentests in the area of GH. One of them is Site Digger v3.0 (Google API licence due is not required)[32], which enables the automatic testing of Google Hacking Data Base on any chosen web page. There are more such tools, like Wikto scanner[33] or online scanners.[34] They operate in a similar way. There are also aggressive tools which imitate the environment of the web page, its errors, loopholes and vulnerabilities in order to entice the attacker, then obtain some information on him that enable counteraction, for example Google Hack Honeypot.[35] An ordinary, untrained and unaware user has limited possibilities and self-defence tools against *Google Hacking*. First of all he can use GH tools on himself to check if and what sensitive data regarding him are publicly available. It is worth checking regularly such bases as *Have I been pwned?*[36] and *We Leak Info,* to find out whether security of our accounts in the net have been broken and published. The first database is available

[31] C. Woodward, *Go Dork Yourself! Because hackers are already dorking you*, in: https://securingtomorrow.mcafee.com/business/google-dorking/ [access: 26 I 2018].

[32] McAfee, *SiteDigger v3.0 Released 12/01/2009*, https://www.mcafee.com/uk/downloads/free-tools/sitedigger.aspx [access: 26 I 2018].

[33] Sectools.org, *Wikto*, http://sectools.org/tool/wikto/ [access: 26 I 2018].

[34] PentestTools.com, *Google hacking,* https://pentest-tools.com/information-gathering/google-hacking, [access: January 2018].

[35] The Google Hack Honeypot, http://ghh.sourceforge.net/ [access: 26 I 2018].

[36] The word "pwned" comes from the online computer player slang (it originated in an online game called Warcraft. It came from a misspelled word "owned" (letters "p" and "o" are next to each other on the computer keyboard). It basically means "to own" but symbolically it refers to online sphere.

at https://haveibeenpwned.com/ and regards the Web pages, in which our accounts' data were entered (for example e-mail address, logins, passwords, other data) because the pages were poorly secured. We search by entering the e-mail address. Currently (June 2018) the database contains more than 5 billion accounts. A more advanced tool can be found at https://weleakinfo.com. It allows to search information by user name, e-mail address, password and its hash, IP, name and phone number. However, the search service is not the only one here. The accounts, which leaked data, can be bought in the network. One-day access is charged only 2 USD.

A more thorough description of preventing the info leakage by GH/GD exceeds the capacity of this article. Nevertheless, it is worth pointing out that professional data protection is possible only after a professional audit made by the so-called pentesters or bughunters. Penetration tests are recently becoming an important sub-branch of practical providing of IT and information security. They are made multidimensionaly, through testing IT security as well as technical and physical security of devices and facilities. Special IT tools have been developed (for example operational systems Kali Linux or Metasploit) and standards of their implementation have been implemented. They are used by both private business and public sectors. However, they are relatively expensive.

The scale and the risk is presented by the FBI information published, which show several tens of thousands of such incidents annually.[37] It is also estimated that in more than three quarters of web pages security loopholes can be found and one in ten web pages has serious loopholes.[38] Google Hacking poses a significant threat to data security. The costs of data collection are much lower than in the case of an attack made with other methods.[39] The competence barrier to break is not too strong – as the author of the term Google Hacking pointed out. The substance of these activities, like in case of hacking activities, is simplicity;[40] they do not require IT qualifications but understanding and learning some commands, as well as some logical and creative thinking.

[37] Quote from: *Roll Call Release. Intelligence for Police, Fire, EMS, and Security Personnel,* 7 July 2014, in: https://info.publicintelligence.net/DHS-FBI-NCTC-GoogleDorking.pdf [access: 26 I 2018].

[38] Quote from: P. Cucu, *How Malicious Websites Infect You in Unexpected Ways. And what you can do to prevent that from happe*ning, „Heimdal Security" of 30 June 2017, https://heimdalsecurity.com/blog/malicious-websites/ [access: 26 I 2018].

[39] M. Laskowski, *Using Google search engine to get unauthorized access to private data*, „Actual Problems of Economics" 2012, No. 132, pp. 381–386.

[40] M. Kassner, *Google hacking: It's all about the dorks*, „IT Security", https://www.techrepublic.com/blog/it-security/google-hacking-its-all-about-the-dorks/ [access: 26 I 2018].

Abstract

The article analyzes the potential of obtaining internet-based information techniques referring to as Google Hacking (GH), that is, the forwarding of Google search queries revealing data not available directly or whose acquisition is unauthorized for ethical reasons, legal reasons or both. Techniques of obtaining information by GH method have been divided into three groups. The first method of obtaining data that does not raise ethical and legal concerns is referred to as open-source, white intelligence, including the search for deleted and archived pages, search for some information about users and other substantive information. The second group of techniques (grey intelligence) – raising ethical concerns – included the acquisition of information left (unconsciously) by the authors and owners of websites, information about the structure of websites and the configuration parameters of www servers. The last group of techniques is the so-called black intelligence – illegal and mostly unethical acts. There subject of analysis was the potential of obtaining secured information, of sensitive personal data and configuration parameters of programs and devices. The text is complemented by the analysis of the possibilities of obtaining information through the FOCA (*Fingerprinting Organizations with Collected Archives*) program, used to automate GH queries, *metadata harvesting* oriented, i.e. mass mining and analysis of meta-data contained in online documents.

Keywords: Google Hacking, FOCA, metadata harvesting, browser**.**