

Metadata

7

INFORMATION IN THIS CHAPTER

- Metadata
- Impact
- Metadata Extraction
- Data Leakage Protection (DLP)

INTRODUCTION

In the last few chapters we have learned extensively about how to find information online. We learned about different platforms, different techniques to better utilize these platforms, and also tools which can automate the process of data extraction. In this chapter we will deal with a special kind of data, which is quite interesting but usually gets ignored, the metadata.

Earlier metadata was a term mostly talked about in the field of information science domain only, but with the recent news circulation stating that National Security Agency has been snooping metadata related to phone records of its citizens, it is becoming a household name. Though still many people don't understand exactly what metadata is and how it can be used against them, let alone how to safeguard themselves from an information security point of view.

The very basic definition of metadata is that it's "data about data," but sometimes it's a bit confusing. So for the understanding purpose we can say that metadata is something which describes the content somehow but is not the part of the content itself. For example in a video file the length of the video can be its metadata as it describes how long the video will play, but it is not the part of the video itself. Similarly for an image file, the make of the camera used to click that picture can be its metadata or the date when the picture is taken as it tells us something related to the picture, but is not actually the content of the picture. We all have encountered this kind of data related to different files at some point of time. Metadata can be anything, the name of the creator of the content, time of creation, reason of creation, copyright information, etc.

The creation of metadata actually started long ago in libraries, when people had information in the form of scrolls but no way to categorize them and find them

quickly when needed. Today in the digital age we still use metadata to categorize files, search them, interconnect them, and much more. Most of the files that reside in our computer systems have some kind of metadata. It is also one of the key components needed for the creation of the semantic web.

Metadata is very helpful in managing and organizing files and hence is used extensively nowadays. Most of the times we don't even make a distinction between the actual content and its metadata. It is usually added to the file by the underlying software which is used to create the file. For a picture it can be the camera that was used to click it, for a doc file it can be the operating system used, for an audio file it can be the recording device. Usually it is harmless as it does not reveal any data which can be sensitive from information security perspective, or is it? We will see soon in the following portion of this chapter.

There are huge number of places where metadata is used, from the files in our systems to the websites on the internet. In this chapter we will mainly focus on extracting metadata from places which are critical from information security view point.

METADATA EXTRACTION TOOLS

Let's discuss about some of the tool which can be used for the metadata extraction.

JEFFREY'S EXIF VIEWER

Exif (exchangeable image file format) is basically a standard used by devices which handle images and audio files, such as video recorder, smartphone cameras etc., It contains data like the image resolution, the camera used, color type, compression etc. Most of the smartphones today contain a camera, a GPS (global positioning system) device, and internet connectivity. In many of the smartphones when we click a picture it automatically tracks our geolocation using the GPS device and embeds that information into the picture just clicked. We being active on social networks share these pictures with the whole world.

Jeffrey's Exif Viewer is an online application (<http://regex.info/exif.cgi>) which allows us to see this Exif data present in any image file. We can simply upload it from our machine or provide the URL for the file. If an image contains the geolocations, it will be presented in the form of coordinates. Exif Viewer is based on the Exif Tool by Phil Harvey, which can be downloaded from <http://www.sno.phy.queensu.ca/~phil/exiftool/>. It not only allows to read the Exif data but also write it to the files. Exif Tool supports a huge list of different formats like XMP, GFIF, ID3, etc., which are also listed on the page.

Basic Image Information

Target file: WP_20140922_10_40_53_Pro.jpg

Camera:	Nokia Lumia 630
Exposure:	Auto exposure, 1/8 sec, f/2.4, ISO 1600
Flash:	Off, Did not fire
Date:	September 22, 2014 10:40:53AM (timezone not specified) (12 hours, 48 minutes, 17 seconds ago, assuming image timezone of 5½ hours ahead of GMT)
Location:	Latitude/longitude: 28° 35' 30.7" North, 77° 22' 17.5" East (28.591863, 77.371538) Location guessed from coordinates: <i>E-88, E-block, Sector 52, New Okhla Industrial Development Area, Uttar Pradesh 201307, India</i> Map via embedded coordinates at: Google , Yahoo , WikiMapia , OpenStreetMap , Bing (also see the Google Maps pane below) Altitude: 176 meters (577 feet) Timezone guess from earthtools.org: 5½ hours ahead of GMT
File:	916 × 1,632 JPEG (1.5 megapixels)
Color Encoding:	WARNING: Color space tagged as sRGB, without an embedded color profile. Windows and Mac browsers and apps treat the colors randomly. Images for the web are most widely viewable when in the sRGB color space and with an embedded color profile. See my Introduction to Digital-Image Color Spaces for more information.

FIGURE 7.1

Jeffrey's Exif Viewer.

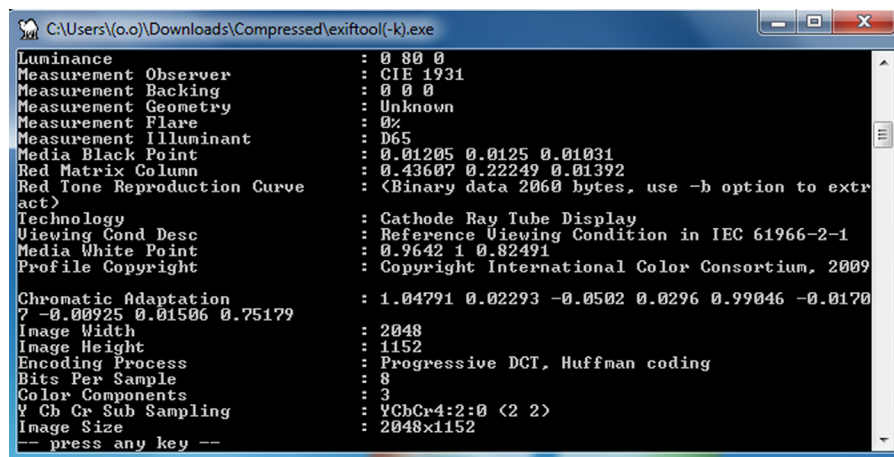


FIGURE 7.2

Exif Tool interface.

Using the geolocation in the images we share, anyone can easily track where we were exactly at the time of clicking it. This can be misused by people with ill intentions or stalkers. So we should be careful if we want to just share our pictures or locations too.

EXIF SEARCH

We just discussed about the Exif and its power to geolocate the content. There is a dedicated search engine which allows us to search through geotagged images, it's called Exif Search (<http://www.exif-search.com/>).

This search engine provides data about the images and pictures from all over the internet. It contains a huge number of searchable Exif images from different mobile devices. Being totally different from traditional image search engines, which tend to just provide us the image as a result, Exif also provides the metadata.

When we search in Exif Search, it searches the image and its information in its own database and provides us the result. Currently it has more than 100 million images with metadata and it's constantly updating its database.

This search engine provides user the freedom to search an image based on location, date, and device type. It also allows us to sort the data based on these date location or device type. Another unique feature of this search engine is that it allows us to force the search engine to fetch us result for only images that contains GPS data. There is a small check box available just below the search bar which does the work for us.

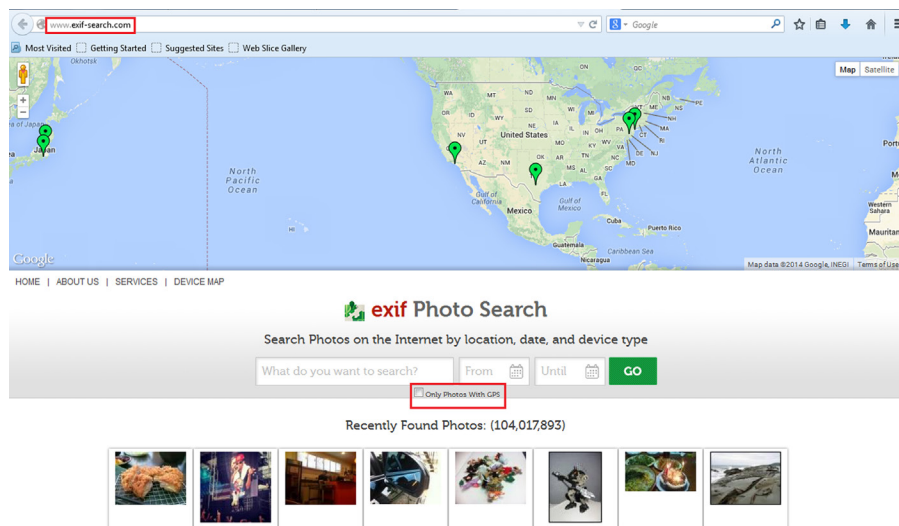
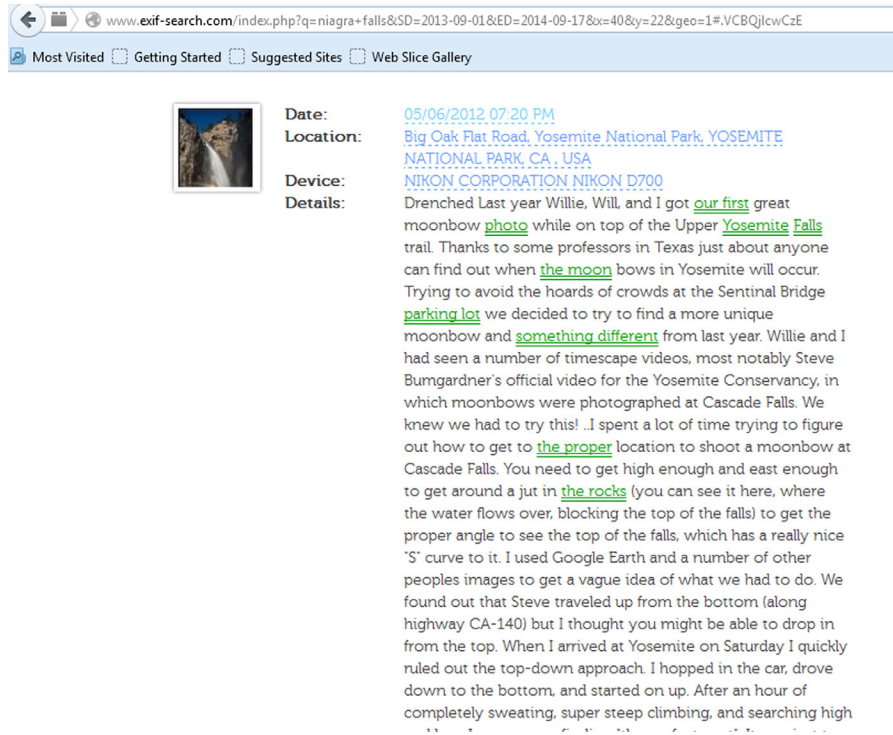


FIGURE 7.3

Exif-search.com interface.

It also supports a huge number of devices. The list can be found <http://www.exif-search.com/devices.php>, some of them are Canon, Nikon, Apple and Fujifilm etc.



www.exif-search.com/index.php?q=niagra-falls&SD=2013-09-01&ED=2014-09-17&cx=40&cy=22&geo=1#.VCBQJlcwCzE

Most Visited Getting Started Suggested Sites Web Slice Gallery


 **Date:** [05/06/2012 07:20 PM](#)
Location: [Big Oak Flat Road, Yosemite National Park, YOSEMITE NATIONAL PARK, CA, USA](#)
Device: [NIKON CORPORATION NIKON D700](#)
Details: Drenched Last year Willie, Will, and I got [our first](#) great moonbow [photo](#) while on top of the Upper [Yosemite Falls](#) trail. Thanks to some professors in Texas just about anyone can find out when [the moon](#) bows in Yosemite will occur. Trying to avoid the hoards of crowds at the Sentinal Bridge [parking lot](#) we decided to try to find a more unique moonbow and [something different](#) from last year. Willie and I had seen a number of timescape videos, most notably Steve Bumgardner's official video for the Yosemite Conservancy, in which moonbows were photographed at Cascade Falls. We knew we had to try this! .I spent a lot of time trying to figure out how to get to [the proper](#) location to shoot a moonbow at Cascade Falls. You need to get high enough and east enough to get around a jut in [the rocks](#) (you can see it here, where the water flows over, blocking the top of the falls) to get the proper angle to see the top of the falls, which has a really nice "S" curve to it. I used Google Earth and a number of other peoples images to get a vague idea of what we had to do. We found out that Steve traveled up from the bottom (along highway CA-140) but I thought you might be able to drop in from the top. When I arrived at Yosemite on Saturday I quickly ruled out the top-down approach. I hopped in the car, drove down to the bottom, and started on up. After an hour of completely sweating, super steep climbing, and searching high

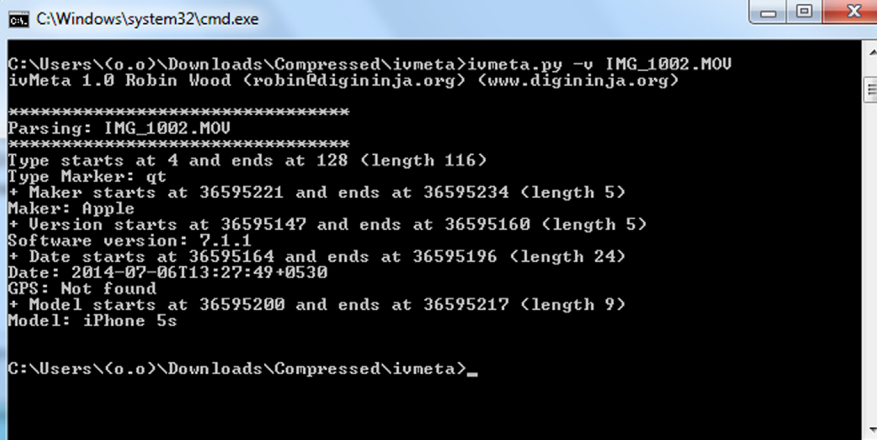
FIGURE 7.4

Exif-search.com sample search result.

ivMeta

Similar to images, video files can also contain GPS coordinates in their metadata. ivMeta is a tool created by Robin Wood (<http://digi.ninja/projects/ivmeta.php>) which allows us to extract data such as software version, date, GPS coordinates, model number from iPhone videos. iPhone is one of the most popular smartphone available and has a huge fan base. With more than a million users, their activity to show the uniqueness of the iPhone standard makes them more vulnerable to metadata extraction. No doubt on the camera quality of the devices and the unique apps to make the pictures and videos look more trendy, iPhone users upload lots of such data content everyday in different social networking sites. Though there is an option available on the device to deactivate geotagging, the by-default setting and the use of GPS allows to create metadata about any image or video taken. In this case this tool comes handy

to gather all the such information from iPhone videos. This tool is a Python script, so running it requires Python installed (2.7+). It can be very helpful in forensic examination of iPhone videos.



```

C:\Windows\system32\cmd.exe
C:\Users\\Downloads\Compressed\ivmeta>ivmeta.py -v IMG_1002.MOU
ivMeta 1.0 Robin Wood <robin@digininja.org> <www.digininja.org>
*****
Parsing: IMG_1002.MOU
*****
Type starts at 4 and ends at 128 <length 116>
Type Marker: qt
+ Maker starts at 36595221 and ends at 36595234 <length 5>
Maker: Apple
+ Version starts at 36595147 and ends at 36595160 <length 5>
Software version: 7.1.1
+ Date starts at 36595164 and ends at 36595196 <length 24>
Date: 2014-07-06T13:27:49+0530
GPS: Not found
+ Model starts at 36595200 and ends at 36595217 <length 9>
Model: iPhone 5s

C:\Users\\Downloads\Compressed\ivmeta>_

```

FIGURE 7.5

ivMeta in action.

HACHOIR-METADATA

Hachoir-metadata is based on hachoir Python library. This is one of the hachoir project used for metadata extraction. As its base library is Python, this tool is also a Python tool which can be used to extract metadata from not only image, audio and video but also archives. This supports more than 30 different formats to extract metadata that is also a unique feature on its own.

Some other features that make this tool stand apart from other similar tools are that it supports invalid and truncated files and its ability to avoid duplicate data. Apart from these, it also provides freedom to the users to filter the metadata by setting priorities to the values. This tool is generally available for different Linux versions and can be downloaded from the URL: <https://bitbucket.org/haypo/hachoir/wiki/Install>.

Some of the popular formats supported by this tool are bzip2, gzip, tar, zip, etc., in archive files; bmp, ico, gif, jpeg, png etc., in images. There is also a popular format supported by this tool, PhotoShop Document (PSD). As Adobe PhotoShop is very popular software for image editing in the multimedia industry, supporting this format is definitely a plus for the users who want to extract metadata. In audio it supports mpeg and real audio, where real audio is the default audio format used in Apple devices. In video it supports flv format. This is again definitely a plus because it is widely used in YouTube, one of the largest video sharing site and it also supports mov, the Apple QuickTime movie support that can be well used in Apple device video forensics. The other popular supported formats are exe, which expands the metadata

extraction to another level by allowing all the Microsoft portable executables. It also supports torrent files, which are the easy solution to most of the data sharing requirements. So torrent metadata extraction is definitely one of its unique feature. Who even would thought of extracting metadata from ttf or true type fonts, but yes this tool also supports ttf format. There are many other formats it supports. we can get the details from the following url: <https://bitbucket.org/haypo/hachoir/wiki/hachoir-metadata>.

This hachoir-metadata is basically a command-line tool, and by default it's very verbose. That means running the same without any switches, it provides lots of information.

```
# hachoir-metadata xyz.png
```

We can also run this tool with multiple and different file formats at a time to get the desired result.

```
# hachoir-metadata xyz.png abc.mp3 ppp.flv
```

When we need only mime details we can use

```
# hachoir-metadata --mime xyz.png abc.mp3 ppp.flv
```

When we need little more information other than mime we can use -type switch

```
# hachoir-metadata --type xyz.png abc.mp3 ppp.flv
```

for exploring the tool for other options we can use

```
# hachoir-metadata --help
```

FOCA

On a daily basis we work with a huge number of files such as DOC, PPT, PDF, etc. Sometimes we create them, sometimes edit, and sometimes just read through. Apart from the data we type into these files, metadata is also added to them. To a normal user this data might seem harmless, but actually it can reveal a lot of sensitive information about the system used to create it.

Most of the organizations today have online presence in the form of websites and social profiles. Apart from the web pages, organizations also use different files to share information with general public and these files may contain this metadata. In Chapter 5 we discussed how we can utilize search engines to find the files that are listed on a websites (E.g. In Google: “site:xyzorg.com filetype:pdf”). So once we have listed all these files, we simply need to download them and use a tool which can extract metadata from them.

FOCA is a tool which does this complete process for us. Though FOCA means seal in Spanish, the tool stands for ‘Fingerprinting Organizations with Collected Archives’. It can be downloaded from <https://www.elevenpaths.com/labstools/foca/index.html>. After downloading the zip file, simply extract it and execute the application file inside the bin folder.

To use FOCA we simply need to create a new project, provide it with a name and the domain to scan. Once this is saved as a project file, FOCA allows us to choose

the search engines and the file extensions that we need to search for. After that we can simply start by clicking on the button “Search All.” Once we click on this button FOCA will start a search for the ticked file types on the mentioned domain, using different search engines. Once this search is complete it will display the list of all the documents found, their type, URL, size, etc.

Now we have the list of the documents present on the domain. Next thing we need to do is download the file(s) by right clicking on any one and choosing the option Download/Download All. Once the download is complete the file(s) is/are ready for inspection. So now we need to right click on the file(s) and click on the Extract Metadata option. Once this is complete we can see that under the option Metadata at the right-hand side bar FOCA has listed all the information extracted from the document(s).

This information might contain the username of the system used to create the file, the exact version of the software application used to create it, system path, and much more which can be very helpful for an attacker. Though metadata extraction is not the only functionality provided by FOCA, we can also use it to identify vulnerabilities, perform network analysis, backups search and much more information gathering, the most prevalent functionality.

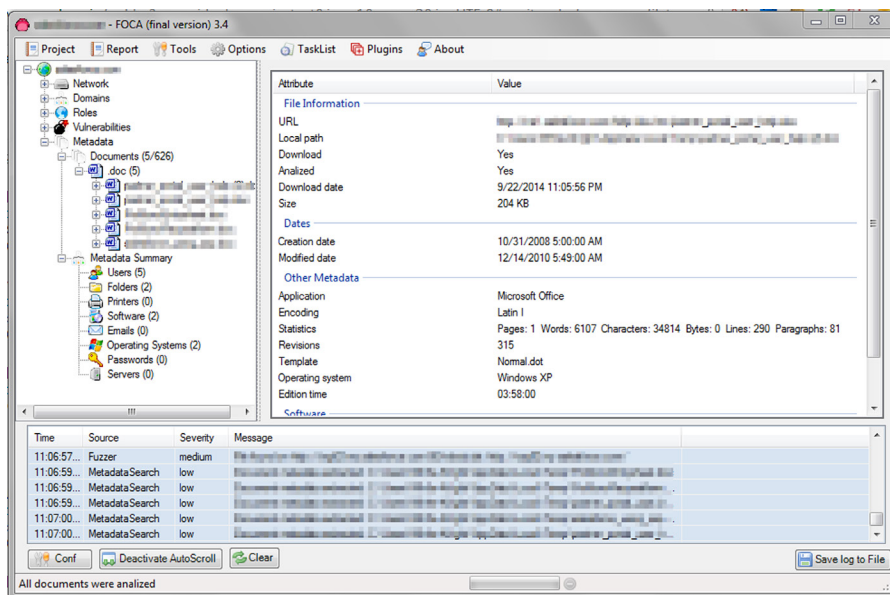


FIGURE 7.6

FOCA result.

METAGOOFIL

Similar to FOCA, Metagoofil is yet another tool to extract metadata from documents which are available online. Metagoofil is basically a Python based command line tool.

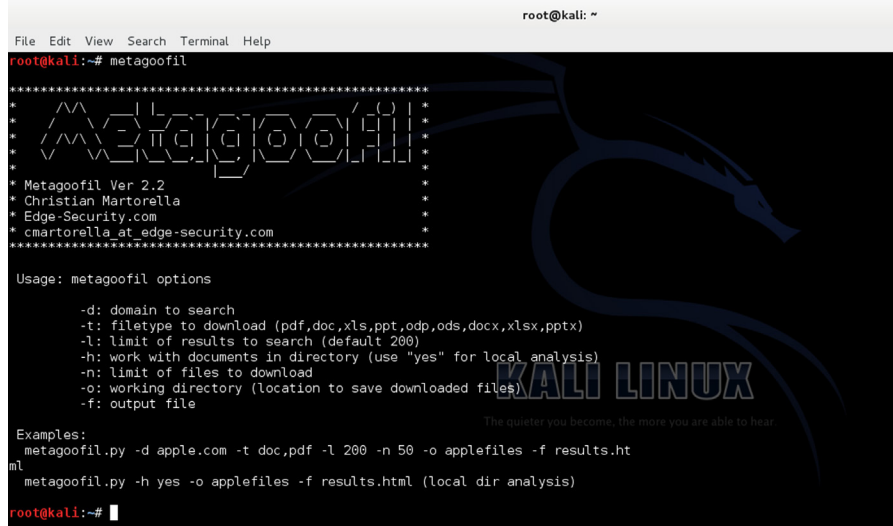
The tool can be downloaded from <https://code.google.com/p/metagoofil/downloads/list>. Using this tool is fairly easy; there are a few simple switches that can be used to perform the task.

The list of the options is as following:

```
Metagoofil options
-d: domain to search
-t: filetype to download (pdf, doc, xls, ppt, odp, ods, docx, xlsx, pptx)
-l: limit of results to search (default 200)
-h: work with documents in directory (use "yes" for local analysis)
-n: limit of files to download
-o: working directory (location to save downloaded files)
-f: output file
```

We can provide the queries such as the one mentioned below to run a scan on target domain and get the result in the form of a HTML file, which can be easily read in any browser:

```
metagoofil -d example.com -t doc,pdf -l 100 -n 7 -o /root/Desktop/meta -f /root/Desktop/meta/result.html
```



```
root@kali: ~
File Edit View Search Terminal Help
root@kali:~# metagoofil
*****
* Metagoofil Ver 2.2
* Christian Martorella
* Edge-Security.com
* cmartorella_at_edge-security.com
*****

Usage: metagoofil options

-d: domain to search
-t: filetype to download (pdf,doc,xls,ppt,odp,ods,docx,xlsx,pptx)
-l: limit of results to search (default 200)
-h: work with documents in directory (use "yes" for local analysis)
-n: limit of files to download
-o: working directory (location to save downloaded files)
-f: output file

Examples:
metagoofil.py -d apple.com -t doc,pdf -l 200 -n 50 -o applefiles -f results.ht
ml
metagoofil.py -h yes -o applefiles -f results.html (local dir analysis)

root@kali:~#
```

FIGURE 7.7

Metagoofil interface.

Similar to FOCA, Metagoofil also performs search for documents using search engine and downloads them locally to perform metadata extraction using various

Python libraries. Once the extraction process is complete the results are simply displayed in the console. As mentioned above these results can also be saved as a HTML file for future reference using the -f switch.

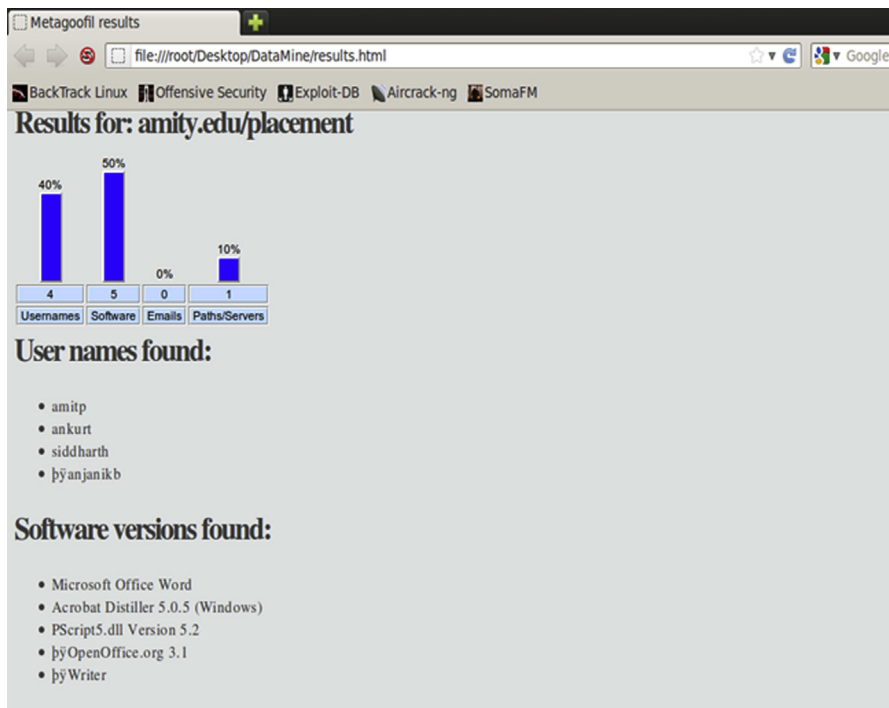


FIGURE 7.8

Metagoofil result.

Similarly there are other tools which can be used for metadata extraction from various different files, some of these are listed below:

- MediaInfo—audio and video files (<http://mediainfo.net/en/MediaInfo>)
- Gspot—video files (<http://gspot.headbands.com/>)
- VideoInspector—video files (<http://www.kcsoftwares.com/?vtb#help>)
- SWF Investigator—SWF/flash files <http://labs.adobe.com/downloads/swfinvestigator.html>)
- Audacity—audio files (<http://audacity.sourceforge.net/>)

IMPACT

The information collected using metadata extraction can be handy and used to craft many different attacks on the victim by stalkers, people with wrong motivations and even government organizations. The real-life scenario can be worse than what

we can expect. As information collected from the above process provide victims' device details, area of interest, and sometime geolocation also, the information such as username, software used, operating system etc. is also very critical for an attacker. This information can be used against the victim using simple methods such as social engineering or to exploit any device-specific vulnerability that harms the victim personally in real life as it also provides exact location where the victim generally spends time.

And all those things are possible just because of some data that mostly nobody cares or some might not even realize its existence, even if they do, then also most of them are not aware where this data can lead to and how it makes their real as well as virtual life vulnerable.

As we have seen that how much critical information is revealed through the documents and files uploaded without us realizing it and what are possibilities of turning this data as critical information against a victim and use them as an attack vector. Now there must be a way to stop this, and it's called as data leakage protection (DLP).

SEARCH DIGGITY

In the last chapter we learned a about advanced search features of this interesting tool. For a quick review Search Diggity is tool by Bishop Fox which has a huge set of options and a large database of queries for various search engines which allow us to gather compromising information related to our target. But in this chapter we are most interested on one of the specific tab of this tool and that is DLP.

There are wide numbers of options to choose from side bar of DLP tab in search Diggity. Some of the options are credit card, bank account number, passwords, sensitive files, etc.

This DLP tab generally is a dependent one. We cannot directly use this. First we have to run some search queries on a domain of our interest then select and download all the files those are found after completion of that search query than provide the path in DLP tab to check whether any sensitive data is exposed to public for that particular domain or not. To do so we can choose either Google tab or Bing tab which means either Google search engine or Bing and in that have to select "DLPDiggity initial" option to start searching for backup, config files, financial details, database details, logs and other files such as text or word document, and many more from that domain of our interest. Though there is a option to only choose some specific suboptions from "DLPDiggity initial" option, from demo prospective let's search for all the suboptions. After completion of the query we will get all the available files in tabular format in a result section of this tool. Select all the files that we got and download the same. It will save all the files in default path and in a folder called DiggityDownloads.

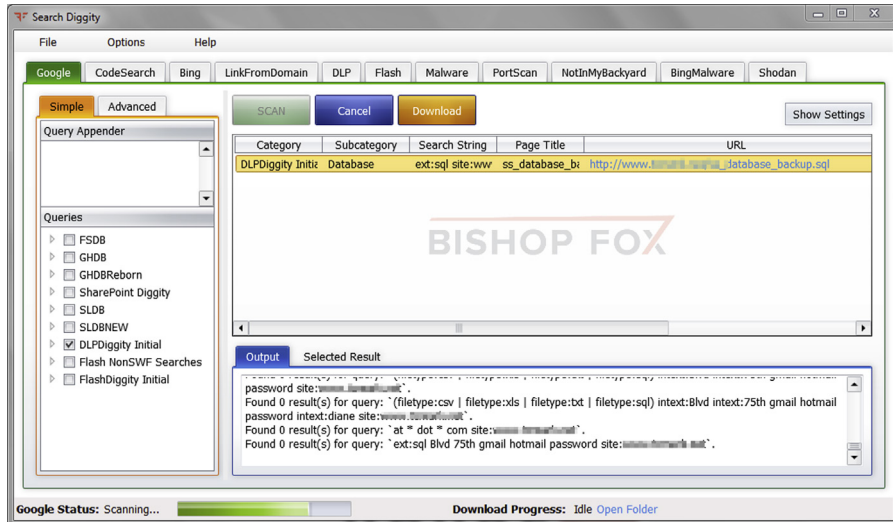


FIGURE 7.9

DLPDiggity Initial scanning.

Now switch the tab to DLP. In the top we can see the default DiggityDownloads path will be present in scan result path. So just select one or more options available in DLP tab. For demo we will select quick Checks option and click on Search to get the result.

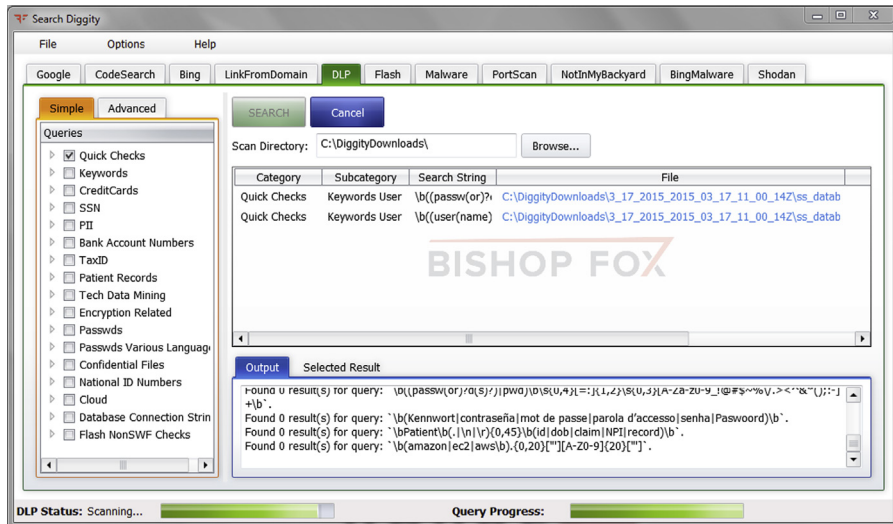


FIGURE 7.10

DLP Quick Checks.

The result sometimes might show scary results such as credit card numbers, bunch of passwords, etc. That is the power of this tool. But our main focus is not about discovery of sensitive files but DLP. So get all the details from the tool's final result. The result shows in an easy and understandable manner, in what page or document what data is available. So that the domain owner can remove or encrypt the same to avoid data loss.

METADATA REMOVAL/DLP TOOLS

As DLP is an important method to avoid data loss. The above example is quite generic to get us some idea about how DLP works. Now as per our topic we are more interested on metadata removal. So there are also different tools available to remove metadata or we can also say them as metadata DLP tools. Some of those are mentioned below.

METASHIELD PROTECTOR

MetaShield Protector is a solution which helps to prevent data loss through office documents published on the website. It is installed and integrated at web server level of the website. The only limitation of this is that, it is only available for IIS web server. Other than that It supports a wide range of office documents. Some of the popular file types are ppt, doc, xls, pptx, docx, xlsx, jpeg, pdf, etc. On a request for any of these document types, it cleans it on the fly and then delivers it. MetaShield Protector can be found at https://www.elevenpaths.com/services/html_en/metashield.html. The tool is available at <https://www.elevenpaths.com/labstools/emetrules/index.html>.

MAT

MAT or metadata anonymization toolkit is a graphical user interface tool which also helps to remove metadata from different types of files. It is developed in Python and utilizes hachoir library for the purpose. As earlier we discussed a bit about hachoir Python library and one of its project in hachoir-metadata portion, this is another project based on the same library. The details regarding the same can be found here <https://mat.boum.org/>.

The best thing about MAT is that it is open source and supports a wide range of file extensions such as png, jpeg, docx, pptx, xlsx, pdf, tar, mp3, torrent etc.

MyDLP

It is a product by Comodo which also provides wide range of security product and services. MyDLP is an one stop solution for different potential data leak areas. In an organization not only documents but also emails, USB devices, and other similar devices are potential source of data leak. And in this case it allows an organization to easily deploy and configure this solution to monitor, inspect, and prevent all the outgoing critical data. The details of MyDLP can be found here. <http://www.mydlp.com>.

OpenDLP

OpenDLP is an open source centrally managed data loss prevention tool released under the GPL. From a centralized web application it can identify sensitive data in different types of systems such as Windows and Unix as well as different types of databases such as MySQL and MSSQL. The project can be found here. <https://code.google.com/p/openslp/>.

DOC SCRUBBER

A freeware to scrub off hidden data from word documents (.doc). Some of its popular features are it allows to scrub multiple doc files at a time. Doc Scrubber can be downloaded from <http://www.javacoolsoftware.com/dsdownload.html>.

REMOVING GEO-TAGS

As we discussed earlier that how geotags can be dangerous for a user in an attacker point of view, as it reveals exact location about a user, here some settings in Picasa can help us to remove these geotags. Picasa, the image organizing and editing application by Google can help to remove geotags from images. The link to the help and support page is <http://support.google.com/picasa/bin/answer.py?hl=en&answer=70822>.

We can also use Exif Tool discussed earlier to remove such data.

Though mainly metadata is used for the organization and linking of data it can also be critical during cyber investigations as well as pentest exercises. As discussed earlier, most of them are harmless but sometimes it can reveal some sensitive data. As many individuals as well as organizations are unaware of its existence, they don't pay much attention to it. The solutions discussed above must be tried to make it easier to mitigate any risk arising from such information.