Information Security

Dark Web Project Report

Deep Web Easy Exploration Tool

Florent Gontharet

MSc. Ethical Hacking

– 2014 –

Table of Contents

Executive Summary

As we started our researches, we discovered an impressive percentage of people not aware of what the misuse or misconfiguration could lead to. The data leakage represents a main threat for anyone's personal data, and could easily be detected and sometimes avoided by the user itself.

As we analysed, we realized the importance to teach to people to be aware of it, to take responsibility and actions to ensure their privacy. Internet is a great tool but misuse can lead to big consequences even more in the case of big companies, as the problem exists as an individual or as a whole company.

We decided to create a tool allowing users to easily perform their own analysis. Our application is a prototype, we targeted the useful capabilities to program and already came with a first version allowing the user to see a first preview of it's data available on Internet.

We used the dark web mechanisms to target the main points for a standard user.

Introduction

A nowadays' problem for anyone on Internet is about personal data. The visible part of the Internet can not show it, but the deep web represents a goldmine for someone capable of time and devotion to dig in this enormous amount of data. Also known as the dark web, because represented as the dark side of the Internet, a lot of information is there, but hidden from a standard user, making it more difficult to explore.

Tor is probably the most famous ambassador of the dark web. Used all over the world, the statistics of the Tor's Project shows an important interest for users to connect such networks: an average of 3 million clients connected during the last winter.

In its cyber-security 2013 Technical Report [5], PwC highlighted an impressive 93% of the companies with more than 250 employees had a security breach in 2012. The percentage goes to 87% for companies with less than 250 employees. This report shows an important change that has to be made by companies in order to ensure their own security. More than one third (36%) of those security breaches were caused by inadvertent human error. It primarily shows that companies should pay more attention to the situation, but most of all, that they need to keep an eye on the dark web, to be aware of information they are leaking.

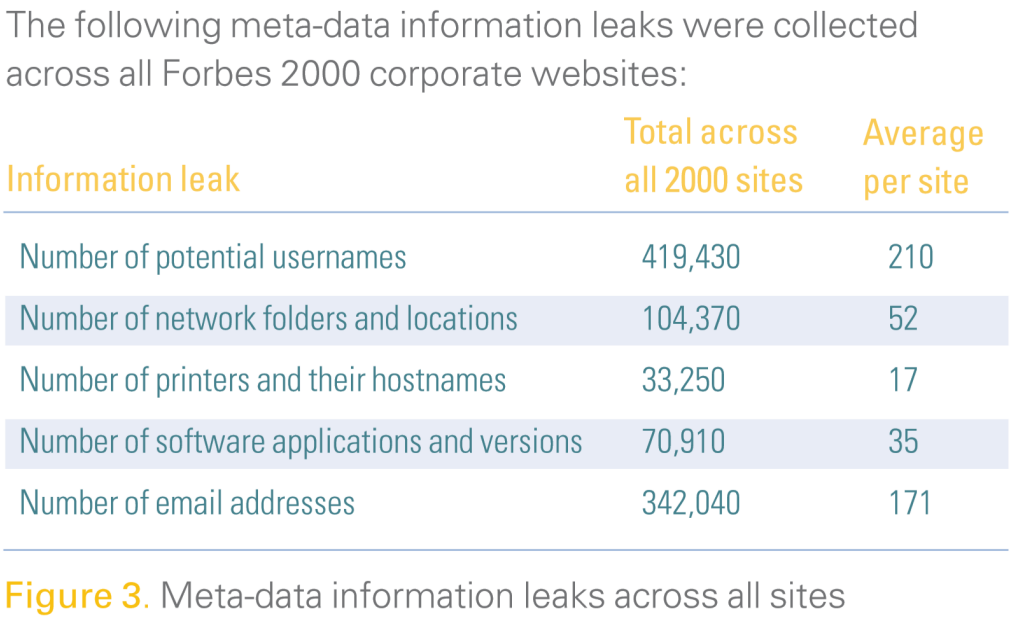
This project has been made in order to highlight the need and analyse it in the creation of a tool, in order to help companies to analyse by themselves how they appear on the Internet. We will have a look at the deep web mechanisms in order to understand the answers given by the tool. Our researches will also give other existing tools that provide their solution.

The decision to build a tool came with the idea that it's a growing problem that people has to be conscientious about. The idea of this tool is not to pretend analysing the dark web in it's deepness, but to provide a quick preview of the type and amount of information discoverable about a company, a person, whoever a bit curious to know what data about them can be found, to make them aware of the dark web, and their personal data.

Literature Review

As titled by The Guardian [1], the deep web represents “The dark side of the Internet”. In an interview, Michael K. Bergman said the dark web can be estimated as two or three times the size of the “regular” web. The amount of information is really too big to perform a manual analysis, it is one of our first conclusion, why we decided to create a tool.

Information in this field has not been easy to find. Companies do not provide such information that easily, and the dark web is not that easy to explore. Among other studies, KPMG “Publish and Be Damned: What does your online corporate profile reveal?” [4]. The study had take in consideration the Forbes 2000 companies to analyse “the leak of potentially dangerous data”:



The following meta-data information leaks were collected across all Forbes 2000 corporate websites:

Information leak	Total across all 2000 sites	Average per site
Number of potential usernames	419,430	210
Number of network folders and locations	104,370	52
Number of printers and their hostnames	33,250	17
Number of software applications and versions	70,910	35
Number of email addresses	342,040	171

Figure 3. Meta-data information leaks across all sites

Figure 1: Forbes Companies - Meta-data information leaks across all companies websites

As we can see, the companies' websites themselves can represent an important problem. What should be well known by the company is in fact totally ignored, and the content available through their website can represents sensible information.

The Symantec State of Security 2011 [6] shows a high interests for private information during cyber crimes:

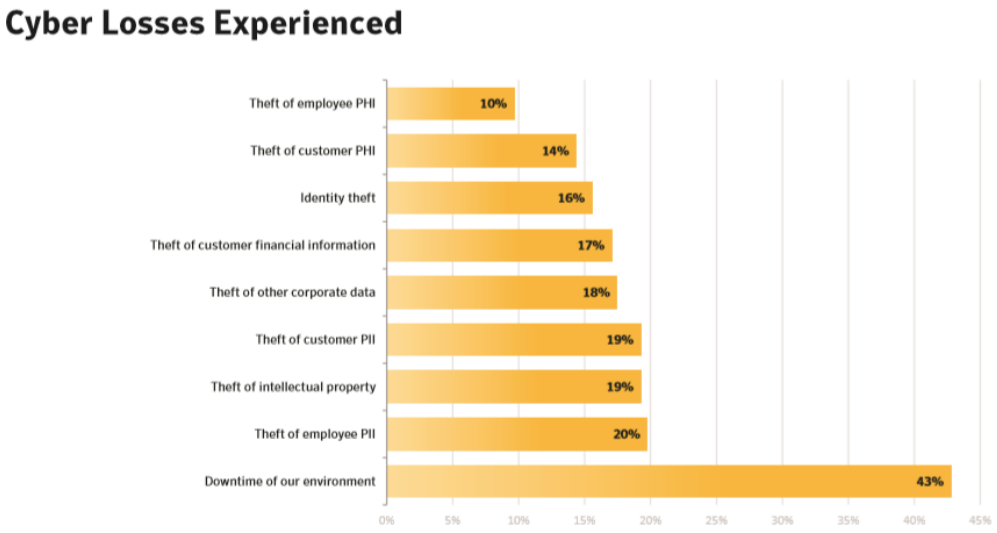


Figure 2: Cyber losses experienced by companies (Symantec 2011)

As we can see, companies are targets, but thieves are highly interested in employees' personal data as well. The company should defend itself, but everyone should also be aware on its own.

KPMG added in an article of 2013 [3]: “The dark web, and sites such as Silk Road, is not yet on the list of high priority threats facing today’s boardrooms; but it may be in the future.”, presenting it as the newest risk to legitimate business.

In its analysis, Dark Web: Exploring and Data Mining the Dark Side of the Web [2], Chen H. give a methodology to collect and analyse dark web information:

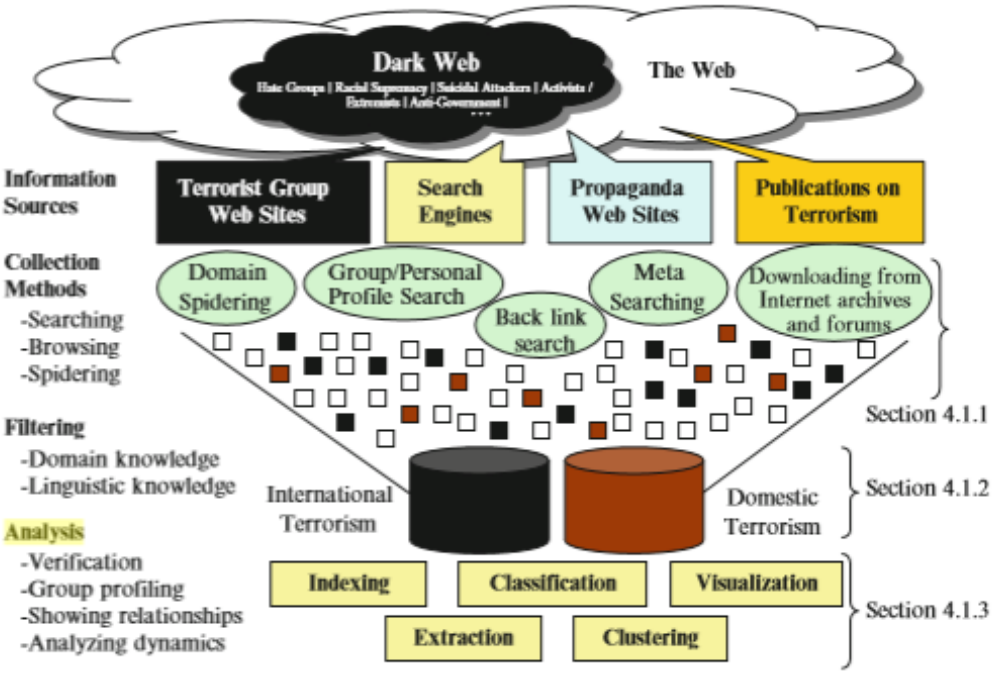


Figure 3: A methodology to collect and analyse the dark web

We will use this methodology for our application, not in its details, but this is the main approach of our exploration: the use of search engines and websites' content to perform an analysis of the data that can be gathered among those tools.

Main Content

In our task of helping persons and companies to keep control over their data, we decided to create an easy-to-use and user-friendly tool. In order to create our application, we used Qt libraries, providing GUI tools and is cross platform. We will use the version based on C++, a really powerful language for programming.

Project settings

As the dark web is hard to explore, our tool is primarily designed to provide an exploring tool. As we saw, alternative search engines or requests can provide a first approach, as they represent an entrance door to the real deep web. As the tool should fit for both companies and individuals, the tool has been made to allow it. A first figure of the main interface, allowing control over the application behaviour:

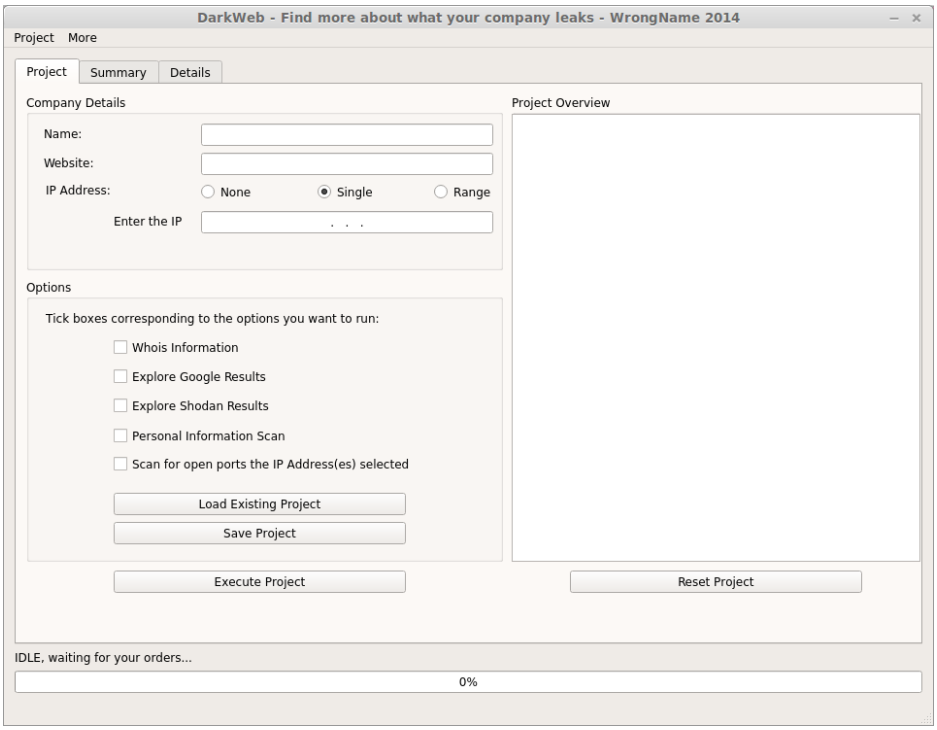


Figure 4: The main interface of our application's prototype

As we can see, we chose to approach the name and website as main elements for our researches. The application has also been designed to perform a port scan, but to match the deadlines and not become to vast on the topic, we decided to consecrate this study to our main goal, and to not implement the scan in the first release. Also, as this feature could be use to perform illegal port scans, we ensured during our experimentations to only target our own network, as it should be done. A port scan can be considered as an attack, or a suspect behaviour, and it shouldn't be performed without knowledge and authorizations.

This main tab, called “Project”, allows our user to create a project he can save, modify, and reload, in order to be more user-friendly. One button “Reset Project” is also here for the same purpose. The white field on the right displays the current information that will be used, as a summary of the configuration. When the “Execute Project” button is pressed, the analysis starts with the different information filled.

As options, the user can choose multiple ones. Here is the complete list and the instructions followed in each case, while the analyse is performed:

Option	Analysis performed
Whois Information	The first way to collect information about a website is the Whois database. Someone using a DNS name has to register, and in some cases those information are kept publicly. This analysis will request the available data and give names, addresses or phone numbers that are linked to the given website.
Explore Google Results	Perform a Google search request to get results. Such request can provide documents or web pages available directly using the famous search engine.
Explore Shodan Results	Shodan is a website providing network analysis results. Such results can provide the user more data regarding hosts publicly visible from the Internet.
Personal Information Scan	As the research performed to explore the Google results regarding a website, the name can also have some interesting findings, as it can appear in some listing of leaked private information. An improvement would be to perform the same using an email address.

Table 1: Options analysis and corresponding instructions

As we can see, the first option is articulated around the problem of a website owner. Today, everyone is susceptible to have a use for a website, and performing a whois is a first step to figure out that personal information can be found. The technical side can make it hard for standard user to have a look to it. Our application uses a PHP class designed to ask directly the databases to obtain the information. We could have use a web service already existing, but as an application to help ensuring privacy, the use of a third service could reveal some troubles. The class has been created by easyDNS Tech. and Mark Jeftovic (maintained by David Saez, available from http://www.phpwhois.org). The use of PHP is a personal preference, ensure a large amount of documentation, and a cross-platform support.

The second option uses the Google Search API, as we want Google results. The free version is limited to some thousands of requests a day, so we can easily perform our few requests. The result is received as Json document, that we will parse and display. The same principle is used for the personal information scan.

Explore Shodan Results uses the provided PHP REST API directly from ShodanHQ. The REST API is free and provides us a tool to research hosts by hostname, as we have that one, we will ask for all the servers linked. The result is a Json document as well, and we will parse and display it to the user.

The Code to perform a request using Qt is presented in Appendix 1: Send a request using Qt. In order to see an example of Json File received from Google, please see Appendix 2: A Json File Example.

Scan for open ports has not been implemented in the current version as the first one already required an important amount of work and is not primarily linked to our researches.

Summary Tab

Once the analysis has been ended, the user is allow to access the Summary tab. It gives the user a list of raw information, as a log of what has been performed, and the output of each of the steps. Here is an example of the results we obtained running the application with example data:

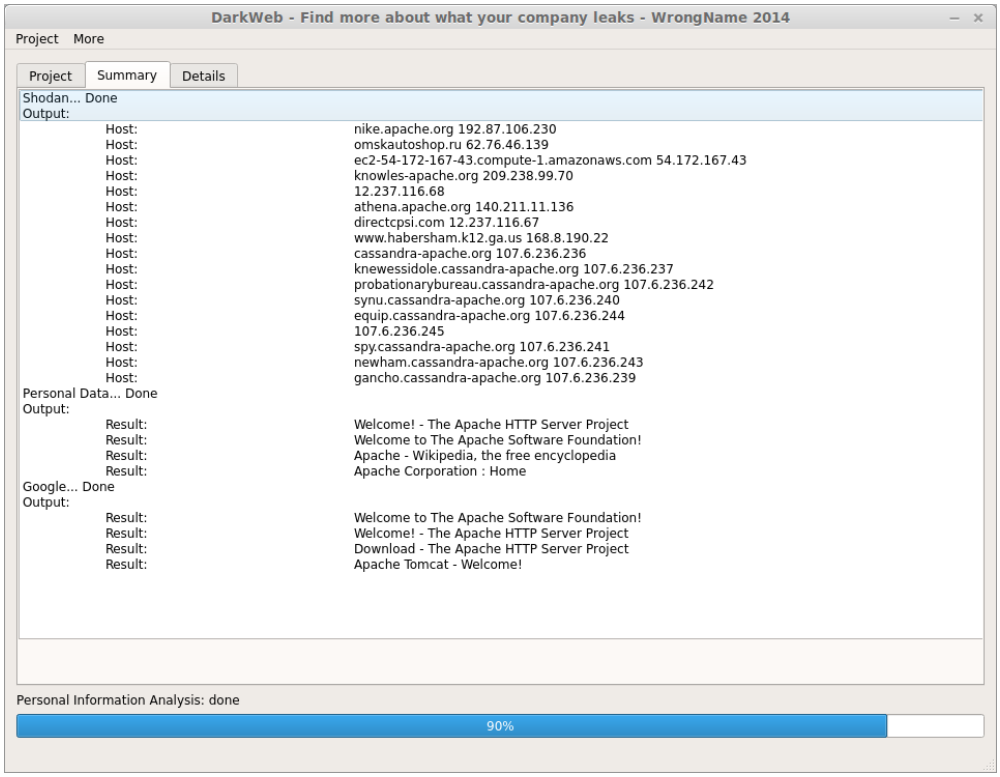


Figure 5: The summary tab, after an example scan

As we can see, information has been found is some categories, and they have been displayed to the user. But to ensure a user friendly application, the next tab will give the user more information.

Details Tab

The details tab is here to give the user more hints about the founded information. A colour indicator gives the user a level of danger for the listed information:

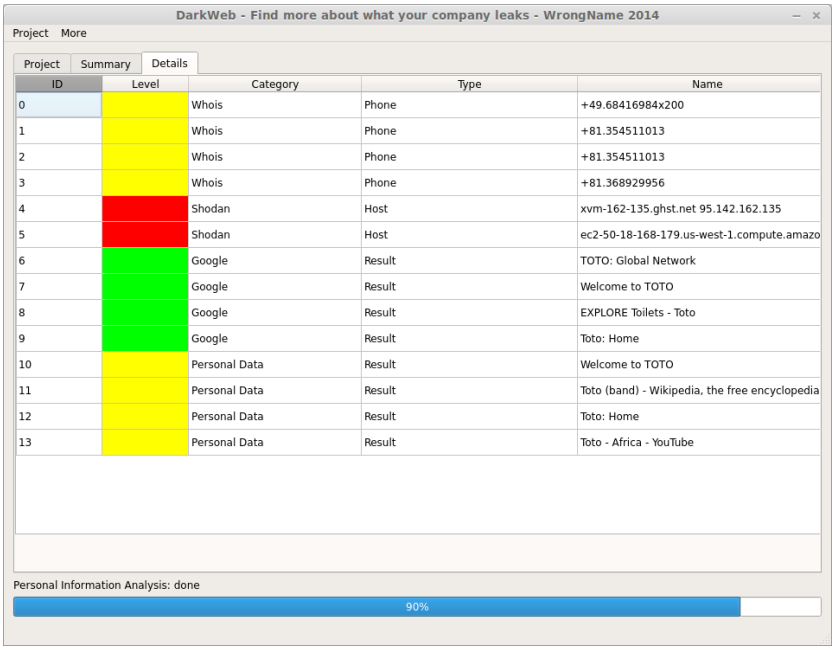


Figure 6: An example of the results details display

This third tab provides the user clear and easy-to-analyse results list. As the figure is from the prototype, it is subject to small changes, as for the algorithm rating the level in example. The main idea is not to drawn the user with a deep analyse of the dark web, but to provide a first preview. But the process cannot be entirely automatised, as the final output still has to be reviewed. But it helps the extraction of data from that web, that can seem dark.

To go deeper: some tools

Some tools already existing can also provide help to someone curious about the data available. Our tool is made to alert a first eye to the question of the dark web, and not to provide a deep and complete analysis. Our project conduct us to discover more tools that can be useful in a such case. We will present them quickly in the following part.

FOCA

A tool for metadata analysis, really complete and powerful, it can provide a lot of results, regarding any kind of file, network, or data. FOCA uses the same principles as our tool, but extends it with an advanced tool, it can be impressive for a standard user to start to get use to using FOCA, as it is more technical, but it is definitely the next step after our tool. Here is a picture of the main window:

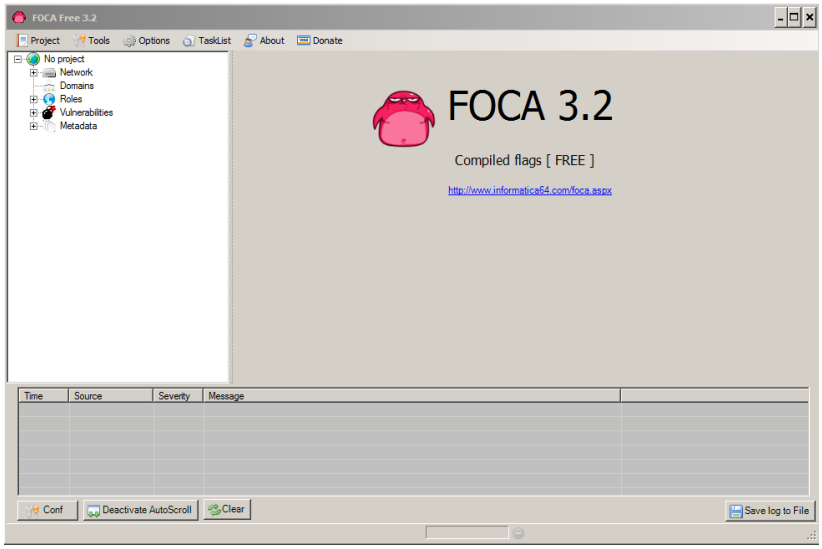


Figure 7: FOCA, a complete tool

OOMetaExtractor

As FOCA shows it, metadata are the next step in the analysis, and as this reports has helping keeping privacy, here is a tool to control metadata of files someone would like to share. It is designed to work with OpenOffice documents, and allows a user to control and modify information available in the file's metadata. Easy to use, it provides a simple interface that can be seen in the next figure:

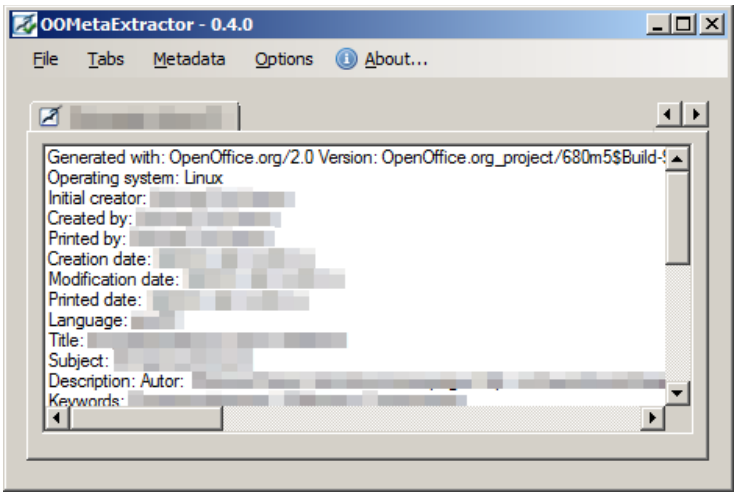


Figure 8: OOMetaDataExtractor, a simple but powerful tool

Conclusion

Even without the port scan, our application is already able to help people to perform a quick and fast preview. The filters to reduce the verbosity of our application are really variable, and we cannot define rules for some results, that can fit any need. Also some results can seem really too much, as the application give them all. Technical solutions can be a first improvement.

Our experimentation highlighted main problems, and other tools to deepen the analysis. The aim of our program is reached and we have discovered an important amount of information about a topic that seems dark and inaccessible.

As a conclusion, we saw that privacy is not only an issue, it is a matter of conscientiousness, as a lot of personal data is simply the result of misuse and misconfiguration. The purpose of this project was to highlight this new problem, and provide a tool to help people taking in consideration the issue. The application has been designed and fits the main points we outlined.

The deep web remain mysterious and vast but a simple and quick approach can already helps people to understand what is going on, and what they should do in order to fix the situation in case of an issue. In order to go further in the analysis, we have seen some methodology and tools that can provide a lot more information. This last point is really important while working with the deep web, as its content can quickly flood the user or interesting information.

To finish this research, we found a way to fill the gap between Internet user and the deep web content, but as the deep web content is growing, it is important to keep an eye on what regards personal information.

Bibliography

[1] Beckett, A. (2009). The dark side of the internet. *The Guardian*. .....[online]

Available at: http://www.theguardian.com/technology/2009/nov/26/dark-side-internet-freenet [Accessed 15 Dec. 2014]

[2] Chen, H. 2011, Dark Web: Exploring and Data Mining the Dark Side of the Web, Springer-Verlag.....[online]

Available at (with credentials): http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-1-4614-1556-5 [Accessed 14 Dec. 2014]

[3] Jones, P. 2013, The Dark Web: The Newest Risk to Legitimate Business. KPMG. .....[online]

Available at: http://www.kpmg.com/uk/en/issuesandinsights/articlespublications/pages/the-dark-web.aspx [Accessed 14 Dec. 2014]

[4] KPMG Report. 2014, Publish and Be Damned: What does your online corporate profile reveal?.....[online]

Available at: http://www.kpmg.com/NL/nl/IssuesAndInsights/ArticlesPublications/Documents/PDF/ITRisk-Consulting/Publish-and-be-Damned.pdf [Accessed 17 Oct. 2014]

[5] PwC. 2013, Information Security Breaches Survey.....[online]

Available at: https://www.pwc.co.uk/assets/pdf/cyber-security-2013-technical-report.pdf [Accessed 14 Dec. 2014]

[6] Symantec. 2011, State of Security.....[online]

Available at: http://www.themavision.fr/upload/docs/application/pdf/2011-10/symantec_-_state_of_security_2011.pdf [Accessed 14 Dec. 2014]

Appendices

Appendix 1: Send a request using Qt

```
#include "website.h"
#include <QDebug>
#include <QUrl>
#include <QtNetwork/QNetworkAccessManager>
#include <QtNetwork/QNetworkRequest>
#include <QtNetwork/QNetworkReply>
#include <QUrlQuery>

// We create our new request, using the given link
Website::Website(QString link)
{
    this->networkMgr = new QNetworkAccessManager();
    setWebsite(link);
}

void Website::setWebsite(QString link)
{
    this->link = link;
}

QString Website::getWebsite()
{
    return this->link;
}

void Website::sendRequest()
{
    // The Network Manager provides us a way to send the request and wait for the answer
    QNetworkReply *reply = this->networkMgr->get(QNetworkRequest(QUrl(this->link)));

    // We will send back the answer to the GUI in order to display the results
    MainWindow* mw = new MainWindow();

    /*
       Qt provides a signals and slots mechanims. When the request's response will be
       received, the function « requestResults() » from the MainWindow class will be called
       to receive it.
    */
    connect(reply, SIGNAL(finished()), mw, SLOT(requestResults()));
}
```

Appendix 2: A Json File Example


```
{
    "responseData": {
        "cursor": {
            "currentPageIndex": 0,
            "estimatedResultCount": "6300000",
            "moreResultsUrl": "http://www.google.com/search?oe=utf8&ie=utf8&source=uds&start=0&hl=en-
GB&q=toto.com",
            "pages": [
                {
                    "label": 1,
                    "start": "0"
                },
                {
                    "label": 2,
                    "start": "4"
                },
                {
                    "label": 3,
                    "start": "8"
                },
                {
                    "label": 4,
                    "start": "12"
                },
                {
                    "label": 5,
                    "start": "16"
                },
                {
                    "label": 6,
                    "start": "20"
                },
                {
                    "label": 7,
                    "start": "24"
                },
                {
                    "label": 8,
                    "start": "28"
                }
            ],
            "resultCount": "6,300,000",
            "searchResultTime": "0.16"
        },
        "results": [
            {
                "GsearchResultClass": "GwebSearch",
                "cacheUrl": "http://www.google.com/search?q=cache:iUMpyfjDUTwJ:www.toto.com",
                "content": "Global Network; About <b>TOTO</b>; Our Environmental Initiatives; CSR
Artivities; \nInvestor Relarions ... Oct 31,2014: <b>TOTO</b> announces financial results for
2Q,\nFY2014.",
                "title": "<b>TOTO</b>: Global Network",
                "titleNoFormatting": "TOTO: Global Network",
                "unescapedUrl": "http://www.toto.com/",
                "url": "http://www.toto.com/",
                "visibleUrl": "www.toto.com"
            },
            {
                "GsearchResultClass": "GwebSearch",
                "cacheUrl": "http://www.google.com/search?q=cache:7zAA3y2V65IJ:www.totousa.com",
                "content": "Manufactures a complete line of residential and commercial plumbing
products.",
                "title": "Welcome to <b>TOTO</b>",
                "titleNoFormatting": "Welcome to TOTO",
                "unescapedUrl": "http://www.totousa.com/",
```

```
                "url": "http://www.totousa.com/",
                "visibleUrl": "www.totousa.com"
            },
            {
                "GsearchResultClass": "GwebSearch",
                "cacheUrl": "http://www.google.com/search?q=cache:MGop-Un7ofgJ:www.totousa.com",
                "content": "Classic configuration and adaptable to \nany budget, two-piece toilets
areÂ ...",
                "title": "EXPLORE Toilets - <b>Toto</b>",
                "titleNoFormatting": "EXPLORE Toilets - Toto",
                "unescapedUrl": "http://www.totousa.com/products/toilets",
                "url": "http://www.totousa.com/products/toilets",
                "visibleUrl": "www.totousa.com"
            },
            {
                "GsearchResultClass": "GwebSearch",
                "cacheUrl": "http://www.google.com/search?q=cache:MXJ_Uct7f_EJ:www.totoofficial.com",
                "content": "116 Results <b>...</b> We are excited to announce that <b>Toto</b> will be
performing at the New Orleans \nHouse of Blues on August 18, and Diamond VIP Packages (concertÂ ...",
                "title": "<b>Toto</b>: Home",
                "titleNoFormatting": "Toto: Home",
                "unescapedUrl": "http://www.totoofficial.com/",
                "url": "http://www.totoofficial.com/",
                "visibleUrl": "www.totoofficial.com"
            }
        ]
    },
    "responseDetails": null,
    "responseStatus": 200
}
```


</ end of Json Example File>