UNIVERSITY OF TARTU
Institute of Computer Science
Cyber Security Curriculum

Karl Mendelman

Fingerprinting a Organization Using Metadata of Public Documents

Master's Thesis (30 ECTS)

Supervisor(s): Olaf Manuel Maennel
Raimundas Matulevicius

Tartu 2018

Fingerprinting a Organization Using Metadata of Public Documents

Abstract:

Many companies and organizations use Internet for their business activities to make information about their products and services more available for customers. Often those organizations and companies share electronic documents on their websites, such as manuals, whitepapers, guidelines, templates, and other documents which are considered as important to share. Documents which are uploaded on organizations' websites can contain extra information, such as metadata.

Metadata is defined as data which describes other data. Metadata associated with documents can contain information about names of authors, creators information, documents general properties, the name of the server, or path where the document was modified. Metadata is added into documents mainly by automated process when document is created, and if documents' metadata is not properly removed before sharing, it could contain sensitive information. Usually people are not aware about metadata existence in documents and could unwillingly leak information about their organization or about themselves. This information can be used for fingerprinting basis or conducting cyber attacks.

In this thesis paper, electronic documents' metadata which are shared on Estonian governmental organizations websites were analyzed. More specifically, three institutions' public documents' metadata were observed in order to identify metadata vulnerabilities that can be used for fingerprinting purposes. To achieve that, a fingerprinting method was developed and utilized against observed websites. This thesis is divided into two different stages, where first stage describes the developed fingerprinting method, and second stage presents the outcomes of metadata analysis with the developed method.

The results of the conducted research showed that almost all documents which were analyzed contained information which could be used for fingerprinting purposes. We processed 2643 documents, where only 12 documents had metadata properly removed. All other documents contained pieces of information that describes environment where document was created and additionally exposed information that could be used for conducting cyber-attacks.

This thesis is written in English and is 77 pages long, including 6 chapters, 41 figures and 26 tables.

Keywords:

Metadata, fingerprinting organization, metadata extraction, information gathering, cyber attacks

CERCS: P170

Organisatsiooni kaardistamine kasutades avalike dokumentide metaandmeid

Lühikokkuvõte:

Paljud ettevõtted ja asutused kasutavad äritegevuseks Interneti, et muuta informatsioon enda pakutavate toodete ja teenuste kohta kättesaadavamaks. Tihtipeale need ettevõtted ja asutused jagavad oma veebilehel elektroonilisi dokumente (näiteks tabelid statistiliste andmetega, juhendid, näited ja õpetused, artiklid, blanketid ja muud dokumendid), mida peetakse vajalikuks jagada. Dokumendid, mis on veebilehtedel kõigile internetikasutajatele vabalt kättesaadavad, võivad sisaldada metaandmeid.

Metaandmed on andmed, mis kirjeldavad teisi andmeid, ehk metaandmed kirjeldavad dokumendi sisu ja dokumendi üldiseid omadusi. Metaandmed on näiteks kasutajanimi, kes dokumendi koostas, salvestas, printis või redigeeris, kuid lisaks ka ajatemplid millal eelpool mainitud tegevusi tehti. Täiendavalt võib dokumentides olla informatsiooni arvutite ja infosüsteemide kohta, kus seda dokumenti töödeldi. Metaandmete lisamine dokumentidele toimub valdavalt automaatselt ning kui metaandmeid dokumendist eemaldatud pole, võib dokumendi metaandmetesse sattuda tundlikku informatsiooni kasutaja ja asutuse kohta. Metaandmete olemasolu dokumendis on paljude kasutajate jaoks teadmata ning nad ei ole teadlikud, et võivad potentsiaalselt lekitada informatsiooni asutuse või süsteemide kohta, kus dokumenti töödeldi. Seda informatsiooni on võimalik kasutada küberrünnakute läbiviimiseks või asutuse kaardistamiseks.

See magistritöö uurib dokumentide metaandmeid, mis on ligipääsetavad Eesti riigiasutuste veebilehtedel ning mis on kõigile Internetikasutajatele vabalt kättesaadavad. Täpsemalt on vaatluse alla võetud kolme riigiasutuse veebilehel olevad dokumentide metaandmed, et välja selgitada, kas nendes peituvat informatsiooni on võimalik kasutada asutuse kaardistamiseks ja võimalike küberrünnakute teostamiseks. Selle täideviimiseks kasutati kahest etapist koosnevat meetodit. Esimene etapp tugines meetodite välja töötamisel, kuidas asutusi kaardistada, kasutades ainult dokumentide metaandmeid. Teine etapp kirjeldas esimeses etapis välja töötatud meetodi rakendamisel saadud tulemuste analüüsist ja järeldustest.

Tehtud analüüsi tulemus näitas, et peaaegu kõik dokumendid sisaldavad metaandmeid, mida on võimalik ära kasutada ühel või teisel viisil asutuse kaardistamiseks või küberrünnakute läbiviimiseks. Magistritöös analüüsisime kokku 2643 dokumenti, millest 12-nel olid metaandmed eemaldatud. Ülejäänud dokumendid sisaldasid informatsiooni kilde, mis kirjeldavad keskkonda kus dokumente on töödeldud ja sisaldasid informatsiooni, mida on võimalik kasutada küberrünnakute läbiviimiseks.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 77 leheküljel, 6 peatükki, 41 joonist ja 26 tabelit.

Võtmesõnad:

Metaandmed, elektroonilised dokumendid, küber rünnakud, kaardistamine

CERCS: P170

Table of Contents

1 Introduction

In today's information age, data is very crucial for every organization. Data is often called the oil of the digital era [1]. Google, Microsoft, Facebook, Apple, Amazon, are giants that deal in data and have enormous power to get to know people's habits, interests, visited places, and etc. Many of the customers' welfare and needs related decisions are made by data analysis. Potential data loss for any organization can have very negative impact financially as well as reputation wise.

Generally, organizations are aware of the information they reveal through different online mediums, but what about the data that is being exposed without the knowledge of the organization, and which could be crucial from security perspective? One place where this issue can appear are public domain webpages, especially documents which are hosted there. Documents have capabilities to store extra information, such as metadata, and those documents can contain hidden information which could be sensitive from security perspective.

Metadata is a structured description of objects containing certain properties useful to the user as well as the program on which the document was created [2]. Classical definition about metadata is data that describes other data, or information that is used to describe other information. Metadata is used in Office applications to store various types of extra data ranging from the document's author's name to the last time the document was printed. From its nature and definition's point of view metadata may seem safe to store within the document. However, metadata may contain very sensitive information about persons who have authored or modified the document [3].

There are several security issues that should be considered when thinking about metadata. Firstly, revealing the name of the document's author can be used in phishing or brute force attacks. Revealing the application's name and version from document metadata may be helpful for conducting potential attacks. For instance, exploits or computer malware often targets specific, known to be vulnerable version of an application or software. In addition, metadata can expose information about the origin of the document, number of authors, and keyboard layouts, which indicate potential nationality.

1.1 Motivation

This thesis describes security related issues that metadata can reveal and how metadata information can expose sensitive information about an organization's infrastructure. The reason for conducting this research is lack of awareness. Based on open literature and on results that are discussed in Chapter 4, organizations and people are not aware what is in the documents they share on the Internet. Potentially they could leak sensitive information about their organization and themselves without noticing it. Metadata can contain internal servers' IP-s, domain names, database queries, and other information which may seem harmless at first [2]. However, it could be essential information for constructing cyber attacks or damaging.

Former National Security Agency (NSA) contractor Edward Snowed described metadata with the following sentence [4]:

"Metadata is extraordinarily intrusive. As an analyst, I would prefer to be looking at metadata than looking at content, because it's quicker and easier, and it doesn't lie."

Metadata provides descriptive information about the contents or assets. By analyzing and processing it, links or patterns between different objects may be exposed. When considering metadata in documents, and especially those documents that are accessible on the Internet for everyone, the metadata can present descriptive information not only by content of a document but also information about document's author or his/her organization.

Conducting metadata extraction and analysis and presenting it, is one way to raise awareness about document creators who share documents on the Internet, and for IT managers to implement certain policies to remove metadata from documents. Introducing metadata is one of the first starting points to mitigate this widely spread issue.

This is the first study that aims at cleaning up the webpages by various ministries of Estonia from compromising metadata to avoid devastating cyber attacks. Estonia has been one of the lead countries in terms of information technologies and cyber security, and in that sense needs to play that role further.

1.2 Problem statement and the contribution

Generation of metadata can be automatic or manual. Microsoft Office applications can add metadata to documents automatically. If people do not remove it manually before sharing on the Internet, it is preserved and might contain information about the systems where the document was created or modified.

In 2007 Oracle made study and analyzed randomly downloaded Microsoft Office documents from various websites [5]. In their study they analyzed 8,846 different documents (Word documents, spreadsheets, presentations) and concluded:

„The results of this study clearly indicate that the issue of metadata and hidden information exposure is very real. The occurrence of this information within documents published to the Web for broad third-party consumption by organizations with large IT resources raises the question of how much sensitive information leaks from organizations every day during the course of normal business.“

The research showed that they managed to extract sensitive information from documents, such as hidden text, embedded objects, comments, paths, network share names, sensitive hyperlinks, sensitive include fields, and usernames. Since they downloaded public documents from several randomly selected websites, it might not be so meaningful. But downloading documents from a certain site or domain and conducting metadata extraction and analysis may prove to be a different story. Since metadata contains information about software versions, printer names, working directories, usernames, operating systems, extracting it and analyzing can expose information about targeted organization's internal network and policies.

Organizations upload many files on their sites for daily business, to make their services more available for customers. Often those documents are sales reports, manuals, templates, guides, or presentations. Those documents contain extra information which can reveal delicate information about the organization, put it in a financial risk or embarrassing situations with costly consequences. Metadata provides private information for basis of fingerprinting and getting compromising information without doing any active scans against networks.

As the metadata's capabilities to store delicate information and the lack of awareness about the risks it can bring, the following hypothesis is posed:

"Metadata of published document on Estonian governmental organization websites leaks compromising information which aids to conduct cyber-attacks against them".

In addition, we describe how this compromising metadata information can be used to create cyber attacks and outline possible attack vectors based on extracted information.

In order to validate or disproof the hypothesis, this study is logically divided into two stages.

First stage describes methodology on how to fingerprint an organization by using only their public documents' metadata. Methodology is divided into three sub stages: document gathering, metadata extraction, and metadata analysis. Methodology is described in more detail in Section 3. While the first stage focuses on introducing the methodology, the second stage presents the results of conducted research, thereby to validate or disproof the hypothesis. Section 4 gives overview of the statistical results of the study. Extracted raw metadata information is in the Appendixes.

1.3 Scope

Metadata security issues do not only occur within the public documents. The same problem is everywhere where documents are shared and proper metadata removal procedures are not implemented. This thesis paper focuses on documents which are hosted on public websites, since the data set is freely available and does not need any extra permissions to gather them.

In this study, the uploaded documents from three Estonian governmental webpages are analyzed. The scope of document types used is the following:

- PDF documents
- Microsoft Office documents

Other document formats are excluded mainly due to the occurrences of other formats being very few, according to the study described in Appendix 1.

Document extensions being analyzed and discussed in this thesis paper are: *pdf, docx, doc, xls, xlsx, ppt, pptx*. Each of those formats are described more thoroughly in Chapter 2.

The findings and the results are presented in a clearly distinct way, in other words, each governmental organization individually.

1.4 Outline

The thesis is structured as follows:

- Section 1: Introduction to the topic, including motivation, problem statement, contribution, and the scope of the study;
- Section 2: This section gives an overview of metadata and about metadata related security incidents, problems and risks metadata can cause. Overview of related works.
- Section 3: This section gives an overview of methodology on how to gather documents on websites, how to extract metadata from documents, and how to analyze it.
- Section 4: This section gives an overview of statistical results of the study. In addition, possible attack vectors were discussed.
- Section 5: This section discusses recommendations about how to mitigate the problem.
- Section 6: This section summarize the results found in the thesis work.

2 Background and related work

This chapter introduces metadata and its terminology, highlighting what it is and why it is used. Furthermore, overview of metadata in documents and the risks which it can expose are discussed.

2.1 Background

Metadata is information which describes other information [6]. The term metadata is used in different ways in different communities. Some use it to refer to machine understandable information, some use it to refer to records that describe electronic resources. Underlying concepts of metadata have been in existence as long as collection of information has been organized. For instance, in mid-18th century, photographers described content of picture, names, and time, in the logic as it done now in modern digital world.

There are four main types of metadata [6]:

- Descriptive metadata – describes resource for purposes like identification and discovery;
- Structural metadata – indicates how compound objects are put together;
- Administrative metadata – provides information to help manage resources;
- Markup languages – mix metadata and content together.

Metadata serves many purposes such as [2][3][6]:

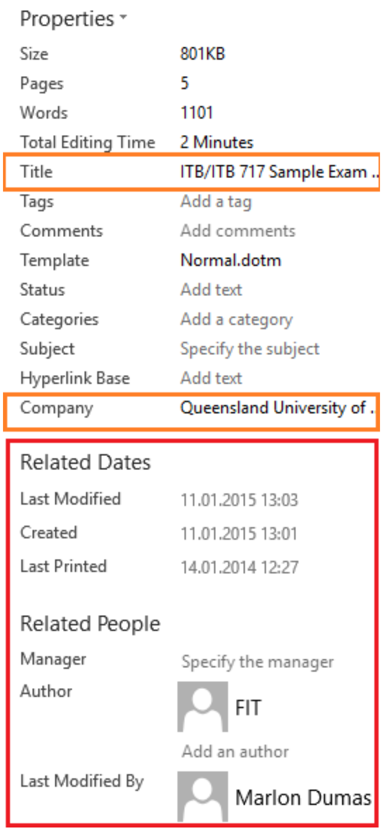
- Helps user to discover resource;
- Organize electronic resources;
- Supports archiving resources;
- Supports preservation of resources.

Metadata has been discussed lately quite actively in the context of electronic information. In that sense, metadata describes location, physical attributes, type, and form of the electronic information. Good example of metadata occurrence in describing electronic information is NSA surveillance project where governmental organization collects metadata about the phone calls – when a call is made, what number they were made to, where they were made from, and how long the calls lasted [7]. Information as such is valuable in the sense of detecting patterns between people and trying to understand their behaviour. More commonly, metadata is associated with documents and files, containing information about the names of authors, creators, properties information about file or document, the name of the server, or path where the file or document was saved. In essence, metadata addresses the underlying data of who, what, when, where, and how [6].

In this study, metadata refers to variety of information types which are found inside electronic documents.

2.2 Metadata in documents

Metadata can be simply described as data that describes other data. Microsoft documents contain variety of metadata which can include author names, document modifier's name, name of the document, person who printed the document, print and save dates, document keywords, comments, hidden information [2][3]. An example of metadata is showed in Figure 1 where document metadata is viewed using MS Word application (inside application navigating file -> properties). This kind of metadata can be added automatically or manually.



Properties

Size	801KB
Pages	5
Words	1101
Total Editing Time	2 Minutes
Title	ITB/ITB 717 Sample Exam ..
Tags	Add a tag
Comments	Add comments
Template	Normal.dotm
Status	Add text
Categories	Add a category
Subject	Specify the subject
Hyperlink Base	Add text
Company	Queensland University of .

Related Dates

Last Modified	11.01.2015 13:03
Created	11.01.2015 13:01
Last Printed	14.01.2014 12:27

Related People

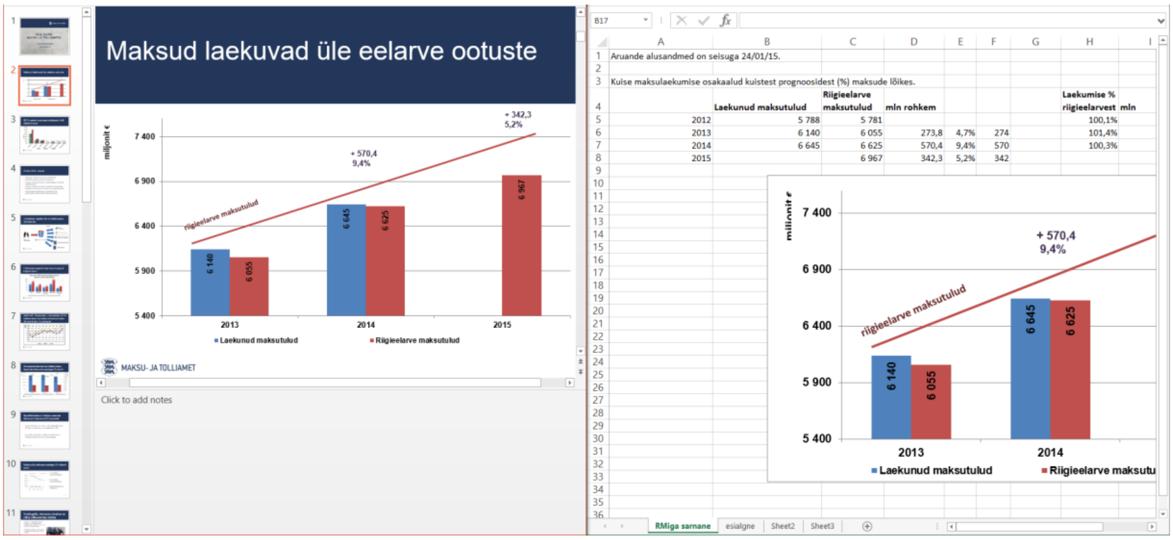
Manager	Specify the manager
Author	FIT Add an author
Last Modified By	Marlon Dumas

Figure 1: From UT website downloaded document metadata observed with MS Word application

When a file is created or modified, document processors populate some descriptive information automatically and it depends on the application the document was created with. Most of the time, the information is there for a good reason. It is needed by authoring and publishing tools to store parameters (for instance, author identifiers, printer settings) that are not immediately part of the document [8]. It enables other tools and applications to communicate with such parameters. While it is good that automatic information propagating processes are working in the background, there is danger that if a user is not aware of the presence of metadata, private or secret information may be revealed unintentionally. The propagation of unnecessary information may also violate organisational security policies.

In many cases, files' metadata contains locations where it was created or modified, giving potentially sensitive information about network shares, paths, and/or locations [2]. Metadata can be examined with the application that created it. However, some of the information that users, applications, or content management tools enrich into files are not observable without extra software or approach. In this study, information which is not accessible through application interface that created the document, is considered as hidden information. For example, author history, comments, track changes, fast save data, embedded objects.

Microsoft Office supports embedding's from other Microsoft applications; data from a spreadsheet can be embedded into a Word document as an external file. Example of this feature is shown in Figure 2. The PowerPoint presentation was downloaded from Estonian Tax and Custom Board's webpage (emta.ee). By disabling read–only protection and observing graph on the 2. slide and then selecting "Change Data" in the graph context menu, an embedded Microsoft Excel source file opens. That file contains all the source data, including formulas, graphs, numbers, raw data. Person who uploaded this document probably assumed that data was inaccessible. Situation shown in Figure 2. can be described as potentially unintended information disclosure which might cause reputational damage to the organization.



Maksud laekuvad üle eelarve ootuste

B17

Aruande alusandmed on seisuga 24/01/15.

Kuise maksulaekumise osakaalud kuistest prognoosidest (%) maksude lõikes.

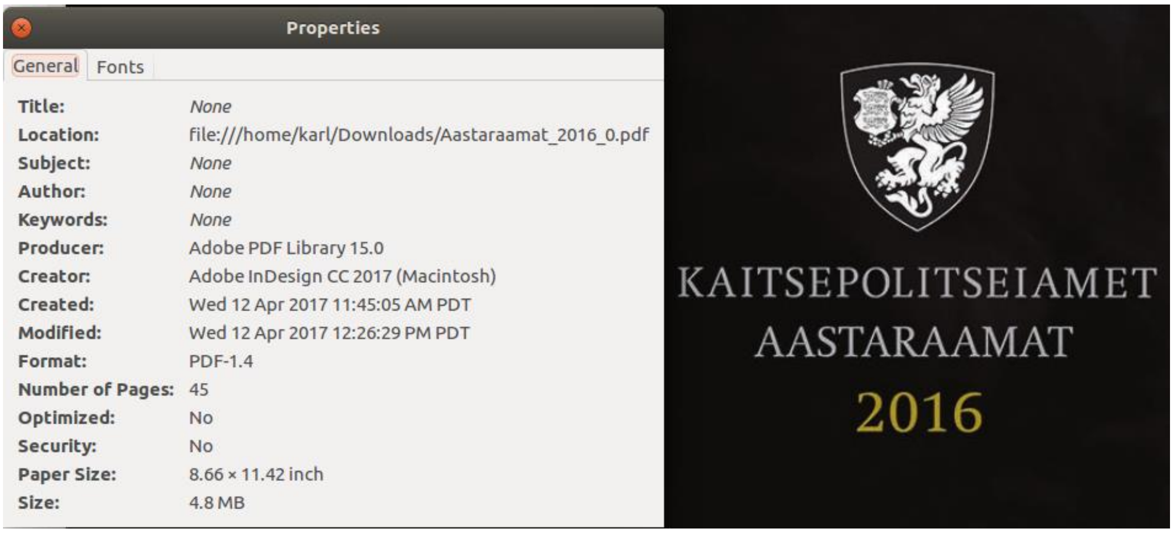
	A	B	C	D	E	F	G	H	I
4		Laekunud maksutulud	Riigieelarve maksutulud	mln rohkem				Laekumise % riigieelarvest	mln
5	2012	5 788	5 781					100,1%	
6	2013	6 140	6 055	273,8	4,7%	274		101,4%	
7	2014	6 645	6 625	570,4	9,4%	570		100,3%	
8	2015		6 967	342,3	5,2%	342			

Click to add notes

Figure 2. PowerPoint presentation file, containing a spreadsheet with an abundance of source data for current and some other calculations

Other parties can access internal spreadsheets in that way and see the calculation formulas, as well as raw data which might be used only in-house. It is ironic that the embedded spreadsheet also contains metadata that can expose the name of the person who made the calculations. Maybe the presentation author does not have the permissions to use that table and using metadata one can determine that this presentation author is violating the rules. For an attacker, embedding's information is useful for constructing attacks. For example, one of the known advanced persistent threat (APT) techniques is the reuse of legitimate documents/attachments. Embedding is a good place to receive that document/attachment, infect it with malware, and send it to the document's author.

Every time a document is opened, edited and saved, metadata is added by the operating system, the application itself, and/or through the use of certain automation features [9]. That means every document shared on the Web has probably some kind of metadata. The question is whether the metadata revealed is harmful or not. For example, the yearbook of 2016, that was published by Estonian Internal Security Service (KAPO), and can be considered as a document made by a very restricted organization, has metadata in it (see Figure 3). Is this metadata harmful or not, one cannot tell, but the fact that metadata exists in documents of such a level is well-proven. It can be read that the document is created with a software named Adobe InDesign and the operating system used was Mac OS.



Properties

Title: None
Location: file:///home/karl/Downloads/Aastaraamat_2016_0.pdf
Subject: None
Author: None
Keywords: None
Producer: Adobe PDF Library 15.0
Creator: Adobe InDesign CC 2017 (Macintosh)
Created: Wed 12 Apr 2017 11:45:05 AM PDT
Modified: Wed 12 Apr 2017 12:26:29 PM PDT
Format: PDF-1.4
Number of Pages: 45
Optimized: No
Security: No
Paper Size: 8.66 × 11.42 inch
Size: 4.8 MB

Figure 3. Metadata of KAPO yearbook

Throughout history there are many cases where seemingly innocuous metadata has caused reputational damage to persons or governments. One famous example is “Dodgy Dossier’s” case in 2003 [10], in which the United Kingdom’s (UK) government placed a report about Iraq’s weapons of mass destruction on its website. The report document was ultimately used by the UK government to justify its involvement in Iraq. The uploaded document was in native MS Word format and contained metadata which showed that the document was drafted by civilians who had plagiarized the information from a university student’s thesis. Furthermore, by deeper analysis of the report document, it was discovered that a large portion of the documents were actually taken from a twelve-year-old PhD thesis [11]. This fact raised some flags about quality, authenticity of the report, and caused reputational damage to the UK government.

Second metadata eye-opening event involved American law firm Venable’s client [12]. Venable was contacted by a company whose vice president had recently resigned. Shortly after his exit, the firm lost a contract with a government organization to a competitor – a competitor working with the former vice president. The vice president of the company was accused of misusing of trade secrets. The defendant and his new firm provided an MS office document (“Sham document”) as evidence for the court; however, they did not take the possible metadata into account. Defendant’s evidence document contained timestamp anomalies: the document was created after the lawsuit was brought to court and it was last saved before it was printed which normally could not happen. Forensic experts discussed that there had been a tool used for editing timestamps of the evidence document. Judge concluded that the document was fraudulent and Venable’s client won the case, receiving 20 million dollars, including sanctions.

The Doggie Dossier and the Venable cases are just a few of the real-world examples for demonstrating that document metadata can contain very sensitive information. Also, embedded spreadsheet shown in Figure 2. proves the fact that metadata can cause problems to people and to corporations. The following chapters give an overview of the most common document formats and their metadata properties.

2.2.1 Metadata in MS office documents

Microsoft Office is the most popular office product in use for corporations and organizations [13]. Applications such as Word, Excel, and PowerPoint, are common applications that generate MS Office documents, spreadsheets, and presentations. In time Microsoft has changed its file formats which affect document structure and characteristics.

Microsoft Office is supporting two types of file formatting for its document creation applications. Microsoft Office versions 1995-2003 used binary format called Object Linking and Embedding (OLE) protocol [14]. In this format, all information is written in streams that are stored in binary file as a linked list of file blocks. With Microsoft Office 2007, MS started to support Office Open XML format (OOXML) [15].

OOXML file format consists of compressed ZIP files called packages. All the contents of the document data, XML-s and other parts, are inside the package [16]. OOXML is an open structure organized in zip archive. Relationship information is used by applications to locate data parts within a package and it is stored inside the package container also.

Microsoft Office supports different file extensions. Microsoft's older versions (until 2003) support *.doc* extension for its documents, *.xls* extension for its spreadsheets, and *.ppt* extension for its presentation documents. Supported file extensions for Microsoft Office 2003 are shown in Table 1[17].

Table 1: Microsoft binary format supported extensions [17]

Word binary format	Extension
Document	.doc
Macro-enabled document	.dot
Microsoft Word Backup Document	.wbk
Exel binary format	Extension
Workbook, spreadsheet	.xls
Template	.xlt
Macro-enabled template	.xlm
PowerPoint binary format	Extension
Presetation	.ppt
Template	.pot
Macro-enabled template	.pps

In following of this thesis paper Microsoft Office OLE format documents are considered as MS binary format documents and the documents with extensions *.doc, .xls, .ppt* are considered as MS binary formatted documents. Other file extensions shown in Table 1 are not in scope of this thesis paper.

Newer formats of Microsoft Office support OOXML file format, which is basically a container file, using industry-standard ZIP format. File extensions of OOXML files are presented in Table 2 [15].

Tabel 2: OOXML file types and extensions [15]

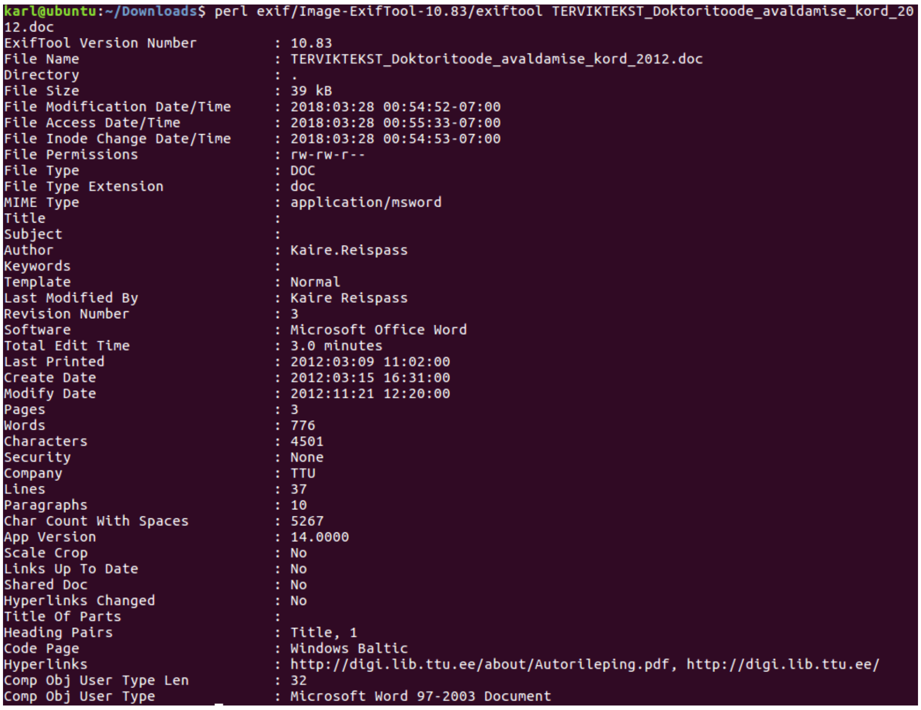
Word XML file type	Extension
Document	.docx
Macro-enabled document	.docm
Template	.dotx

Macro-enabled template	.dotm
Exel XML file type	Extension
Workbook	.xlsx
Macro-enabled workbook	.xlsm
Template	.xltx
Macro-enabled template	.xltm
Non-XML binary workbook	.xlsb
Macro-enabled add-in	.xlam
PowerPoint XML file type	Extension
Presentation	.pptx
Macro-enabled presentation	.pptm
Template	.potx
Macro-enabled template	.potm
Macro-enabled add-in	.ppam
Show	.ppsx
Macro-enabled show	.ppsm
Slide	.sldx
Macro-enabled slide	.sldm
Office theme	.thmx

In the following of this thesis paper file extensions *.docx*, *.xlsx* and *.pptx* are processed; other file extensions are ignored. It is due to the existence of other file extensions being slight on the Web, and them not being very popular document types that are hosted on companies' websites. OOXML format documents in this thesis are considered as documents with extensions *.docx*, *.xlsx* and *.pptx*.

Microsoft Office documents have functionalities to store extra information about themselves, describing the document author, timestamps of when the document was created and edited; also when printed and what application the document was processed with. The easiest way to examine that kind of information is using the application the document was created with, an example is shown in Figure 1. Observing the document metadata of Microsoft Office versions 1995-2003 (MS binary format) is more complicated than with newer version of Microsoft Office documents.

For the MS binary files all the data is written in streams that are stored in the binary file as linked lists of file blocks [2]. Metadata is stored, for the most part, in Summary Information and Document Summary information stream within the file, which means the metadata of MS binary documents are not easily viewed. The main options to see metadata of those types of documents are with hexadecimal viewers or with the application used to create the document. A very good tool for observing metadata is *ExifTool* by Phil Harvey [18] which is platform independent tool working on Perl library. *ExifTool* supports different file formats, including MS OOXML, PDF, and MS Binary formats. Observing metadata of randomly downloaded document on TTÜ website with the extension of .doc (*TERVIKTEKST_Doktoritoode_avaldamise_kord_2012.doc*), *ExifTool* prints out the following output:



```
karl@ubuntu:~/Downloads$ perl exif/Image-ExifTool-10.83/exiftool TERVIKTEKST_Doktoritoode_avaldamise_kord_20
12.doc
ExifTool Version Number         : 10.83
File Name                       : TERVIKTEKST_Doktoritoode_avaldamise_kord_2012.doc
Directory                       : .
File Size                       : 39 kB
File Modification Date/Time     : 2018:03:28 00:54:52-07:00
File Access Date/Time           : 2018:03:28 00:55:33-07:00
File Inode Change Date/Time     : 2018:03:28 00:54:53-07:00
File Permissions                : rw-rw-r--
File Type                       : DOC
File Type Extension             : doc
MIME Type                       : application/msword
Title                           :
Subject                         :
Author                          : Kaire.Reispass
Keywords                        :
Template                        : Normal
Last Modified By                : Kaire Reispass
Revision Number                 : 3
Software                        : Microsoft Office Word
Total Edit Time                 : 3.0 minutes
Last Printed                    : 2012:03:09 11:02:00
Create Date                     : 2012:03:15 16:31:00
Modify Date                     : 2012:11:21 12:20:00
Pages                           : 3
Words                           : 776
Characters                      : 4501
Security                        : None
Company                         : TTU
Lines                           : 37
Paragraphs                      : 10
Char Count With Spaces          : 5267
App Version                     : 14.0000
Scale Crop                      : No
Links Up To Date                : No
Shared Doc                      : No
Hyperlinks Changed              : No
Title Of Parts                  :
Heading Pairs                   : Title, 1
Code Page                       : Windows Baltic
Hyperlinks                      : http://digi.lib.ttu.ee/about/Autorileping.pdf, http://digi.lib.ttu.ee/
Comp Obj User Type Len          : 32
Comp Obj User Type              : Microsoft Word 97-2003 Document
```

Figure 4: Output of ExifTool

Analyzing the same document with Microsoft Word application interface, it does not present all the metadata information, such as application information of document creator, which *ExifTool* is able to show in its output. *ExifTool* is used widely by forensic investigators and it is also used in a popular malware checking website Virustotal. There are many other tools which can be used to investigate metadata of MS binary documents, such as "hachoir-metadata", "libextractor", and "bintext". To make metadata analysis more effective, third-party tools are reasonable to use when extracting metadata from MS binary formats. MS OOXML format is therefore open by its structure and gives many opportunities to observe metadata.

OOXML documents contain two XML files inside their container that contain metadata. Those two XML files are known as *app.xml* and *core.xml*, located in *docProps* directory. OOXML Word document structure and location of *docProps* directory is presented in Figure 5. Insights to *app.xml* and *core.xml* are presented in Appendix 2. *App.xml* contains properties about application which created the document as well as information about keywords, revisions, editing time, etc. *Core.xml* contains properties about the document itself, such as timestamps, author who created and modified it.

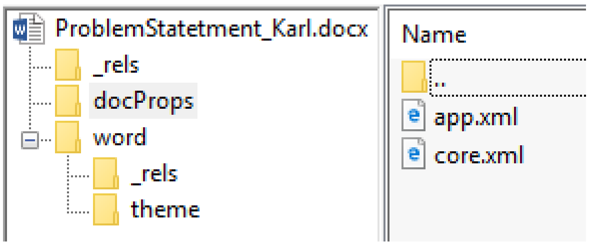


Figure 5: OOXML zip container content

Core structures of OOXML file inside a container vary and it is depending on the document type. Figure 6. shows OOXML container's default structure. Most complicated structure is for presentation files that are generated with PowerPoint application.

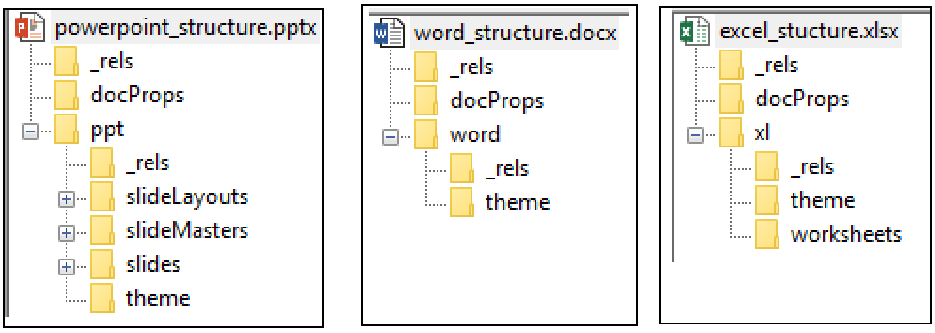


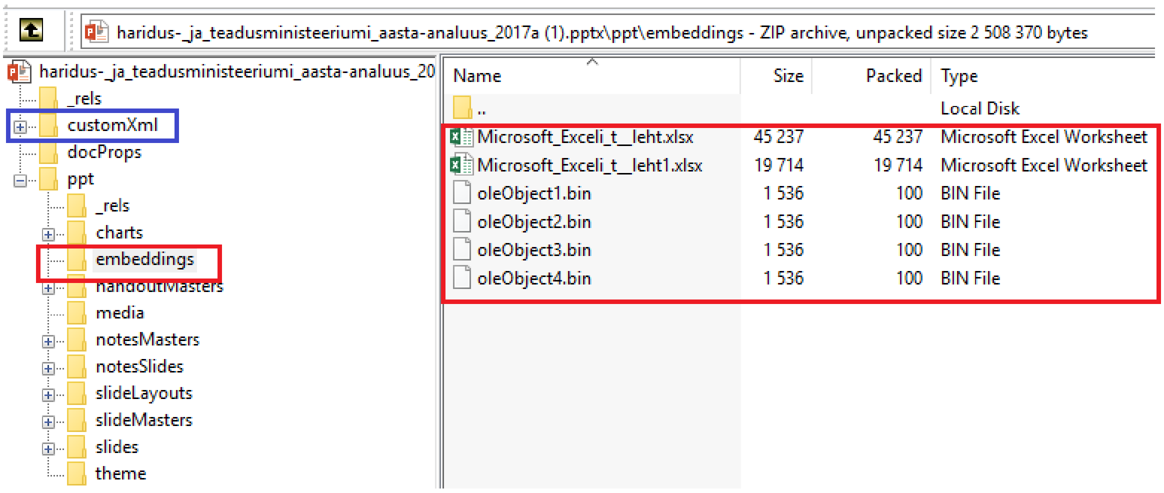
Figure 6: The structures of different types of OOXML files

OOXML documents have features that support business processes and data integration with documents [19]. The feature is called Custom XML and it is very powerful, enabling interoperability with other systems. It has no content restrictions, only syntactical restrictions, meaning it has to be in a well-formed XML format. That means that if the metadata or the Custom XML information is not removed from the document, it could contain compromising information about the organization services which procces that document. As well as a Custom XML feature, OOXML documents support embeddings. In Figure 2 there is one good example of the embedding's feature. In embeddings one can find pictures, videos, other OOXML files, and binary files. Embeddings are usually the result of document author's actions; the problem is, however, that usually the document author is unaware of the existence of embedding in that form.

There are several metadata fields that can be extracted from MS documents. The fields vary depending on the MS document format. The core of the metadata fields is the same in both file formats (OOXML, OLE):

- Creator – The creator or author of the document;
- Created Date – The date when document was created;
- Modified Date – The date when document was modified and saved;
- Application – Application name that created the document;
- App Version – Version of application that created the document;
- Last Modified By – name of the user who modified the document last;
- Company – organization or company which created the document;
- Printer – information about printers which were used for printing the document.

Some of the metadata information is not viewable by Office application interfaces, which means the users are likely not aware about full information that their document contains. This information as mentioned in previous chapters is considered as hidden information. Hidden information can be comments, revision history, and track changes. By copy-pasting charts or graphs to *.pptx* presentation from worksheet, the entire worksheet could be added into the OOXML container, but for the user it is presented only as the graph or chart. Analogous situation is shown in Figure 2. The user might not see the links between the graph and the worksheet and, when sharing the document with other parties, accidently causes data leak. Example of embeddings existent in OOXML document structure is shown in Figure 7. A random presentation document (.pptx) was downloaded from the Ministry of Education and Research webpage. Red rectangles present content of embeddings. One can see that two external Excel Worksheets and four binary object files are in the embeddings .

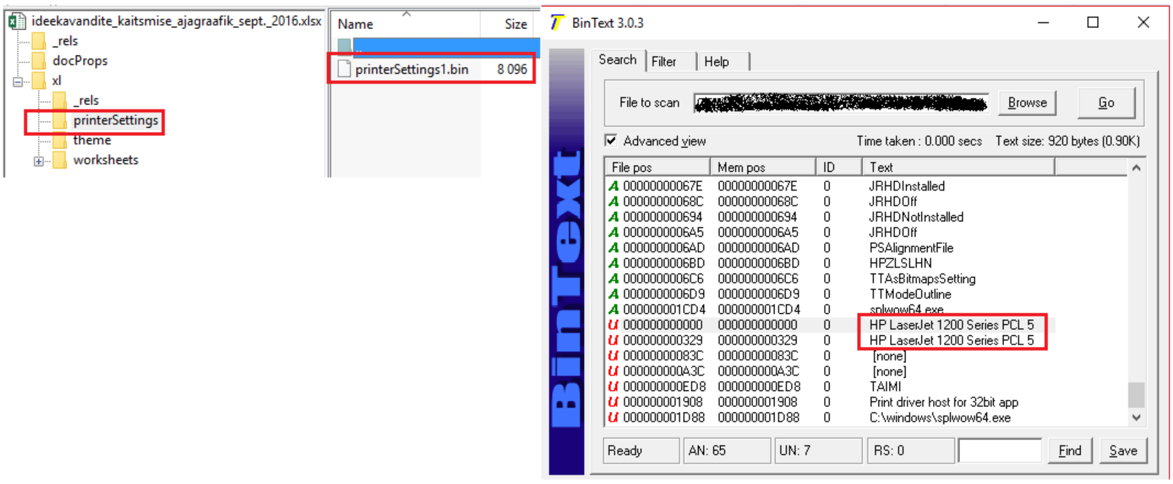


haridus-_ja_teadusministeeriumi_aasta-analuus_2017a (1).pptx\ppt\embeddings - ZIP archive, unpacked size 2 508 370 bytes

Name	Size	Packed	Type
..			Local Disk
Microsoft_Exceli_t__leht.xlsx	45 237	45 237	Microsoft Excel Worksheet
Microsoft_Exceli_t__leht1.xlsx	19 714	19 714	Microsoft Excel Worksheet
oleObject1.bin	1 536	100	BIN File
oleObject2.bin	1 536	100	BIN File
oleObject3.bin	1 536	100	BIN File
oleObject4.bin	1 536	100	BIN File

Figure 7: Example of the existence of embeddings in OOXML document structure

In addition to embeddings and Custom XML features, OOXML and MS binary documents contain printer information about the printer that was used for printing the document. Figure 8. Demonstrates one way how to extract printer information from inside the OOXML file. Printer information is stored in a binary file and it contains the name of the printer and driver information. The binary file can be viewed with hexadecimal viewers. In Figure 8, extracted printer name is *“HP LaserJet 1200 Series PLC5*”.



Name	Size
printerSettings1.bin	8 096

File to scan [redacted]

☑ Advanced view   Time taken : 0.000 secs   Text size: 920 bytes (0.90K)

```
  File pos       Mem pos        ID   Text
A 00000000067E   00000000067E   0    JRHDInstalled
A 00000000068C   00000000068C   0    JRHDOff
A 000000000694   000000000694   0    JRHDNotInstalled
A 0000000006A5   0000000006A5   0    JRHDOff
A 0000000006AD   0000000006AD   0    PSAlignmentFile
A 0000000006BD   0000000006BD   0    HPZLSLHN
A 0000000006C6   0000000006C6   0    TTAsBitmapsSetting
A 0000000006D9   0000000006D9   0    TTModeOutline
A 000000001CD4   000000001CD4   0    splwow64.exe
U 000000000000   000000000000   0    HP LaserJet 1200 Series PCL 5
U 000000000329   000000000329   0    HP LaserJet 1200 Series PCL 5
U 00000000083C   00000000083C   0     [none]
U 000000000A3C   000000000A3C   0     [none]
U 000000000ED8   000000000ED8   0    TAIMI
U 000000001908   000000001908   0    Print driver host for 32bit app
U 000000001D88   000000001D88   0    C:\windows\splwow64.exe
```

Ready   AN: 65   UN: 7   RS: 0

Figure 8: Example of OOXML printer extraction

Metadata is stored in documents for good. It aids in the collaboration and production proccesses of documents among many people. Added comments and track changes features help to produce quality documents. Automatically added date fields help to find and open recently proccessed documents (browsing recent files from Word application) from the quick access of Office applications. The problem with document metadata is that many users are not well-versed in what information is saved with their documents as they share and distribute them.

2.2.2 Metadata in PDF documents

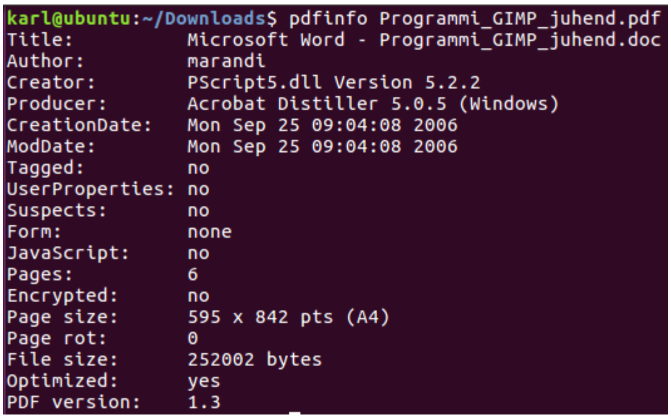
In RFC3778 that describes Portable document format (PDF) has sentenced: “PDF was originally designed as a way to communicate and view printed information electronically across wide variety of computers, devices, and operating systems” [20]. Nowadays it is a popular file format to distribute electronic documents. The extension of PDF document is *.pdf*.

PDF's main goal is to allow users to exchange and view unmodifiable documents. PDF preserves the formatting from the file it was created from, which also makes PDF an excellent file format for sharing and printing. No matter which program, software, device, or operating system is used for opening a PDF file, it always looks the same [20]. Those are some of the reasons why PDF's existence has such a high percentage on the Web.

For document sharing on the web, PDF format is also preferred because of its strict structure and metadata properties. In the PDF generating process, PDF creators normally strip all the hidden information that the original file contained. However, sometimes it is not the case.

Metadata within PDF documents can be stored in two ways [21]: In a document information dictionary or in a metadata stream. Contents of the metadata originating from the document information dictionary, are described below (containing author information, timestamps, etc.). Metadata stream is represented in XML and it is visible in plain text only if the tools are PDF aware. The specific format of XML is defined as Extensible Markup Platform (XMP) [21]. The purpose of this format is to allow different programs to process PDF files and add their own types of metadata information.

Metadata information about the documents can be examined in a way similar to Microsoft documents, using user interface. In Figure 3 there is shown how metadata looks in the user interface when observing a PDF document with the PDF reader application. To see all the information about a PDF document, tools can be utilized. There are two commonly used tools available: *ExifTool* which was introduced in chapter 2.2.1, and command line tool *pdfinfo*. *ExifTools'* output is shown in Figure 4. Tool *pdfinfo* returns similar information as *ExifTool* but sometimes some of the metadata fields are not extracted by *pdfinfo* utility. Output of *pdfinfo* can be seen in Figure 9. Examined PDF document was downloaded from UT website at random. According to metadata, document is quite old (from 2006) and document author is "Marandi".



```
karl@ubuntu:~/Downloads$ pdfinfo Programmi_GIMP_juhend.pdf
Title:          Microsoft Word - Programmi_GIMP_juhend.doc
Author:         marandi
Creator:        PScript5.dll Version 5.2.2
Producer:       Acrobat Distiller 5.0.5 (Windows)
CreationDate:   Mon Sep 25 09:04:08 2006
ModDate:        Mon Sep 25 09:04:08 2006
Tagged:         no
UserProperties: no
Suspects:       no
Form:           none
JavaScript:     no
Pages:          6
Encrypted:      no
Page size:      595 x 842 pts (A4)
Page rot:       0
File size:      252002 bytes
Optimized:      yes
PDF version:    1.3
```

Figure 9: Output of pdfinfo metadata viewer

PDF metadata is added when document is created, modified, or saved. A PDF file can contain metadata such as title, author, producer, and creation and modification timestamps. As shown in output of *pdfinfo*, several metadata fields exist inside pdf documents. The core metadata fields used in this thesis are [21]:

- Author – contains the name of the person who created the document;
- Creator – contains the name of the application that was originally used for creating/converting document to PDF format;
- Producer – contains the name of the application that was used to convert the document to PDF from another format, if conversion took place;
- CreationDate – Contains the date and time when the document was created;

- ModDate – contains the date and time the document was modified.

PDF documents have less capabilities of storing metadata comparing to MS Office documents. However, information that can be stored in metadata fields are still compromising and in certain circumstances can cause problems. The following chapter discusses the risks of metadata.

2.3 Risks

Hidden information in electronic documents can pose serious risks and often people are not aware of that kind of danger. The intention of metadata is to help with document creation, editing, and collaboration: for making it faster and more reliable. But when metadata is ignored, third-party people may get unauthorized access to privileged information that could be used against you or your organization.

The problem is not the existence of metadata in documents, but that it is difficult to fully identify and remove it. Metadata that is left in documents can easily be viewed by anyone with access to these documents. Due to majority of people not being fully aware of the metadata existence, they can unwittingly send confidential information outside of their organization or publish it on the Internet where everyone has access to it. Sharing documents that contains sensitive metadata with co-workers in an internal network seems not a very harmful action, but if one of the co-workers should send that document to a partner company via e-mail, it may be a different story, resulting possible compromise of all the person names and comments of those who were working with that document.

In addition, people and organizations usually do not know when, and for what their document metadata is used, and who uses it. It is impossible to control that flow when documents are outside of the corporate perimeter. Throughout history there are many cases where document metadata has played a significant role, like in court cases, but there is no information available on whether the document metadata has been used in preparation of a cyber-attack or for Open Source Intelligence (OSINT) purposes. In that sense, when talking about metadata risks, the facts that metadata provides useful information to attackers about target organization users, software, and internal services, should be taken into account. Decreasing feasible attack noise, preparing attacks and selecting targets more accurately.

Metadata inside documents comes in many forms and has different values depending on the document format and structure. Understanding the risks and the impact that metadata exposes, each of metadata fields and information they contain has to be discussed separately.

Below there is a list of metadata types that are found in documents (in MS Office and PDF documents) and the risks each type poses in a cyber threat's perspective. Metadata types are chosen considering common metadata properties as well as other fields and information which can be found inside documents.

Document creator/author information

Applies to: Microsoft Office and PDF documents.

The risk: Names of document authors who saved or created the document are stored inside metadata. Saving that kind of information within document poses several risks including sensitive personal information and usernames exposure. Organizations often use first name and surname combination as usernames or as local system credentials. Exposing those names through metadata can raise many threat vectors for the company and it might help conduct brute-force attacks against the organization's services that are available on the Internet (for example webmail, cloud services).

In some organizations, workers names and occupations are hidden and are not publically available on the contact list of the website. Through document author metadata fields those names can possibly leak. In addition, this author information can reveal possible e-mail target lists for sending spare phishing e-mails.

Comparing document creator information with other metadata fields, such as timestamps, company name fields, and software version, gives the attacker knowledge about the software that was used and the time when it was used. If the document modification date says that document was modified yesterday by that person, then one can be quite sure that the victim has that version of software running in its systems.

Document author information exposes and opens plenty of attack vectors which can be used against people and organizations.

File dates and timestamps

Applies to: Microsoft Office and PDF documents.

The risk: When a document is created, modified timestamps about that event are saved into the document's metadata. Releasing this information with the document raises little or no direct security concerns, but it gives descriptive information of document and its contents in that time moment. For example, if a document contains server location or compromising information about the organization, it is possible to determine the time period when this information is accurate. Also if a document was uploaded to a corporate webpage and metadata exposes timestamps and author of the document, then most probably this document author works in that company.

Local and network paths

Applies to Microsoft Office (Word, Excel, PowerPoint) documents.

The risk: Microsoft Office documents have abilities to store local or network paths within them, exposing several risks, including local services and personal information exposure. Personal information is typically found in the file path text. The local and network paths of documents are usually added to the documents where they are modified. If a document is edited in a network share and saved, the file path information can disclose path to that network share. When a document is edited on a local computer then paths refer to a local computer and can disclose username and the operating system.

Network path's information could disclose sensitive information about the internal network, also about document directories or folders. It means that from the network path one can learn how folder naming structure is done- in other words, the directory hierarchy. This information provides a view into corporate network topology which leaves the organization's network open to risk of intrusion.

If a path directory or folder name contains sensitive information, the risk of sensitive information leakage can occur outside of the organization; for example, exposing the names of projects, departments that are doing them, and clients. When a document was edited on a network share, the path to that document can expose potential file server name. For example, an attacker could prepare ransomware to target that server in that organization and this ransomware does not have to scan local networks, because the location of the server is already exposed.

When documents are edited and modified using web applications or document managing platforms, the file path in the metadata can disclose information about the organization's services. Also, if it happens to be an internal service then this information exposes internal

DNS namespace. Internal domain name helps the attacker to conduct more accurate phishing attacks, for example, fake login page with prefilled form (backlash internal DNS name and username).

In addition to network paths, the local path also exposes several risks. Local paths where the document was edited contains full path to that document or to the template. This means it exposes the operating system, logged in user information, as well as hard drive mappings or software names. For example, if a document is edited in Outlook and then shared with other parties, local path information can contain full path to the Outlook cache directory, exposing the organization's use of Outlook as the e-mail application.

The following list presents some examples of path information that Microsoft Office documents contain. Paths are extracted from documents downloaded from microsoft.com website.

- `C:\Users\Luann\Documents\Social\Batch 4\`
- `https://microsoft-my.sharepoint.com/personal/johale_microsoft_com/Documents/New Use Case Templates/_ALL formatted for upload/`
- `C:\Users\IBM_ADMIN\Desktop\Deliverables\DEMO\2015\WA\Inventory & MarketPlace\Data Def & cue docs\Deb & Suman\WIP_2\DONE_ 28Aug\`
- `U:\Misc\`

Printer information

Applies to: Microsoft Office documents. Printer setup information is often stored within a Microsoft Office document.

The risk: Organizations and companies' IT managers usually name printers in a way that they are easily distinguishable from others printers and also by physical locations. Documents that include printer setup information carry a risk of disclosing sensitive printer path information which can contain printer's physical location and model information. Since printer names are described usually in a way that they contain physical location information, for example, "HP MFP printer_second_floor_room23", then this information can be used for exposing the document creator's physical locations. This carries out risks associated with personnel location exposure.

In addition, printer names could contain print server location or file paths that disclose sensitive file path information and provides information about network topology. Attackers can read internal network information without penetrating the systems. Matching printer information with document creator's information, it exposes that document author has permissions and access to that resource. From attackers' perspective, a print server is a valuable target, since many documents from different resources are printed through those servers, which in turns means that a lot of sensitive data might go through them.

Printer setup information can include printer's model name, which represents few concerns; however, this information can be used by attackers sending phishing e-mails with attachments or links refering to infected printer drivers.

The following lines show an example of printer information which can be inside the document. Printer names are extracted from documents downloaded randomly from microsoft.com website:

- `\\red-prn-xrx\b110-3270-a`
- `\\PRN-CORP4.redmond.corp.micro`
- `\\rfrandsen\HP LaserJet 400 M4`

One can read out that one of the printer is located in Redmond and probably document author works there or visited that place in some point.

Application and software information

Applies to: Microsoft Office and PDF documents

The risk: Microsoft and PDF documents store inside themselves information about applications that were used to create them. Software information exposes several risks that can be used for cyber attacks. Firstly, if an attacker knows the softwares that is used, it can help conduct more targeted approach to the victim. If metadata exposes the software's name, version, and timestamps, the attacker can construct malware according to that information, reducing exploit choice and increasing success rate.

Secondly, application names and versions could expose information about the environment where people are working. For example, if metadata says that the document is created with MS applications, most probably the target operating system is Windows. Thirdly, correlating software versions with time, it is possible to determine the update cycle of the organization and find out if outdated software is used.

Below there are some examples of software versions that exist in documents, those in particular are extracted from documents hosted on microsoft.com webpage.

- `Acrobat Distiller 5.0.5 (Windows)`
- `FrameMaker 6.0`
- `pdfTex-1.40.13`
- `Microsoft® Word 2010`
- `PDF-XChange 4.0.193.0 (Windows Seven Ultimate x64 (Build 7600))`

Embedded Objects

Applies to: Microsoft Office documents.

The risk: Microsoft Office allows embeddings, meaning that objects are allowed to be created inside a document. A case of a simple use of embeddings would be when a user is editing a Word document and copies a chart from an Excel document to the Word document. Word will show the user the chart that was copied but underneath the visible Word document contains the Excel worksheet where all the data is stored in a format that can be read by anyone. This feature poses several risks. Firstly, embedded files contain their own metadata which can be extracted. Secondly, the embedded Excel table might contain sensitive information and is meant for corporate use only. Also, that table might be originating from secret networks, which exposes a high risk for the organization.

Thirdly, the risk of reusing embedded objects or OOXML documents can occur. That means the attacker could send a prepared attachment to the document creator or any other targeted personnel and have the same table (which is for corporate internal use only) attached, thereby infecting the user computer with malware and for target user perspective it seems very truthful. Example of embedded objects is shown in Figure 7.

Custom information

Applies to: Microsoft Office documents

The risk: Custom properties are often used by applications to associate metadata with a document. For example, document management systems could use custom properties to assist document categorization or some additional information. Depending on the implementation, information that can be in custom properties could range from innocuous to highly sensitive. Also, custom information could contain descriptive information about internal services. The following example presents some of the information that existed in custom metadata fields:

- <Client_x0020_E-mail **david.appel@microsoft.com** </Client_x0020_E-mail>
- <Account_x0020_Contact_x0020_Mobile_x0020_Phone **425-233-2120** </Account_x0020_Contact_x0020_Mobile_x0020_Phone
- <Account_x0020_Contact **Erin Arnold** </Account_x0020_Contact>

Documents were downloaded from Microsoft.com website; the custom information exposes phone numbers, contact names, and e-mail addresses.

Document Properties

Applies to: Microsoft Office documents, including PDF documents.

The risk: Document properties are details about the document that help identify it. Document properties contain usually several fields, such as title, subject, author, manager, company, keywords, and comments. For this thesis paper we exclude author and manager information from Document Properties, since the risks those fields can expose are discussed already in previous points.

Document properties generally presents few risks. This is because they are a mirror of some visible content from the document. However, some of the metadata fields that the document properties contain might expose some risks. Field named "Company" helps to bind the document with a certain organization, meaning if the document is found somewhere in the Web, company field could possibly indicate where that document is originating from. In some cases, company name field exposes internal domain namespace.

Comments information exposes personal information exposure if the comments are not removed. Comments are usually meant for collaboration and, if released, can leak information that was not intendent to be there. For example, descriptions of some internal services or references. The severity of this threat depends highly on the content of comments.

For example, some of the document property fields, extracted from documents that were downloaded from microsoft.com webpage:

- <Company>Microsoft</Company>
- <Company>Infosys Technologies Limited</Company>

2.4 Related work

Jeffery R. Jones introduced in his research paper documents' metadata and the security issues metadata can cause [2]. The paper gives an overview of metadata and its fields in different types of documents, such as Microsoft Office, OpenOffice, and PDF documents. Paper also introduces tools and places where forensic investigators can find information for investigation. Jeffery R. Jones concluded that examination of documents metadata can lead to discovery of the following information: documents' author names; names of contributors as well as their recommended changes and comments; network storage path locations, user IDs of the document author; as well as computer specific information, such as the GUID [2].

Larry Pesce from the SANS institute published a whitepaper which introduced metadata extraction and information gathering approaches [22]. The paper discussed that information gathering can be done by documents metadata analysis. Those electronic documents can be found from among documents on public websites, from e-mail, or using Google Search. The author described how to utilize Google search engine for finding documents on targeted websites and how to use Google search engine operators for exposing sensitive information.

He concluded that document metadata has a valuable place in information gathering and auditing programs, and most organizations do not realize that they have some form of exposure.

In 2009 Chema Alonso and Enrique Rando described in their whitepaper the tools and techniques how to fingerprint an organization [23]. The structure of that whitepaper is similar to Larry Pesce's paper [22], but some additional techniques are described as well. In general, the whitepaper gives a very good overview how to extract metadata from Office and OpenOffice documents and what tools and techniques to use for information gathering. In addition, the authors introduced a tool called FOCA which stands for "Fingerprinting Organization with Collected Archives". It is an automated tool for downloading documents published on websites, extracting metadata, and analyzing data.

A detailed overview of risks of metadata and hidden information is described in Oracle's whitepaper, which was published in 2007 [5]. Oracle performed a study to educate users and organizations about the risks associated with information that is commonly exposed when documents are shared. The methodology of this study was downloading documents from randomly selected websites and analyzing metadata of those documents. This was followed by pinpointing the issues found in the documents using a study format containing five categories: Target Element Name, Description, Risk, Study Findings, and Recommendation. Oracle suggested some implementation opportunities to clean documents from metadata.

Hanno Langweg from Norwegian Information Security Laboratory published a paper where he examined Microsoft Office document metadata [24]. He conducted the "July 22nd Terrorist Manual" analysis to determine if style changes can be spotted in text which would indicate different authorship. The author checked revision numbers, changes in formatting, keyboard layout changes, language of metadata paragraphs, and generated of table of contents. The methods described in the paper introduced a new angle how to analyze metadata in Office documents, even when there are no document properties available.

Muhammad Ali Raffay described in his thesis how to hide and detect data in Microsoft Office files [25]. In other words, stenography using MS OOXML files was introduced. The paper gives a very detailed overview of OOXML structure and its capabilities. Due to the structure of OOXML files, extra information can be added inside the document structure and it is not detected by the application that opens it. If extra data is inserted, for example, inside an xml file that is part of the OOXML file, the end user cannot notice the presence of extra information. At the end of this thesis paper an algorithm which detects stenography inside OOXML documents was introduced.

Simson L. Garfinkel introduced in his paper how to recover hidden information from Office files [26]. Complex document formats such as Microsoft formats and PDF can contain information that is hidden but recoverable. This can be the result of embedding files, cropping pictures, highlighting text, or adding media files into documents. The paper included examples of privacy leakages in history that were caused by metadata. Microsoft Office has a tool called "Inspector" which finds and removes all sensitive metadata. However, according to this paper it is not enough for removing all the sensitive information. According to Simson L. Garfinkel, one solution to mitigate metadata privacy issues and exposures is to modify tools so that underlying data model is in line with what is presented in the user interface – in that way it is harder for the end users to produce documents which contain hidden information.

Randal Farrar stated in his paper that every Microsoft Office document contains some kind of metadata [9]. Every time a document is opened, edited, and saved, metadata is added by the operating system, the application itself, and through the use of certain automation features. If metadata removal procedures are not in place in organizations, it is a very high probability to gain sensitive or harmful information from documents processed by those organizations or people. To solve metadata issues, a Metadata policy has to be implemented in organizations that involves several topics, including educating people about metadata.

A very large study was conducted by a group of people [3] where they analyzed over 15 million distinct documents downloaded from the Internet. The motivation for the research was to identify social cliques of users that collaborate in the production of documents by correlating the document author field found in document metadata. In addition, the extracted amount of metadata showed that the existence of metadata in documents is relatively frequent. The study highlighted several privacy risks involved in sharing documents that carry sensitive metadata information.

The current chapter gave an overview of metadata terminology and presented where metadata in documents can be found and which tools to use. In addition, the risks that metadata exposure could bring were discussed and overview of related works in this field was described. The following chapter describes methods on how to fingerprint an organization using the documents hosted on target organization’s webpage.

3 Methodology for conducting metadata analysis of publically available documents

Chapter 3 discusses the contribution made in this thesis by introducing fingerprinting method in subsection 3.1 for gathering documents from public websites and conducting metadata analysis.

Document collection and metadata analysis aims to validate or disproof the hypothesis set in this thesis about whether documents contain compromising metadata for conducting attacks against governmental entities and whether there is a possibility to understand the organization's internal processes and services.

The explained method was used against certain organizations' websites to validate the hypothesis. The results and analysis will be presented in Chapter 4.

3.1 Fingerprinting method

Fingerprinting method consist of three logical stages: document collecting, metadata extraction, and metadata analysis. Workflow of those stages is presented in Figure 10.

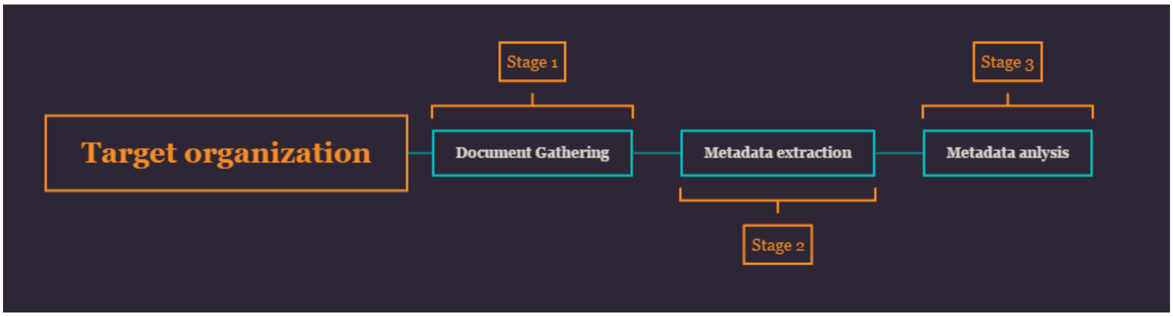


Figure 10: Fingerprinting method workflow

The first stage, document gathering aims to collect electronic documents from selected websites. Document gathering is done by using search engines' functionalities such as search operators. Utilizing search engines functionalities we are able to determine if the data set is available and exists for downloading. Search engines' queries can be specified for finding documents with certain file extensions. The returned query results from the search engines are downloaded with a web browser plugin such as Download Manager. We did not use in this thesis any of automated tools that automatically scrap the documents from the websites, nor custom scripts. It is mainly because of the issues with websites' integrity and availability which may occur when scanning websites. In addition, we do not visit the webpages manually and search for documents. All the document downloads were done based on search engine queries. To increase the document findings from the webpages, multiple search engines were used (Google, Yandex, Bing). The following document extensions were downloaded: *pdf, doc, docx, xls, xlsx, ppt, pptx.* Duplicate documents were deleted using diff function and MD5 hash function.

The second stage of the fingerprinting method is metadata extraction. The first prerequisite of metadata extraction is the existence of documents that were gathered in stage one. All the collected documents are examined using different tools, including manual examination. Each document format is analyzed separately:

- PDF documents are analyzed with *ExifTool*;
- MS Binary documents (*.doc, .xls, .ppt*) are analyzed with *ExifTool* and FOCA;

- OOXML documents (*.docx, xlsx, pptx*) are analyzed with *ExifTool*, *FOCA* and with manual examination.

Metadata grabbing with *ExifTool* is automated with bash scripts. Due to the structure of OOXML documents, novel techniques are used for extracting sensitive information, utilizing manual examination. Metadata fields which *ExifTool* is capable of extracting (author information, timestamps, versions, etc.) are stored into local elasticsearch database for further processing. Some of the metadata properties in documents are ignored (for example, keywords, number of words, titles). The aim of extraction is to gather all that information that can cause dataleaks about the organization and its assets.

The third stage of fingerprinting method is metadata analysis. Extracted metadata in previous stage is analyzed manually and the aim is to identify the targeted organization's assets which aid in conducting cyber attacks and exposing internal information.

In the following subsections all the three stages are described more deeply.

3.2 Stage 1 - Document collecting

It is necessary to gather large collections of documents in order to carry out the metadata analysis [3]. Metadata analysis can be done when the documents are stored into local systems (workstations, servers). The more documents we have, the more opportunities to gain sensitive information from metadata. In a classic penetration framework this document collecting stage is called reconnaissance phase [27]. This phase is usually the first step for attackers, including penetration testers, to gather information about the target and its systems. It is a starting point to attackers, giving them ideas and knowledge about who their victims are. In our case, the starting point are the electronic documents which are uploaded to the target organization's website.

In general, we assume that there is no direct access to the websites' files directory nor administrative privileges on the victim's webserver. That means files which are hosted on corporate websites have to be gathered some other way. We try to utilize an approach where documents are collected remotely without any extra permissions from the website, in the same conditions as a potential attacker or penetration tester would have. Since visiting the website and scrolling through all subpages is a time consuming approach, we considered to use the help of search engines.

Document gathering is done by using search engines and their functionalities of finding documents. For increasing document findings we use three different search engines. Figure 11. presents the overview of the document collecting structure.

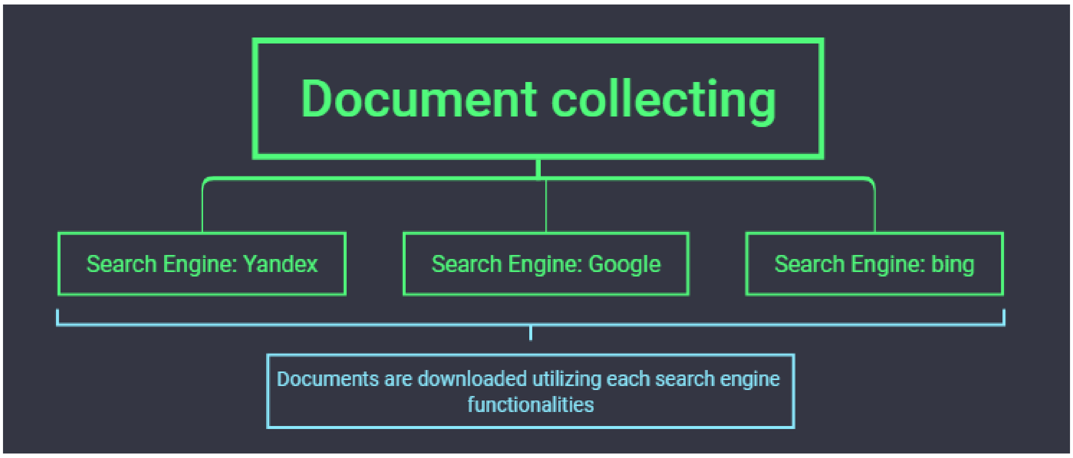


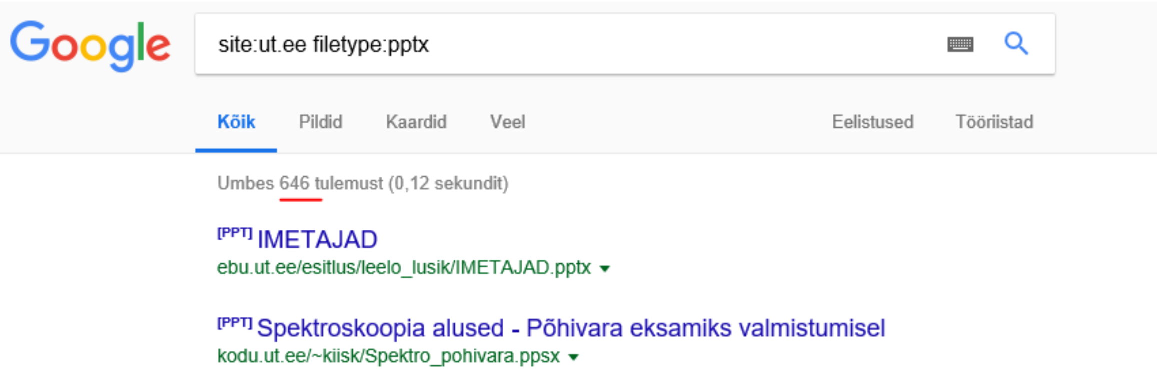
Figure 11: Document gathering done using three different search engines separately

Nodes shown in Figure 11 present search engines that were used in this study. We decided to use three different search engines and the decision was made upon statistical examination of randomly generated query results. Those three search engines shown in Figure 11 returned most responses in our examination. Documents are queried using each search engine separately and the results of the query responses are downloaded into local systems where the duplicates are removed after wards. Below, overview of the search engines and conducted queries is presented.

Search engines

When a document is uploaded in an official corporate website, it becomes available for people as well as software systems – search engines. There are many search engines available, such as Google, Yahoo, Yandex, Bing, Exlead, etc. [28]. Each of them has their own specialty but by design they are software systems which help search information from the Web. Search engines crawl through the Web using web spiders and index wide range of parameters, for instance keywords and backlinks.

When searching the Web with Google search engine, it returns in most cases thousands or even millions of responses if the desired object is indexed. Those results can be optimized by using search operators the search engines provide. Google has several search operators which are useful for conducting information gathering. For example, if one wants to find files from University of Tartu (UT) website with extension "pptx", Google search operators "*site*" and "*filetype*" can be utilized. Results of this conducted query is shown in Figure 12. This method is called in other words "Google Hacking"[26]. This is a powerful functionality for searching documents with certain file extensions. Manually scrolling through UT webpage and its subpages, searching for a *"pptx"* files would take a lot of time, but with a search engine it took only 0.12 seconds (Figure 12).



site:ut.ee filetype:pptx

Umbes 646 tulemust (0,12 sekundit)

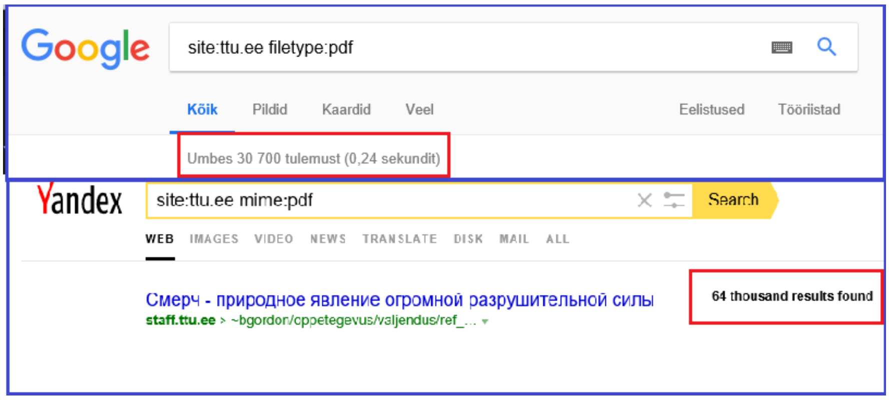
[PPT] IMETAJAD
ebu.ut.ee/esitlus/leelo_lusik/IMETAJAD.pptx

[PPT] Spektroskoopia alused - Põhivara eksamiks valmistumisel
kodu.ut.ee/~kiisk/Spektro_pohivara.ppsx

Figure 12: Google search operators

Google search operator "site" finds results associated with the domain name which is passed in for the query, including also the subdomains. Figure 12 presents that Google found *pptx* files from the UT subdomains. If a targeted website has aliases then it is necessary to query documents from all of its top level domains for maximizing document findings. Downloading process for the search engine queries is quite simple, it is only needed to click on the link the search engine finds. However, downloading manually by clicking links the search engines respond with is a quite time consuming activity. For optimizing download process, web browser plugins, such as Download Managers, are used. Download Managers download all the results that are found by the search engine.

Since each search engine is different by design and by web indexing capabilities, three separate search engines are used for information gathering. For example, searching PDF documents from TTU webpage, Yandex returns 64000 results but Google only 30700, which is less than half of what Yandex has found. Figure 13 presents mentioned example; response amounts are shown in red rectangles.



site:ttu.ee filetype:pdf

Umbes 30 700 tulemust (0,24 sekundit)

site:ttu.ee mime:pdf

Смерч - природное явление огромной разрушительной силы
staff.ttu.ee > ~bgordon/cppetegevus/valjendus/ref_...

64 thousand results found

Figure 13: Yandex vs Google

Each search engine was used separately for querying the Web with specific search operators. Search operators and queries that were used are described below.

Google:

```
site:targetsite.ee filetype:docx
site:targetsite.ee filetype:doc
site:targetsite.ee filetype:xlsx
site:targetsite.ee filetype:xls
site:targetsite.ee filetype:pptx
site:targetsite.ee filetype:ppt
site:targetsite.ee filetype:pdf
```

Yandex:

```
site:targetsite.ee mime:docx
site:targetsite.ee mime:doc
site:targetsite.ee mime:xlsx
site:targetsite.ee mime:xls
site:targetsite.ee mime:pptx
site:targetsite.ee mime:ppt
site:targetsite.ee mime:pdf
```

Bing:

```
site:targetsite.ee filetype:docx
site:targetsite.ee filetype:doc
site:targetsite.ee filetype:xlsx
site:targetsite.ee filetype:xls
site:targetsite.ee filetype:pptx
site:targetsite.ee filetype:ppt
site:targetsite.ee filetype:pdf
```

The search operator “site:targetsite.ee” is fictive and has to be replaced with the targeted organization’s domain name, for example “site:ut.ee” if UT would be our target. All the downloaded documents are compared with MD5 hash function and the duplicates are removed.

Following subchapter presents metadata extraction techniques and approaches, in order to extract compromising information out of the documents.

3.3 Stage 2 - Metadata extraction

The method of conducting metadata extraction from documents was decided upon examination of specifications of each document format, as well as manual examination of hundreds of publically available documents on the Internet. Review of literature and investigation of different documents gave an understanding that there is no single tool available on the Internet that can possibly extract all meaningful metadata and hidden information for fingerprinting purposes. That means there is no single solution for extracting metadata, therefore, some other approaches are needed in order to carry out this study. In Figure 14, an overview of metadata extraction steps is presented. Documents are analyzed on the basis of document format, depending on which different tools were utilized.

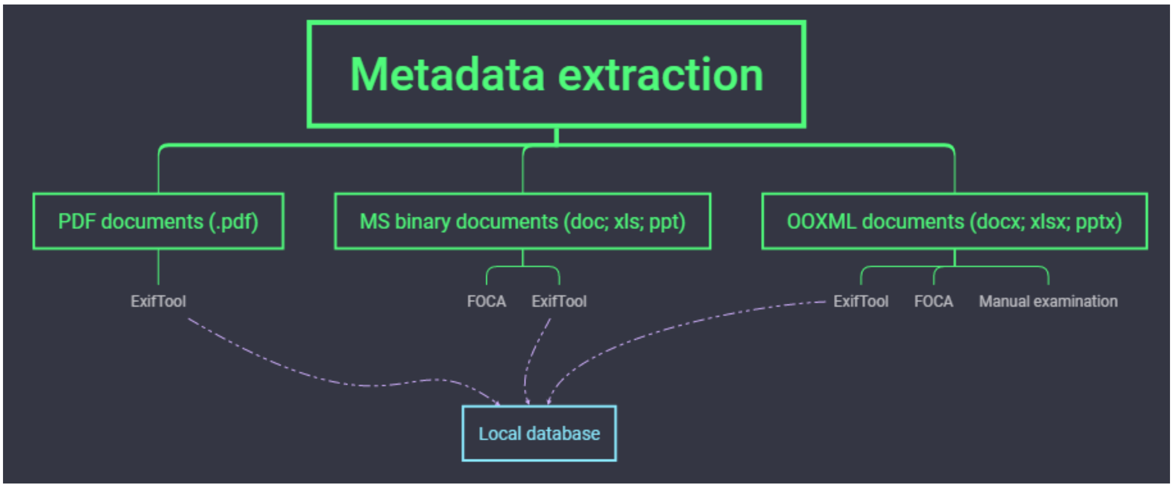


Figure 14: Metadata extraction

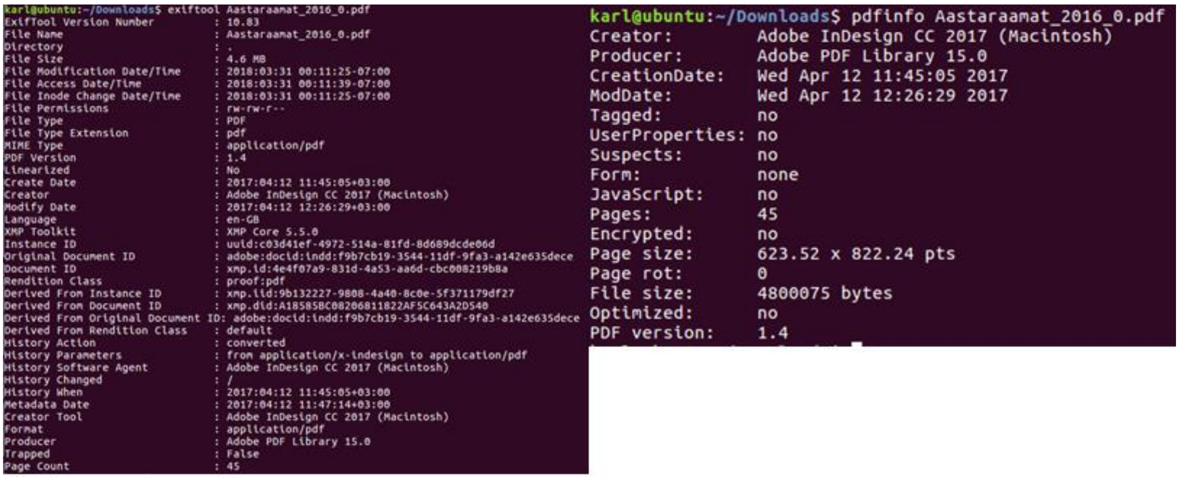
All of the document formats are analyzed using *ExifTool* whose output is saved into local elasticsearch database. Metadata fields which *ExifTool* extracts and stores into database, differ slightly by each file format. Those extracted metadata fields are introduced in the following sections. In addition to *ExifTool*, FOCA is used for analyzing MS Office documents. The aim of using FOCA is to extract printer information from MS Office documents. FOCA has functionalities to analyze and extract all the document types that were discussed before, but the tool misses a lot of information, especially when considering OOXML documents. In addition, OOXML documents are analyzed utilizing manual examination approaches. Novel methods on how to conduct OOXML document analysis are introduced in Chapter 3.3.3.

The aim of this stage is to scrap as much compromising information from the documents as possible. We want to fingerprint an organization, that means all information related to that organization is important. Next subchapters present methods and techniques on how metadata extraction is done in this study.

3.3.1 PDF documents

The amount of documents each website contains can range from one to thousands. This means that looking at documents' metadata by using the application interface and checking the properties there can be very time consuming. For faster metadata extraction third-party tool is used.

There are several useful tools available which help to extract metadata from PDF documents. We decided to use *ExifTool* for analyzing PDF documents because of the tool's capability to extract more metadata and present it in a readable output. Figure 15 shows metadata extraction comparison between *ExifTool* and *pdfinfo* (*pdfinfo* was introduced in chapter 2). The following figure shows comparison between *ExifTool* and *pdfinfo* default outputs.



```
karl@ubuntu:~/Downloads$ exiftool Aastaraamat_2016_0.pdf
ExifTool Version Number         : 10.83
File Name                       : Aastaraamat_2016_0.pdf
Directory                       : .
File Size                       : 4.6 MB
File Modification Date/Time     : 2018:03:31 00:11:25-07:00
File Access Date/Time           : 2018:03:31 00:11:39-07:00
File Inode Change Date/Time     : 2018:03:31 00:11:25-07:00
File Permissions                : rw-rw-r--
File Type                       : PDF
File Type Extension             : pdf
MIME Type                       : application/pdf
PDF Version                     : 1.4
Linearized                      : No
Create Date                     : 2017:04:12 11:45:05+03:00
Creator                         : Adobe InDesign CC 2017 (Macintosh)
Modify Date                     : 2017:04:12 12:26:29+03:00
Language                        : en-GB
XMP Toolkit                     : XMP Core 5.5.0
Instance ID                     : uuid:c03d41ef-4972-514a-81fd-8d689dcde06d
Original Document ID            : adobe:docid:indd:f9b7cb19-3544-11df-9fa3-a142e635dece
Document ID                     : xmp.id:4e4f07a9-831d-4a53-aa6d-cbc008219b8a
Rendition Class                 : proof:pdf
Derived From Instance ID        : xmp.iid:9b132227-9808-4a40-8c0e-5f371179df27
Derived From Document ID        : xmp.did:A18585BC08206811822AF5C643A2D540
Derived From Original Document ID: adobe:docid:indd:f9b7cb19-3544-11df-9fa3-a142e635dece
Derived From Rendition Class    : default
History Action                  : converted
History Parameters              : from application/x-indesign to application/pdf
History Software Agent          : Adobe InDesign CC 2017 (Macintosh)
History Changed                 : /
History When                    : 2017:04:12 11:45:05+03:00
Metadata Date                   : 2017:04:12 11:47:14+03:00
Creator Tool                    : Adobe InDesign CC 2017 (Macintosh)
Format                          : application/pdf
Producer                        : Adobe PDF Library 15.0
Trapped                         : False
Page Count                      : 45
```

```
karl@ubuntu:~/Downloads$ pdfinfo Aastaraamat_2016_0.pdf
Creator:        Adobe InDesign CC 2017 (Macintosh)
Producer:       Adobe PDF Library 15.0
CreationDate:   Wed Apr 12 11:45:05 2017
ModDate:        Wed Apr 12 12:26:29 2017
Tagged:         no
UserProperties: no
Suspects:       no
Form:           none
JavaScript:     no
Pages:          45
Encrypted:      no
Page size:      623.52 x 822.24 pts
Page rot:       0
File size:      4800075 bytes
Optimized:      no
PDF version:    1.4
```

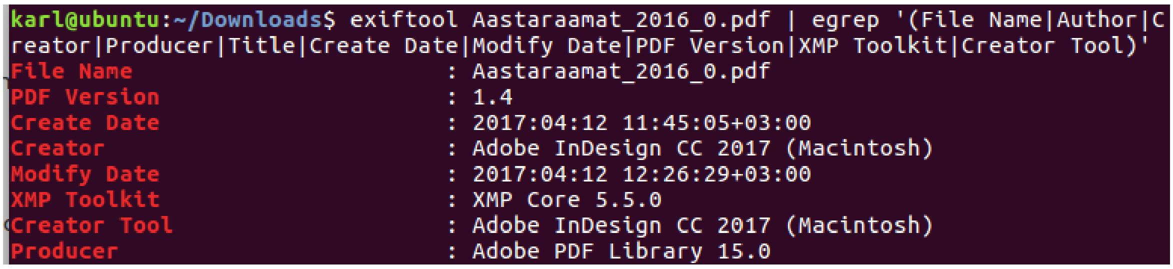
Figure 15: *ExifTool* and *pdfinfo* output comparison

It can be noticed that *ExifTool* prints more information to its output, for example, metadata fields such as "XMP toolkit" and "Creator Tool". The difference is that *pdfinfo* does not print on its output default information that is in the metadata stream. This is the reason why we prefer *ExifTool* in this research.

From *ExifTool* output, metadata fields which are meaningful for fingerprinting purposes are extracted to a local database, other fields are ignored. The following metadata fields are extracted:

- Author
- Creator
- Producer
- Title
- Created Date
- Modify Date
- PDF version
- XMP Toolkit
- Creator Tool

For filter out those metadata fields from *ExifTool* output, command-line search utility *egrep* is used. Figure 16 shows an example of filtered metadata fields from the output of *ExifTool* which is shown in Figure 15.



```
karl@ubuntu:~/Downloads$ exiftool Aastaraamat_2016_0.pdf | egrep '(File Name|Author|C
reator|Producer|Title|Create Date|Modify Date|PDF Version|XMP Toolkit|Creator Tool)'
File Name                       : Aastaraamat_2016_0.pdf
PDF Version                     : 1.4
Create Date                     : 2017:04:12 11:45:05+03:00
Creator                         : Adobe InDesign CC 2017 (Macintosh)
Modify Date                     : 2017:04:12 12:26:29+03:00
XMP Toolkit                     : XMP Core 5.5.0
Creator Tool                    : Adobe InDesign CC 2017 (Macintosh)
Producer                        : Adobe PDF Library 15.0
```

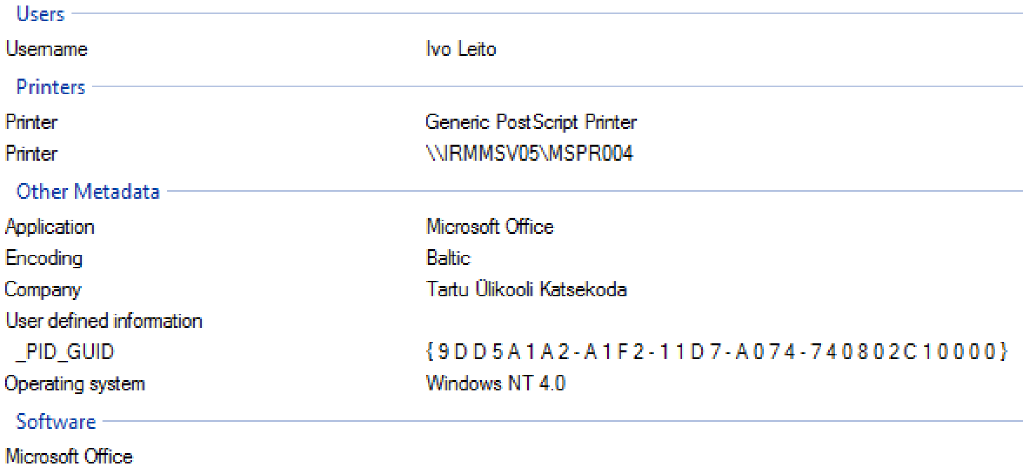
Figure 16: Selecting metadata fields using *egrep*

For automating the metadata extracting process, bash script is used for scanning thru all the downloaded PDF documents and extracting the metadata fields which were discussed previously. The data is saved into local elasticsearch database for further analysis.

3.3.2 MS Office older format (doc, xls, ppt)

MS Office older versions support binary format where all the data is written in streams and it is stored in a binary file. To examine parts or contents of that binary file, some extra tools are needed for observation. As mentioned in chapter 2.2.2, several tools are available that can be utilized for extracting metadata from MS binary formats.

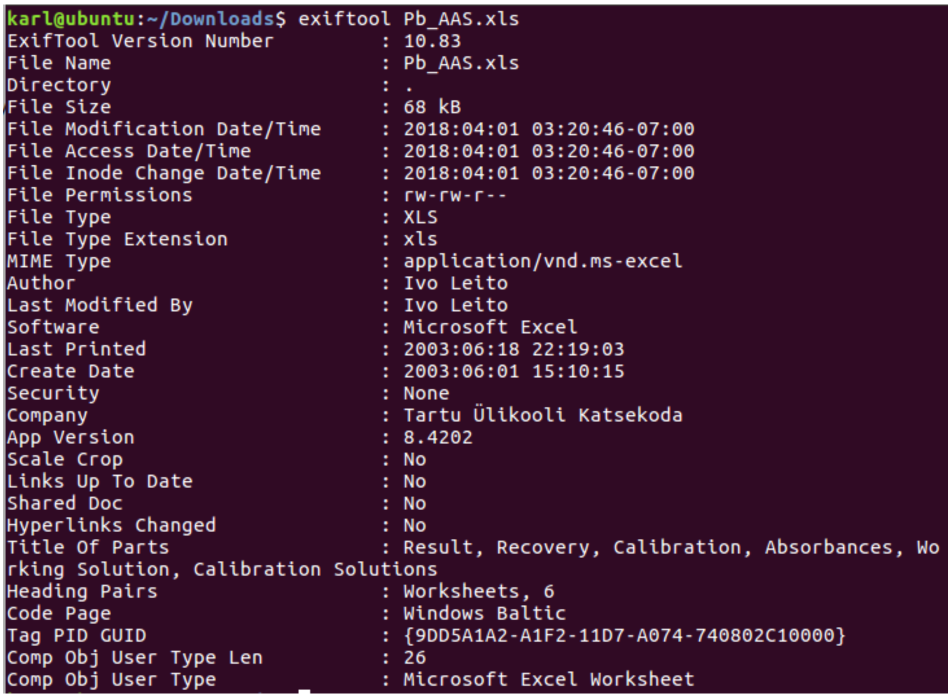
In this thesis two third-party applications are chosen for extracting metadata from the MS binary formatted documents. First tool is *ExifTool* already introduced in previous chapters, and second tool is FOCA. FOCA is used for extracting printer information from MS binary format documents. Figure 17 presents metadata information that is extracted with FOCA from a document randomly downloaded from the UT webpage. FOCA managed to extract printer information form the *.xls* document as well as operating system information.



Users	
Username	Ivo Leito
Printers	
Printer	Generic PostScript Printer
Printer	\\IRMMSV05\MSPR004
Other Metadata	
Application	Microsoft Office
Encoding	Baltic
Company	Tartu Ülikooli Katsekoda
User defined information	
_PID_GUID	{9DD5A1A2-A1F2-11D7-A074-740802C10000}
Operating system	Windows NT 4.0
Software	
Microsoft Office	

Figure 17: Metadata extraction from spreadsheet using FOCA

However, FOCA sometimes misses timestamps of when the document was created, modified, or printed. Timestamp information is important for understanding, for example, what time that version of software was used in the company. Also, timestamp information gives an overview of the relevance of the extracted information. For information as such, *ExifTool* tool is used. Figure 18 shows *ExifTool's* output of the extracted information from the MS binary format document.



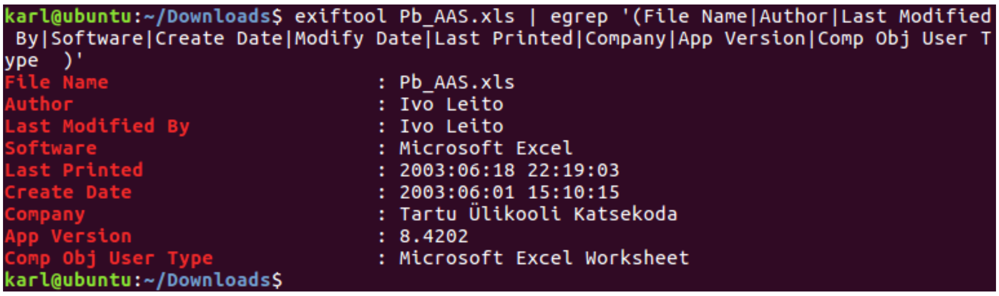
```
karl@ubuntu:~/Downloads$ exiftool Pb_AAS.xls
ExifTool Version Number         : 10.83
File Name                       : Pb_AAS.xls
Directory                       : .
File Size                       : 68 kB
File Modification Date/Time     : 2018:04:01 03:20:46-07:00
File Access Date/Time           : 2018:04:01 03:20:46-07:00
File Inode Change Date/Time     : 2018:04:01 03:20:46-07:00
File Permissions                : rw-rw-r--
File Type                       : XLS
File Type Extension             : xls
MIME Type                       : application/vnd.ms-excel
Author                          : Ivo Leito
Last Modified By                : Ivo Leito
Software                        : Microsoft Excel
Last Printed                    : 2003:06:18 22:19:03
Create Date                     : 2003:06:01 15:10:15
Security                        : None
Company                         : Tartu Ülikooli Katsekoda
App Version                     : 8.4202
Scale Crop                      : No
Links Up To Date                : No
Shared Doc                      : No
Hyperlinks Changed              : No
Title Of Parts                  : Result, Recovery, Calibration, Absorbances, Wo
rking Solution, Calibration Solutions
Heading Pairs                   : Worksheets, 6
Code Page                       : Windows Baltic
Tag PID GUID                    : {9DD5A1A2-A1F2-11D7-A074-740802C10000}
Comp Obj User Type Len          : 26
Comp Obj User Type              : Microsoft Excel Worksheet
```

Figure 18: Output of ExifTool

From *ExifTool*'s output we gather the metadata fields that are important for fingerprinting purposes. Those fields are:

- Author;
- Last Modified By;
- Software;
- Create Date;
- Modify Date;
- Last Printed;
- Company;
- App Version;
- Comp Obj User Type

Named metadata fields are filtered out and saved into the database. The approach is the same with PDF documents where command-line utility *egrep* is used. Extracting process is automated with bash script which searches MS binary files from the downloaded folders and extracts metadata when specific documents are found. Figure 19 presents filtered metadata fields's output that is saved into the database.



```
karl@ubuntu:~/Downloads$ exiftool Pb_AAS.xls | egrep '(File Name|Author|Last Modified
 By|Software|Create Date|Modify Date|Last Printed|Company|App Version|Comp Obj User T
ype  )'
File Name                       : Pb_AAS.xls
Author                          : Ivo Leito
Last Modified By                : Ivo Leito
Software                        : Microsoft Excel
Last Printed                    : 2003:06:18 22:19:03
Create Date                     : 2003:06:01 15:10:15
Company                         : Tartu Ülikooli Katsekoda
App Version                     : 8.4202
Comp Obj User Type              : Microsoft Excel Worksheet
karl@ubuntu:~/Downloads$
```

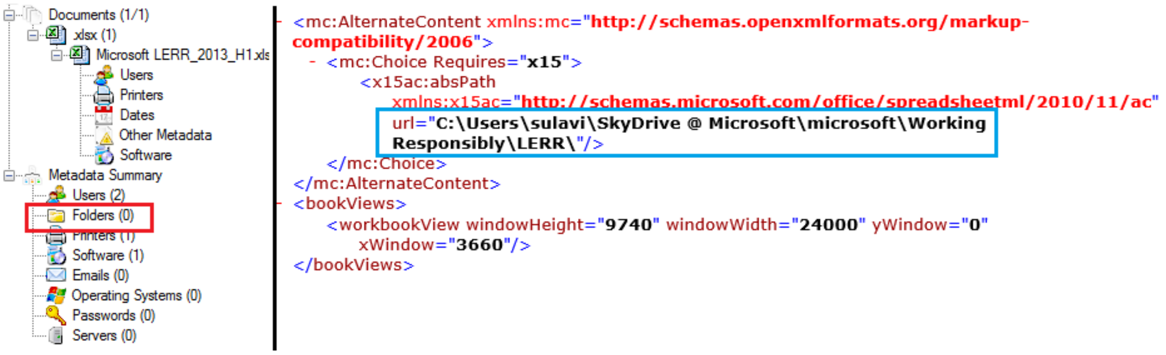
Figure 19: Exiftool filtered output of *.xls* file

Printer information is extracted from MS binary documents using FOCA tool. The results of FOCA output is analyzed manually.

3.3.3 MS OOXML format (docx, xlsx, pptx)

Microsoft OOXML format is more accessible compared to older MS binary formats. When documents are saved in this format, metadata can be viewed and accessed by viewing the source XML files with XML or text editors.

Most of the forensic tools and metadata extractors read information only from *core.xml* and *app.xml* which miss a lot of information. Since metadata can be propagated automatically into document during its creation/modifying process or by other applications, then custom information could exists very often in other parts of OOXML files. For example, a random spreadsheet document was downloaded from microsoft.com website containing path information that was not located in *app.xml* or *core.xml*. Figure 20 shows metadata extraction using the tool FOCA which did not detect any file paths inside OOXML file (marked with red rectangle). However, the document contained a file path in its *workbook.xml* file which is a part of the OOXML container. The file path was extracted manually by observing contents of the OOXML files; the path is presented in Figure 20 inside the blue rectangle.



```
- <mc:AlternateContent xmlns:mc="http://schemas.openxmlformats.org/markup-
compatibility/2006">
  - <mc:Choice Requires="x15">
      <x15ac:absPath
        xmlns:x15ac="http://schemas.microsoft.com/office/spreadsheetml/2010/11/ac"
        url="C:\Users\sulavi\SkyDrive @ Microsoft\microsoft\Working
        Responsibly\LERR\"/>
    </mc:Choice>
</mc:AlternateContent>
- <bookViews>
    <workbookView windowHeight="9740" windowWidth="24000" yWindow="0"
      xWindow="3660"/>
</bookViews>
```

Figure 20. Extraction of local file path

In this thesis, all kinds of information describing file paths or softwares, or any information that is useful for fingerprinting purposes, is essential. To maximize meaningful metadata extraction, it should be taken into consideration that different approaches are needed for getting all the important data from documents. Figure 21 shows the workflow of metadata extraction from OOXML files. Extraction process is logically divided into three phases:

- Common metadata extraction using *ExifTool*;
- Searching specific directories, which indicates the existence of custom/compromising information;
- Conducting string searches from unpacked OOXML files.

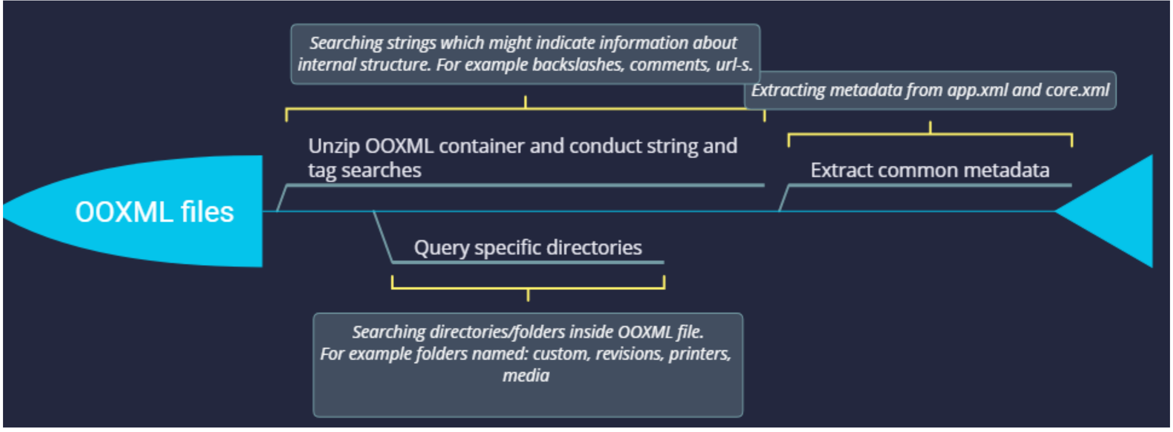

Searching strings which might indicate information about internal structure. For example backslashes, comments, url-s.
Extracting metadata from app.xml and core.xml
Unzip OOXML container and conduct string and tag searches
Extract common metadata
OOXML files
Query specific directories
Searching directories/folders inside OOXML file. For example folders named: custom, revisions, printers, media

Figure 21: Examination of OOXML file workflow

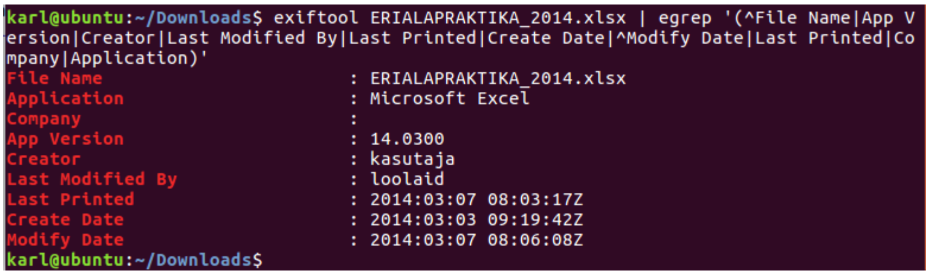
Phases which are shown in Figure 21 are described in more detail below.

Extracting common metadata

Common metadata fields are extracted using *ExifTool* much like with the PDF and MS binary documents. The following metadata fields are selected from *ExifTool* output and saved into the local database.

- Creator
- App Version
- Last Modified By
- Last Printed
- Create Date
- Modify Date
- Company
- Application

Metadata is gathered from *app.xml* and *core.xml* files inside the OOXML container. An example of a filtered output OOXML document is shown in Figure 22. The document is downloaded from UT website at random and analyzed with the method described above.



```
karl@ubuntu:~/Downloads$ exiftool ERIALAPRAKTIKA_2014.xlsx | egrep '(^File Name|App V
ersion|Creator|Last Modified By|Last Printed|Create Date|^Modify Date|Last Printed|Co
mpany|Application)'
File Name                       : ERIALAPRAKTIKA_2014.xlsx
Application                     : Microsoft Excel
Company                         :
App Version                     : 14.0300
Creator                         : kasutaja
Last Modified By                : loolaid
Last Printed                    : 2014:03:07 08:03:17Z
Create Date                     : 2014:03:03 09:19:42Z
Modify Date                     : 2014:03:07 08:06:08Z
karl@ubuntu:~/Downloads$
```

Figure 22: Output of E*xifTool* OOXML metadata

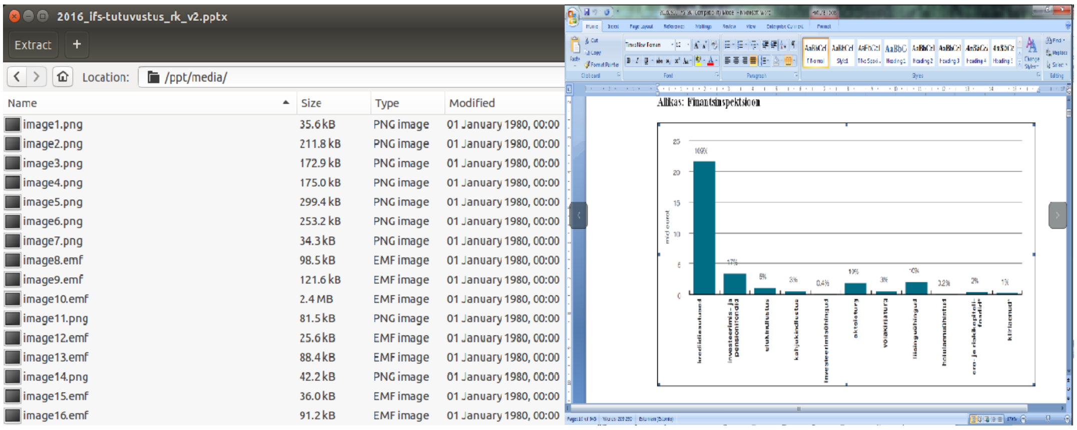
The extraction process is automated with bash script which scans the download folder for OOXML files and when the file is found metadata extraction utilized. Printer information is scraped with FOCA and this information is processed manually.

Querying specific directories

Second phase of the extraction process is folder searching inside of the OOXML file and manual examination. The analysis of several documents resulted in a decision that the only appropriate way to understand the contents of data that can be inside OOXML file is utilizing manual examination. Inside OOXML files there are several folders which potentially could contain sensitive information essential for fingerprinting purposes.

The following list presents the folders that could contain information needing additional manual investigation in order to gather sensitive information. The list is illustrated with the examples of extracted findings from random documents on the Web for proof of concept purposes.

- \media – contains media files, such as pictures and videos. In some cases media folder contains screenshots from the user's desktop, exposing operating system and software names, for example, information from taskbar icons. An example of *pptx* file which contains several files in its media folder, is shown in Figure 23. Examining the image, it can be deduced that the document author is probably using Microsoft Office 2007.



2016_ifs-tutuvustus_rk_v2.pptx

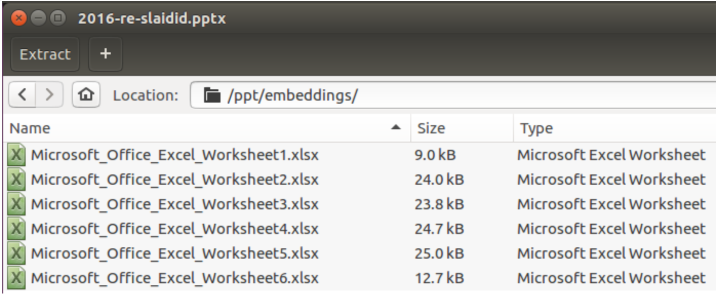
Extract +

Location: /ppt/media/

Name	Size	Type	Modified
image1.png	35.6 kB	PNG image	01 January 1980, 00:00
image2.png	211.8 kB	PNG image	01 January 1980, 00:00
image3.png	172.9 kB	PNG image	01 January 1980, 00:00
image4.png	175.0 kB	PNG image	01 January 1980, 00:00
image5.png	299.4 kB	PNG image	01 January 1980, 00:00
image6.png	253.2 kB	PNG image	01 January 1980, 00:00
image7.png	34.3 kB	PNG image	01 January 1980, 00:00
image8.emf	98.5 kB	EMF image	01 January 1980, 00:00
image9.emf	121.6 kB	EMF image	01 January 1980, 00:00
image10.emf	2.4 MB	EMF image	01 January 1980, 00:00
image11.png	81.5 kB	PNG image	01 January 1980, 00:00
image12.emf	25.6 kB	EMF image	01 January 1980, 00:00
image13.emf	88.4 kB	EMF image	01 January 1980, 00:00
image14.png	42.2 kB	PNG image	01 January 1980, 00:00
image15.emf	36.0 kB	EMF image	01 January 1980, 00:00
image16.emf	91.2 kB	EMF image	01 January 1980, 00:00

Figure 23: Contents of media folder exposes software version

- \embeddings – contains embedded or external content. Embeddings can be other files, such as OOXML documents or objects which are inserted into the document. For example, in Figure 24, embedded spreadsheet files are shown that are embedded in the PowerPoint presentation. The presentation file, shown in Figure 23, contains 6 different embedded files and those files usually preserve their metadata. That increases the possibility of extracting very compromising information about the observed organization.



2016-re-slaidid.pptx

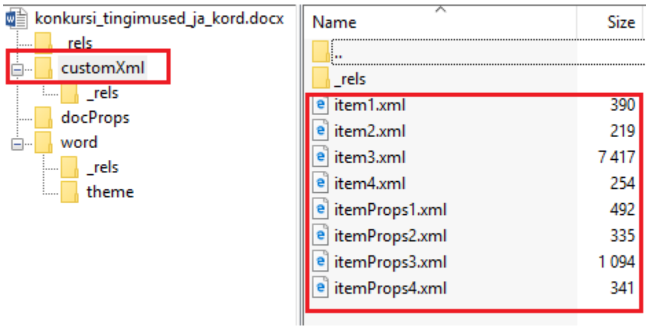
Extract +

Location: /ppt/embeddings/

Name	Size	Type
Microsoft_Office_Excel_Worksheet1.xlsx	9.0 kB	Microsoft Excel Worksheet
Microsoft_Office_Excel_Worksheet2.xlsx	24.0 kB	Microsoft Excel Worksheet
Microsoft_Office_Excel_Worksheet3.xlsx	23.8 kB	Microsoft Excel Worksheet
Microsoft_Office_Excel_Worksheet4.xlsx	24.7 kB	Microsoft Excel Worksheet
Microsoft_Office_Excel_Worksheet5.xlsx	25.0 kB	Microsoft Excel Worksheet
Microsoft_Office_Excel_Worksheet6.xlsx	12.7 kB	Microsoft Excel Worksheet

Figure 24: Content of random document embeddings

- \customXml – contains custom information which is inserted by the user or other applications, for example, document content management systems. Information found in that folder is ranged from innocuous to highly sensitive, depending on the implementation.

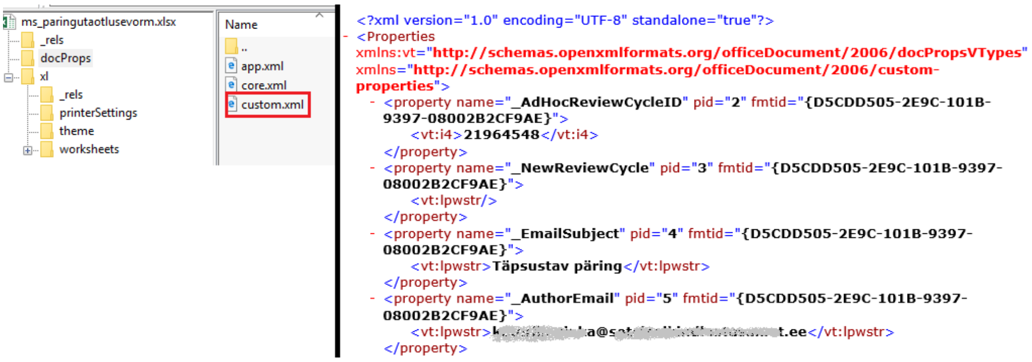


```
konkursi_tingimused_ja_kord.docx
  _rels
  customXml
    _rels
  docProps
  word
    _rels
    theme
```

Name	Size
..	
_rels	
item1.xml	390
item2.xml	219
item3.xml	7 417
item4.xml	254
itemProps1.xml	492
itemProps2.xml	335
itemProps3.xml	1 094
itemProps4.xml	341

Figure 25: CustomXml appearing in word document

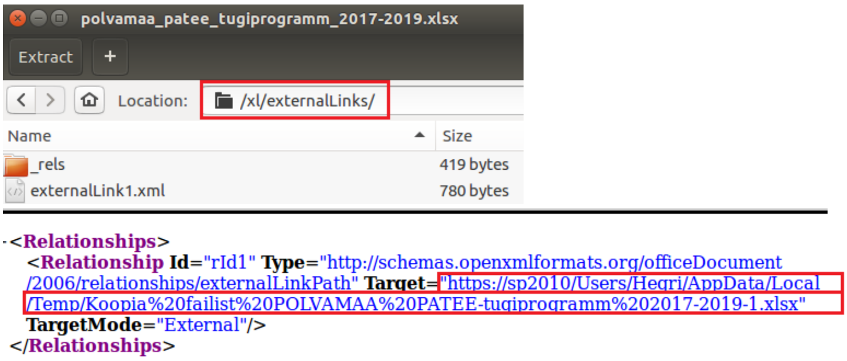
- \custom – Contains a custom.xml file which can hold any custom document properties added by the user or developer, or through custom logic. Very often the custom.xml is located in *docProps* directory. Information in the custom.xml file can be used for guessing or deriving internal services. For example, in Figure 26, custom.xml contains information about e-mail accounts.



```
<?xml version="1.0" encoding="UTF-8" standalone="true"?>
- <Properties
  xmlns:vt="http://schemas.openxmlformats.org/officeDocument/2006/docPropsVTypes"
  xmlns="http://schemas.openxmlformats.org/officeDocument/2006/custom-
  properties">
  - <property name="_AdHocReviewCycleID" pid="2" fmtid="{D5CDD505-2E9C-101B-
    9397-08002B2CF9AE}">
        <vt:i4>21964548</vt:i4>
    </property>
  - <property name="_NewReviewCycle" pid="3" fmtid="{D5CDD505-2E9C-101B-9397-
    08002B2CF9AE}">
        <vt:lpwstr/>
    </property>
  - <property name="_EmailSubject" pid="4" fmtid="{D5CDD505-2E9C-101B-9397-
    08002B2CF9AE}">
        <vt:lpwstr>Täpsustav päring</vt:lpwstr>
    </property>
  - <property name="_AuthorEmail" pid="5" fmtid="{D5CDD505-2E9C-101B-9397-
    08002B2CF9AE}">
        <vt:lpwstr>[redacted]</vt:lpwstr>
    </property>
```

Figure 26: Custom.xml contains information about e-mails

- \externallinks – contains information about relationships with other resources. Resources can be file absolute paths, external and internal services, or file servers. For example, in Figure 27 an absolute path to a local computer is shown, exposing the user name and operating system that were used.

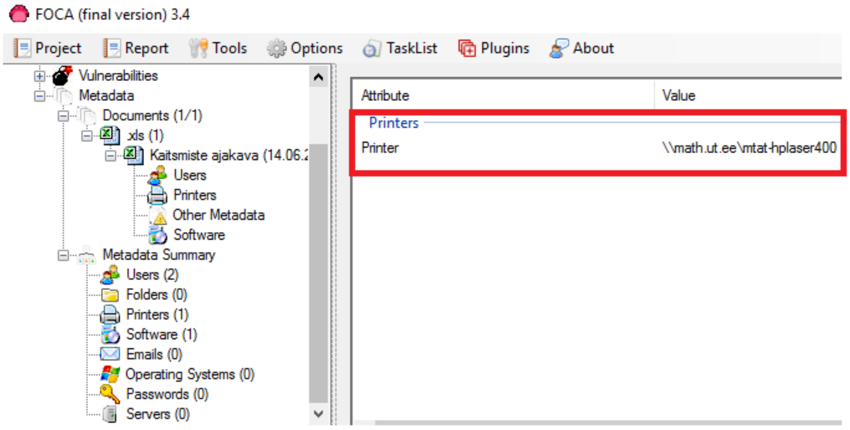


Name	Size
_rels	419 bytes
externalLink1.xml	780 bytes

```
-<Relationships>
   <Relationship Id="rId1" Type="http://schemas.openxmlformats.org/officeDocument
   /2006/relationships/externalLinkPath" Target="https://sp2010/Users/Hegri/AppData/Local
   /Temp/Koopia%20failist%20POLVAMAA%20PATEE-tugiprogramm%202017-2019-1.xlsx"
   TargetMode="External"/>
 </Relationships>
```

Figure 27: Directory externallinks exposing link to external resource

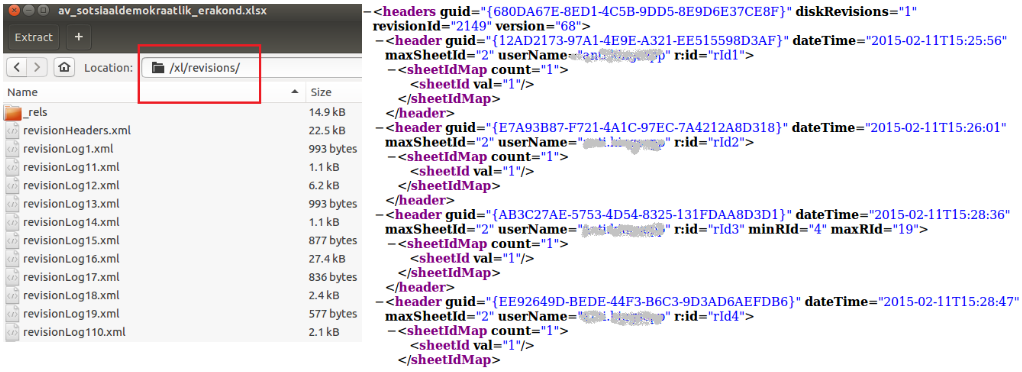
- printerSettings – contains information about the printer's name and drivers. Folder contents are binary files. For example, printer information extracted from the document "*Kaitsmiste ajakava (14.06.2016).xls*"that was downloaded from cs.ut.ee website.



Attribute	Value
Printers	
Printer	\\math.ut.ee\mtat-hplaser400

Figure 28: Printer information extracted with FOCA

- \revision – contains XML files which describes revisions and its related information. Folder contains timestamps about revisions and about author who made it. Figure 29 shows revision related files of presentation file.



Name	Size
_rels	14.9 kB
revisionHeaders.xml	22.5 kB
revisionLog1.xml	993 bytes
revisionLog11.xml	1.1 kB
revisionLog12.xml	6.2 kB
revisionLog13.xml	993 bytes
revisionLog14.xml	1.1 kB
revisionLog15.xml	877 bytes
revisionLog16.xml	27.4 kB
revisionLog17.xml	836 bytes
revisionLog18.xml	2.4 kB
revisionLog19.xml	577 bytes
revisionLog110.xml	2.1 kB

```
–<headers guid="{680DA67E-8ED1-4C5B-9DD5-8E9D6E37CE8F}" diskRevisions="1"
 revisionId="2149" version="68">
 –<header guid="{12AD2173-97A1-4E9E-A321-EE515598D3AF}" dateTime="2015-02-11T15:25:56"
   maxSheetId="2" userName="[redacted]" r:id="rId1">
  –<sheetIdMap count="1">
      <sheetId val="1"/>
    </sheetIdMap>
   </header>
 –<header guid="{E7A93B87-F721-4A1C-97EC-7A4212A8D318}" dateTime="2015-02-11T15:26:01"
   maxSheetId="2" userName="[redacted]" r:id="rId2">
  –<sheetIdMap count="1">
      <sheetId val="1"/>
    </sheetIdMap>
   </header>
 –<header guid="{AB3C27AE-5753-4D54-8325-131FDAA8D3D1}" dateTime="2015-02-11T15:28:36"
   maxSheetId="2" userName="[redacted]" r:id="rId3" minRId="4" maxRId="19">
  –<sheetIdMap count="1">
      <sheetId val="1"/>
    </sheetIdMap>
   </header>
 –<header guid="{EE92649D-BEDE-44F3-B6C3-9D3AD6AEFDB6}" dateTime="2015-02-11T15:28:47"
   maxSheetId="2" userName="[redacted]" r:id="rId4">
  –<sheetIdMap count="1">
      <sheetId val="1"/>
    </sheetIdMap>
```

Figure 29: Revisions exposing name of the revision's author and timestamp

Folders that are mentioned above can be found using *ExifTool* verbose mode or unpacking all OOXML files and using search utilities. *ExifTool* verbose mode (flag –V for activating verbose mode) returns in its output an unzipped structure of the OOXML file, presenting all its XML and data files in a directory tree. By filtering that output using *egrep* one can find folders discussed above, if they appear.

Unpacking OOXML files and conduct string searches

Final phase of metadata extraction from OOXML files are specific string searches from unpacked OOXML files. In the analysis of *.docx*, *.pptx*, and *.xlsx* files we notice that metadata, such as network and local path information as well as other custom information, can be stored inside other parts of OOXML files. For finding that information we were considering using text search utility *egrep* and search for specific strings.

Before searching strings and XML tags, the documents have to be in an unpacked form. For unpacking the documents we used bash script which scans the folder, and if it finds files with the extensions *.docx*, *.xlsx*, *.pptx* then the document will be unzipped.

The following strings are searched:

- "\" or "\\" – finding backslashes for exposing possible directories, server addresses, paths, network shares;
- "@" exposing possible e-mail accounts associated with the document. Regular expression can be utilized: "@\w+\.\w+";

- "absPath" – exposing path where the document was modified or created. Usually indicates the path were document is modified;
- "cmAuthor" and <author> – exposing the names of people who have commented the document.
- <Url> <\URL> tags – could contain internal services addresses;
- ".intra", ".sise" – could contain internal DNS names;
- "File:///" - contains file path information;
- "https" – exposes potential web services that are used in organization;
- "Descry=" – exposes information about local paths, referring to other sources such as pictures or other documents.

If some of the custom information is found in documents, one can specify the string searches upon the string found and utilize regular expressions.

This phase of the OOXML extraction presents a novel method and string values in order to find sensitive information from OOXML container documents.

3.4 Stage 3 - Metadata analysis

In the previous subchapter we mentioned that metadata extracted with *ExifTool* were stored into a local elasticsearch database. We use that database for making the metadata analysis quicker and more convenient. This raises effectiveness of processing the metadata which was gathered. In addition, we use Kibana graphical interface to make collected information more readable.

For the summary of the metadata analysis we want to get answers to the following aspects:

- Services
    - External – services that are described in metadata and are accessible over the Internet;
    - Internal – services that run only in corporate networks;
- Domain namespace – internal domain namespace;
- Servers – file servers, print servers, domain controllers;
- Softwares, application – what software runs in observed organization environments;
- User roles – finding the website content manager, for example;
- Operating systems.

Outcome of this stage of fingerprinting method is to analyze extracted metadata information and find sensitive information which can be used for fingerprinting purposes and for conducting possible cyber attacks against observed organization. In addition, to determine possible links and relationships between people and organizations who collaborate in document processing.

This chapter described a method of fingerprinting an organization by just using the metadata of electronic documents hosted on their website. The fingerprinting method consists of three separate stages – document collecting, metadata extraction, and metadata analysis. We introduced novel methods and approaches on how all the meaningful information can be extracted from documents. The next chapter presents the results and analysis of the outcomes of the fingerprinting method which was used against three different governmental organizations' webpages.

4 Results and analysis

This chapter presents the results of three Estonian governmental organization's public document metadata analysis. The analysis is conducted by three stages that were discussed in Chapter 3.

The fingerprinting method discussed in Chapter 3 is utilized against Estonian ministries' webpages in order to validate or disproof the proposed hypothesis. Those ministries were chosen upon collaboration with the ministries' chief of information officers. We asked permissions to conduct this study and share our results with them. For privacy reasons, those ministry names are hidden and ministries are named as follows: Ministry A, Ministry B, and Ministry C. The actual results of the extracted information is showed in Appendixes 3-5.

The first stage of this analysis is document gathering. Documents are downloaded from governmental organizations' webpages and analyzed. All the documents are publically available for everyone and there is no access to webservers' upload or data folders directly. No automated tools are used; all the documents are gathered by utilizing search engine features. Documents with the following extensions are searched and downloaded*: pdf, docx, doc, xlsx, xls, pptx, ppt*. Other document formats are ignored. The second stage is metadata extraction from collected documents. The third and final stage presents the results of stage 1 and 2 outcomes. The ultimate goal is to gain sensitive information about organizations' IT infrastructure and information about end users devices which aids in building attacks against those organizations and fingerprinting them.

For conducting this analysis two different virtual machines are used: Ubuntu 16.04 and Windows 8.1. All the tools that are used are open source and available for everyone.

The following subsections will present the outcomes of three Estonian governmental organizations' metadata analysis with statistical data, and the conclusions.

4.1 Documents gathered

This section presents the results of collected documents from observed websites. This is the first stage of fingerprinting method. Documents were collected from the organizations' websites that are associated with the organization's domain and indexed by search engines.

Results of gathered documents are shown in Table 3.

Table 3: Documents collected form observed ministries

Ministry A		**Ministry B**		**Ministry C**	
PDF:	275	PDF:	345	PDF:	960
DOCX:	56	DOCX:	4	DOCX:	22
doc:	71	doc:	20	doc:	189
xls:	84	xls:	1	xls:	374
xlsx:	125	xlsx:	0	xlsx:	9
ppt:	14	ppt:	0	ppt:	73
pptx:	19	pptx:	0	pptx:	2
Total: 644		**Total:** 370		**Total:** 1633	

The results of document occurrences in organizations' websites validate the research which is presented in Appendix 1 – PDF documents are most popular electronic document format available on websites. We can notice that document occurrences depend highly on the behaviour of a particular organization. For example, Ministry B is more restricted in sharing documents on their website, especially MS Office documents, but Ministry C has five times more documents hosted on their website. A better visual overview is presented in Figure 30.

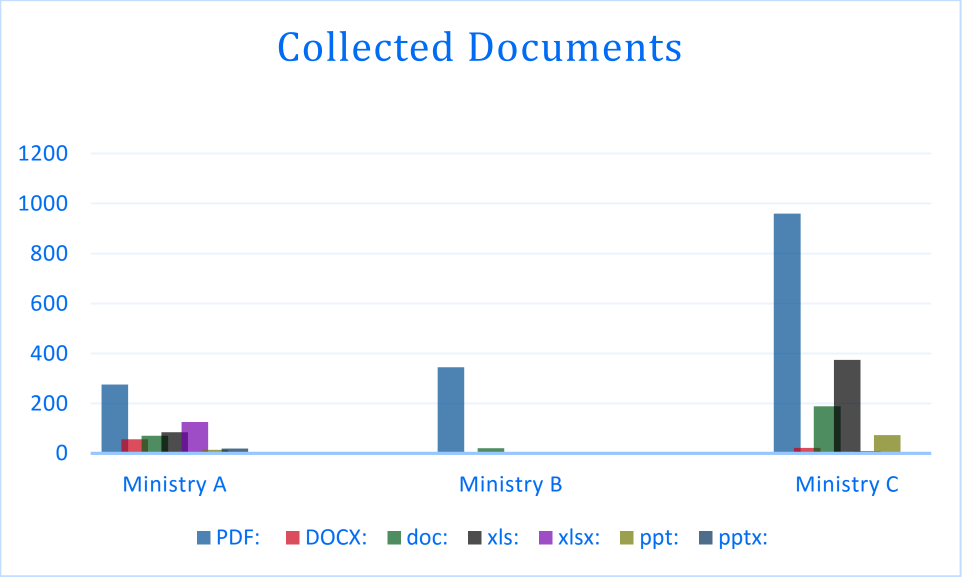


Figure 30: Gathered documents by document extension

The majority of MS documents that were found in observed sites were in older MS binary format, which can indicate that either those documents are old and uploaded years ago, or those document templates are reused and the format is preserved that way.

The amount of collected documents was expected since governmental entities have to be transparent according to laws and have to provide service to citizens which often means document sharing on their websites. The data set is reasonable and gives prerequisites for the next stages of this analysis. We managed to collect 2643 electronic documents in total.

4.2 Metadata extracted

The second stage of the fingerprinting method is metadata extraction. Metadata is extracted from gathered documents in previous section. Statistical results are presented in the following chapters and the actual extracted metadata information is presented in Appendixes 3-5.

4.2.1 Ministry A

The following tables present the results of scraping metadata from public documents from Ministry A website. The total amount of documents which were analyzed is 644. The results for each file format are presented by its format type. Table 4 presents results of extracted metadata fields from PDF documents.

Table 4: Extracted metadata from Ministry A PDF documents

Metadata field	Rate %	Files Affected:
File Name	100,0	267
Author	81,3	217
Creator	99,6	266
Producer	93,6	250
Title	30,0	80
Create Date	100,0	267

Modify Date	99,3	265
PDF version	100,0	267
XMP Toolkit	22,5	60
Creator Tool	22,5	60

We managed to extract 217 author's names from 267 PDF documents, which could be potential usernames of computer accounts. Most frequently occurred information that we managed to extract were timestamps, information about document author, application which processed the document, and PDF version (Metadata fields: Create Date, Modify Date, Creator, Producer and Author). The results presented in Table 4 indicate that the metadata of PDF documents is not properly removed and every document contains some of the metadata fields which can be used for fingerprinting.

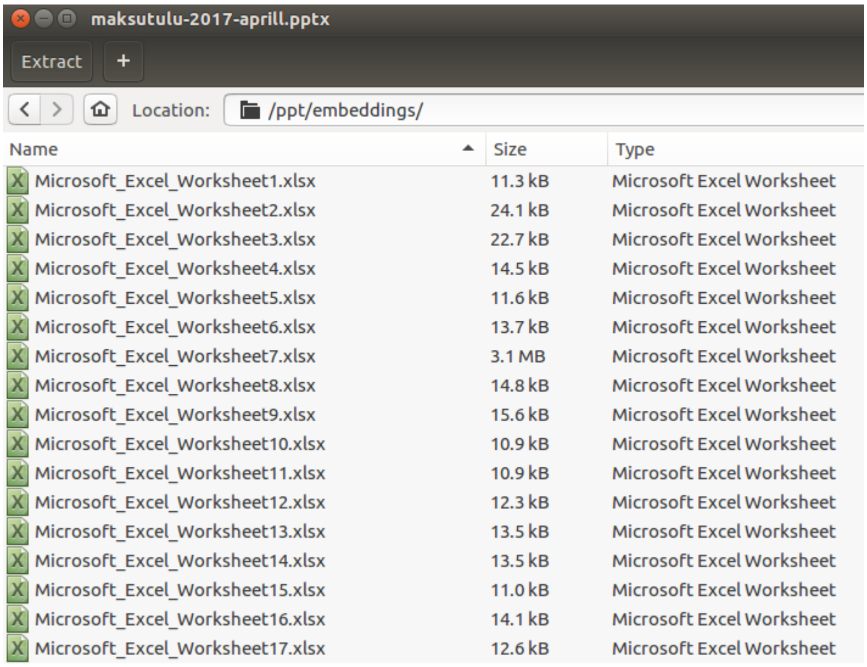
In addition to PDF documents, we managed to download 369 Microsoft Office documents from Ministry A webpage. MS Office documents were with extensions *docx*, *doc*, *xls*, *xlsx*, *ptt*, *pttx*. The following Table.5 shows metadata extraction results of MS documents.

Table 5: Extracted metadata from Ministry A MS Binary files

MS Binary format (doc, xls,ppt)			MS OOXML format (docx, xlsx, pttx)		
Metadata field	Rate %	Files Affected:	Metadata field:	Rate %	Files Affected:
File Name	100,0	170	File Name	100,0	199
Author	100,0	170	Creator	95,5	190
Last Modified By	100,0	170	Last Modified By	95,5	190
Software	97,6	166	Last Printed	44,7	89
Create Date	100,0	170	Application	100,0	199
Modify Date	100,0	170	Create Date	100,0	199
Company	91,8	156	Modify Date	100,0	199
App Version	100,0	170	Company	78,9	157
Comp Obj User Type	73,5	125	App Version	100,0	199
Printer Information	53,5	91	Printer Information	58,8	117
Paths	1,2	2	Embedding's	6,5	13

The majority of the documents contained author information as well as additional names in "Last Modified By" and "Last Printed" metadata field (shown in Table 5), which expose potential usernames and personnel names. Compared to PDF documents, MS documents contain more extra information in metadata fields and the rates of metadata occurrences are higher. More than half (56%) of the documents contained printer information and we were able to extract the local path information from two of the MS binary formatted documents. In addition, many documents contained "Company" information which gives additional confidence about where the document is originating from.

From MS OOXML documents, the existence of embeddings were discovered in 13 documents (shown in Table 5). From the 13 documents containing embedded extra content, we were able to extract 140 different files: 127 spreadsheet documents (with xlsx and xls extensions) and 13 binary objects. Surprisingly, the number of found documents inside embeddings is high, almost as many as were gathered from the website. Those embedded spreadsheets contain their own metadata and could potentially contain sensitive information. Embeddings were found from Word and PowerPoint documents. One example of embedding occurrence in a document is presented in Figure 31. The PowerPoint presentation file contained 17 embedded spreadsheet documents.



Name	Size	Type
Microsoft_Excel_Worksheet1.xlsx	11.3 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet2.xlsx	24.1 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet3.xlsx	22.7 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet4.xlsx	14.5 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet5.xlsx	11.6 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet6.xlsx	13.7 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet7.xlsx	3.1 MB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet8.xlsx	14.8 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet9.xlsx	15.6 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet10.xlsx	10.9 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet11.xlsx	10.9 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet12.xlsx	12.3 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet13.xlsx	13.5 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet14.xlsx	13.5 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet15.xlsx	11.0 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet16.xlsx	14.1 kB	Microsoft Excel Worksheet
Microsoft_Excel_Worksheet17.xlsx	12.6 kB	Microsoft Excel Worksheet

Figure 31: PowerPoint presentation containing 17 embedded spreadsheets

Each worksheet has its own metadata. In addition, those spreadsheets can potentially contain sensitive corporate data, which could expose internal workflow, processes, calculations, and data.

The results presented in Table 4 and Table 5 are extracted using *ExifTool*. As was explained in Chapter 3.3.3, we utilized novel manual examination methods for extracting metadata from OOXML files. Table 6 presents the results of manual examination.

Table 6: Results of manual examination of Ministry A OOXML documents:

Hidden information findings	Unique	Total
Local and network paths, directories	41	169
Additional document authors name	39	85
URL-s	3	22

We were able to manually extract additional information from comments, revisions, and custom fields. The results of manual extraction are significant; the information we managed to gather exposes internal file servers and services.

According to the results shown in Tables 4-6 we determine that MS Office documents contain more descriptive information about the observed organization's internal assets. Most of the compromising information we managed to extract from OOXML documents and especially from presentation files.

The results showed that all the documents contained some metadata and metadata common fields were not removed. Actual contents of extracted results are presented in Appendix III.

4.2.2 Ministry B

The following tables present the results of scraping metadata from public documents of Ministry B website. Total amount of documents which were analyzed are 370.

Table 7: Extracted metadata from Ministry B PDF documents

Metadata field	Rate %	Files Affected:
File Name	100,0	345
Author	80,0	276
Creator	90,4	312
Producer	93,3	322
Title	51,3	177
Create Date	95,1	328
Modify Date	90,7	313
PDF version	99,1	342
XMP Toolkit	37,7	130
Creator Tool	34,5	119

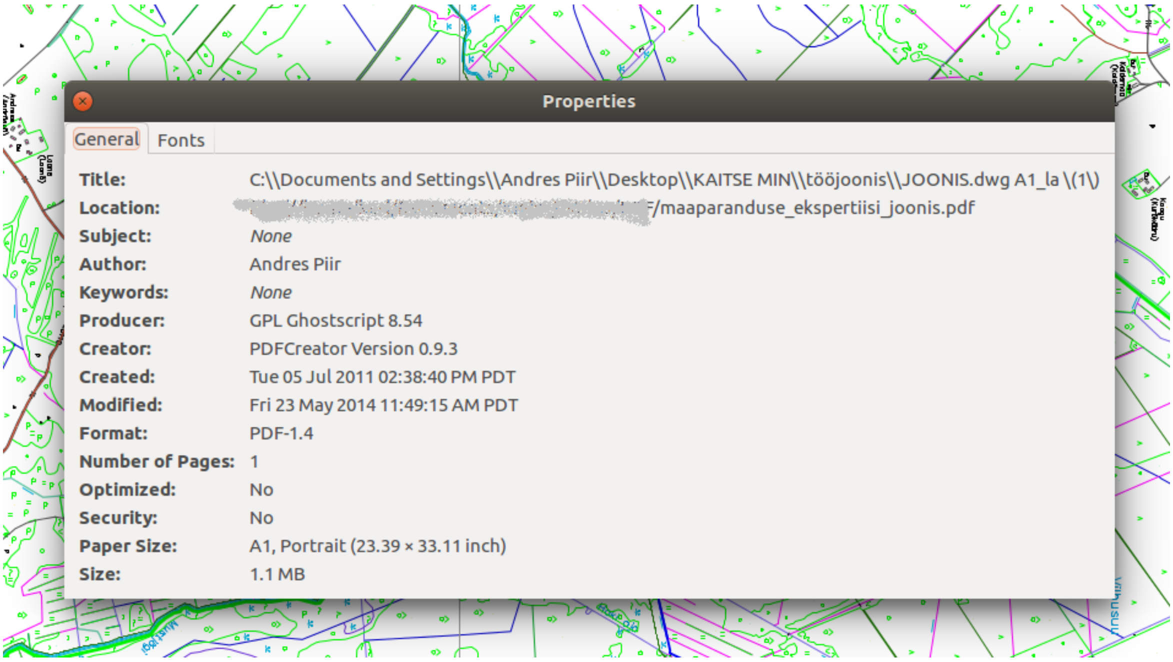
As shown in Table 7, the majority of documents which existed on Ministry B website were in PDF format. We were able to detect 12 PDF documents which had properly removed metadata and only the metadata field "PDF version" existed in metadata.

We also managed to gather some of the MS Office documents from the observed website and the results of extracted metadata are presented in Table 8.

Table 8: Extracted metadata from Ministry B MS Office documents

MS Binary format (doc, xls,ppt)			MS OOXML format (docx, xlsx, pttx)		
Metadata field	Rate %	Files Affected:	Metadata field:	Rate %	Files Affected:
File Name	100,0	23	File Name	100	4
Author	91,3	21	Creator	100	4
Last Modified By	91,3	21	Last Modified By	100	4
Last Printed	39,1	9	Last Printed	25	1
Software	100,0	23	Application	100	4
Create Date	100,0	23	Create Date	100	4
Modify Date	100,0	23	Modify Date	100	4
Company	26,1	6	Company	100	4
App Version	100,0	23	App Version	100	4
Printer Information	4,3	1	Printer Information	0	0
Paths	4,3	1	Embedding's	0	0

From the MS binary format documents we were able to detect one printer and one local file path. From OOXML files we did not detect embeddings or custom information that might expose compromising information. An interesting finding was the file path information which was inside the PDF metadata field "Title". This is shown in Figure 32.



Title:	C:\\Documents and Settings\\Andres Piir\\Desktop\\KAITSE MIN\\tööjoonis\\JOONIS.dwg A1_la \(1\)
Location:	[redacted]/maaparanduse_ekspertiisi_joonis.pdf
Subject:	*None*
Author:	Andres Piir
Keywords:	*None*
Producer:	GPL Ghostscript 8.54
Creator:	PDFCreator Version 0.9.3
Created:	Tue 05 Jul 2011 02:38:40 PM PDT
Modified:	Fri 23 May 2014 11:49:15 AM PDT
Format:	PDF-1.4
Number of Pages:	1
Optimized:	No
Security:	No
Paper Size:	A1, Portrait (23.39 × 33.11 inch)
Size:	1.1 MB

Figure 32: Metadata found in “Title” field, exposing file local path information

Comparing results shown in Table 8 with Ministry A’s results it can be concluded that they are demure. We were not able to detect many local and network path informations and it is mainly due to the occurrences of MS Office documents being few. MS Office documents have more capabilities to store path information and embeddings.

4.2.3 Ministry C

The following tables present the results of scraping metadata from public documents of Ministry C website. Total amount of documents which were analyzed are 1633.

Table 9: Extracted metadata from Ministry C PDF documents

Metadata field	Rate %	Files Affected:
File Name	100,0	959
Author	79,9	766
Creator	87,4	838
Producer	99,6	955
Title	79,2	760
Create Date	99,3	952
Modify Date	96,1	922
PDF version	100,0	959
XMP Toolkit	22,5	216
Creator Tool	19,3	185

All the PDF documents that we analyzed contained metadata which could be used for fingerprinting purposes. Since the amount of PDF documents we located on their website is large, we were able to extract a lot of metadata. We extracted 838 user information and 955 document producer application names with version info. Those amounts of metadata give a very good overview of the software portfolio used in the organization and of the authors and users who are working with the documents.

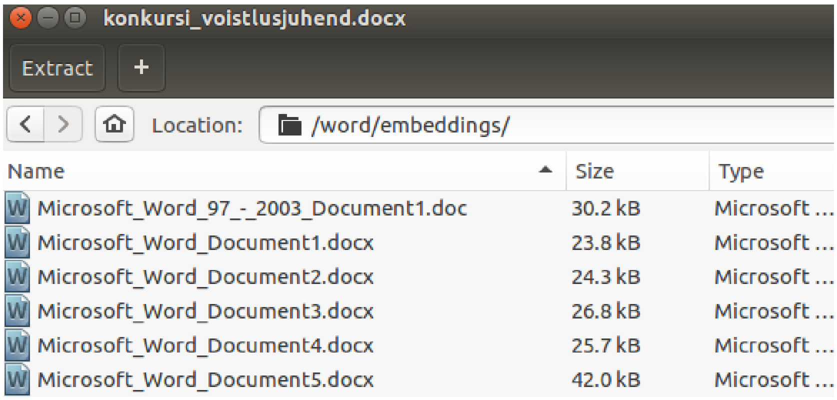
We also gathered significant amount of MS Office documents. Most of the documents were in older MS binary format.

Table 10: Extracted metadata from Ministry of C MS Office documents

MS Binary format (doc, xls,ppt)			MS OOXML format (docx, xlsx, pttx)		
Metadata field	Rate %	Files Affected:	Metadata field:	Rate %	Files Affected:
File Name	100,0	636	File Name	100,0	33
Author	46,4	295	Creator	100,0	33
Last Modified By	90,6	576	Last Modified By	100,0	33
Last Printed	40,7	259	Last Printed	36,4	12
Software	49,2	313	Application	100,0	33
Create Date	99,8	635	Create Date	100,0	33
Modify Date	100,0	636	Modify Date	100,0	33
Company	43,6	277	Company	90,9	30
App Version	100,0	636	App Version	100,0	33
Printer Information	12,9	82	Printer Information	9,1	3
Paths	10,4	66	Embedding's	15,2	5

We were able to extract a lot of metadata from the MS office documents. Every document contained metadata. Most of the sensitive metadata we extracted from the documents originated from the MS binary and spreadsheet documents.

We found five OOXML documents containing embeddings. From the embeddings we managed to extract 14 different files: 10 files with *.docx* extension, 1 binary file, 1 file with *xlsx* extension, and 1 file with *.doc* extension. One OOXML Word document contained 5 different Word files which were embedded into it (shown in Figure 33). Word files embedded to Word files is a relatively rare occasion. All embeddings preserved their own original metadata and we were able to additionally extract it.



Name	Size	Type
Microsoft_Word_97_-_2003_Document1.doc	30.2 kB	Microsoft ...
Microsoft_Word_Document1.docx	23.8 kB	Microsoft ...
Microsoft_Word_Document2.docx	24.3 kB	Microsoft ...
Microsoft_Word_Document3.docx	26.8 kB	Microsoft ...
Microsoft_Word_Document4.docx	25.7 kB	Microsoft ...
Microsoft_Word_Document5.docx	42.0 kB	Microsoft ...

Figure 33: Word files embedded to Word document

From manual examination of OOXML documents we managed to extract information which is showed in Table 11.

Table 11: Metadata gathered with manual examination from Ministry C OOXML documents

Hidden information findings	Unique	Total
Local and network paths, directories	10	25
Additional document authors name	3	3
URL-s	0	0

The number of OOXML documents on Ministry C website were not very large, we managed to extract 25 path information from 33 documents. Almost every OOXML document contained information about path where it was modified, which is impressive. We did not detect any URL-s from metadata which refers to internal service.

In this section we processed 2643 electronic documents and extracted metadata from them. Only 12 documents had metadata properly removed, even timestamps. Most of compromising information were found in MS OOXML documents. We managed to extract 154 files and documents from OOXML embedding's. This is number is significant, those embedded files could possible contain sensitive internal information and we got additional metadata information from those embedded documents. Most embedding's were found in presentation files, which to the nature of the document promotes embedding's existents. All the printer information is originating from MS spreadsheet documents. This is same for both versions of MS Office. We also detect some of the printer information from presentation files and from PDF documents. In PDF documents printer information were described in "producer" metadata field.

PDF documents are one of the most popular document format which is shared in public websites. One reason for that is when converting document to PDF format it strips most of the metadata. We analyzed 1850 PDF documents which almost all contained its core metadata fields. PDF documents gave very good overview about software names and versions, also about document authors in the observed organization.

We detected most of custom information from Word documents.*(.docx).* Word documents have included custom fields in its OOXML structure which contained information about document management systems parameters as well as URLs of document management service.

According to the metadata extraction results, we can conclude that most dangerous documents in terms of metadata extraction are OOXML presentation documents. We were able to find one document which contained 17 embeddings. Next section presents manual analysis of extracted metadata.

4.3 Analysis of extracted metadata

This section describes analysis of documents metadata which were gathered and extracted from previous stages. Outcome of this section is the manual analysis of extracted information to determine whether observed organizations leaks information which aids third-parties to conduct cyber attacks against them. In addition this section validates our hypothesis which is posed for this thesis.

4.3.1 Analysis of Ministry A

The final stage of the fingerprinting method is the analysis of the extracted information which were gathered in the previous stages. Table 12 presents conclusive results of the Ministry A metadata analysis based on manual observation of extracted metadata.

Table 12: Analysis results of Ministry A

Findings	**Description of findings**
Detected services	**Internal** We were able to detect two internal services, and according to path descriptions from metadata, those services are for document management/sharing services. One of the services was using HTTP protocol, which means if the service has some kind of authentication implemented, it is a good place to sniff passwords. From the local file path information we detected an e-mail client application and we can assume which e-mail services the organization most likely uses. **External** We found references to external webserver located in Estonia. We cannot determine whether it belongs to Ministry A or not, but that information can be used for cyber activities.
Doman namespace	From the local and network paths and from document common metadata we were able to detect four different internal domain names.
Servers	We discovered nine different servers. Three of the detected servers are most probably Windows fileservers. In addition, one of the servers stores the organization's users roaming profiles. Another six servers are print servers.
Software, application	Most frequently used Office software is Microsoft Office 2007 and 2013. We determined that mentioned software is the main program for creating PDF documents. Overall, names of the software that are used for document creation were detected.
Printer information	We were able to detect 6 different print servers. In addition, we detected printer vendors that are used in that organization,

	and driver information. Printer information indicates that there are different physical locations for the organization, or personnel are printing documents in multiple places. Furthermore, from printer information we are able to conclude that the organization is located in a high building and some of the personnel are working on the 14h floor.
Personnel information, roles	We detected 178 unique user information. Majority of that information was extracted from document author and creators fields; however, some of the information were extracted from embeddings, local paths, and comments. We detected two possible persons who could be websites content managers.
Relationships between persons, companies, organizations.	Identified possible relationships of certain users between different organizations. From metadata some of documents contained other governmental organization metadata. In addition links between those different organization personnel are detected.
Operating systems	We detected that mainly Windows based operating systems is used. In addition some of Windows XP workstations and Macintosh operating systems identified.

Some of the more interesting findings show patterns between people and organization. From the “Company” metadata field we detected other organizations’ names in documents that were uploaded in Ministry A’s website. This means that those documents are exchanged between organizations and the metadata exposes the documents’ actual authors. Therefore, we are able to see the patterns between organizations and people who share documents with each other, and this opens many potential attack vectors. In addition, some of the documents which contained a different name on the “Company” metadata field instead of where they were originating from, contained some sensitive metadata of their own. This means that documents which were uploaded onto Ministry A website contained other governmental organization’s documents metadata filled with some compromising information, such as information about internal servers and domain names.

The compromising information about links between documents’ authors come from “last modified by”, “last printed”, and “creator” metadata fields. Correlating these fields with “company” metadata field, we were able to detect in which company document author is most probably working. For example, in Figure 34 possible links between the organizations and author of the document are presented. For drawing such a graph, Maltego software is used [29]. Maltego software is used widely by penetration testers for gathering information about observed targets and to detect links between several entities. In Figure 34 we searched links between “author” and “company” metadata fields.

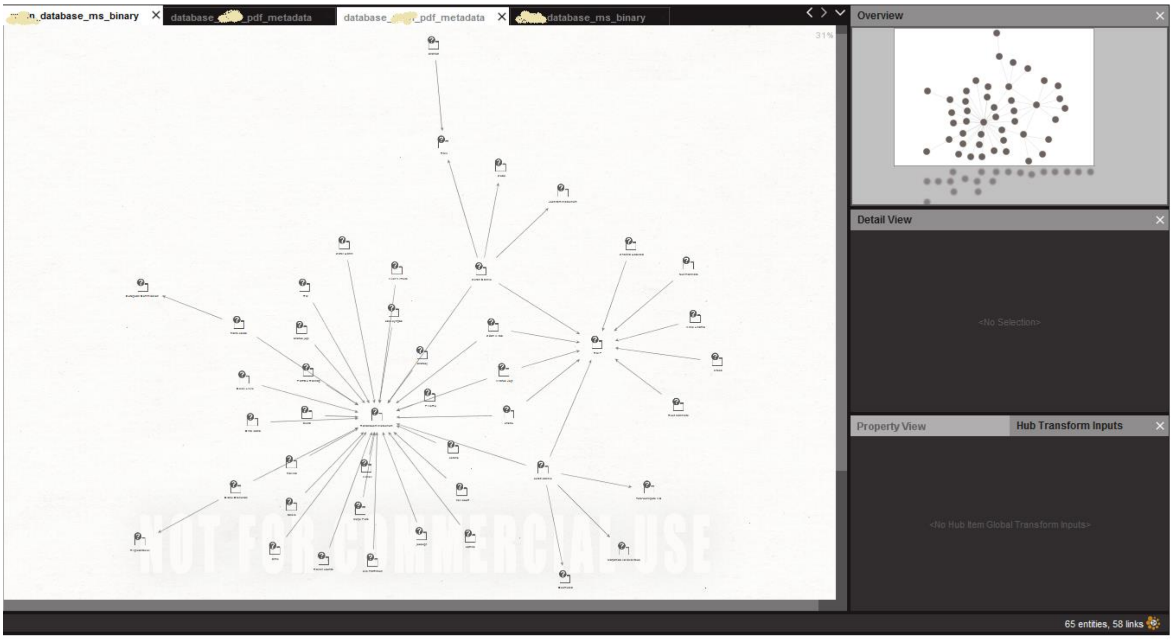

[redacted]n_database_ms_binary
database[redacted]pdf_metadata
database_[redacted]_pdf_metadata
[redacted]database_ms_binary
31%
NOT FOR COMMERCIAL USE
Overview
Detail View
<No Selection>
Property View
Hub Transform Inputs
<No Hub Item Global Transform Inputs>
65 entities, 58 links

Figure 34: Visualizing links between document authors and organizations with Maltego

Those possible links between document authors and organizations are useful information for cyber criminals who could send spoofed e-mails based on the exposed links between people. In addition, attackers might approach the target organization through a partner organization’s systems.

We managed to bind some of the users network share names with the internal server. If a user had network mapping “X:\” we were able to detect the server location in the internal network. Since we saw patterns of network mappings among a certain amount of people and the timestamps indicated nearly the same time period, we were able to understand network mappings which are automatically mapped with a computer in the internal domain.

Findings presented in Table 12 show that documents that are uploaded on organizations’ sites are full of compromising data. The information we managed to gather from metadata is intrusive and it can be used for building cyber-attacks against a particular organization. We managed to fingerprint many assets of the Ministry A’s internal network without having access to it and we did not do any active scan. Hypothesis posed in this thesis is validated - those findings aid attackers in picking targets and attacking techniques more accurately. Attack vectors are discussed more deeply in section 4.4.

4.3.2 Analysis of Ministry B

The final stage of the fingerprinting method is the analysis of extracted information which were gathered in previous stages. Table 13 presents conclusive results of the Ministry B metadata analysis based on manual observation and analysis of the extracted metadata.

Documents which we managed to gather from Ministry B website did not contain very many Microsoft Office documents. This is one of the reasons why we did not get that much information about their internal services, printers, or assets. Metadata analysis is depending highly on the document formats a particular website contains, most compromising are OOXML documents.

Table 13: Summary of metadata analysis about Ministry B

Findings	**Description of findings**
Detected services	**Internal** We detected possible e-mail services. One extracted local file path describes an e-mail client application, which lets us assume what e-mail server is most probably used.
Doman namespace	We detected one possible internal domain name from a PDF's "title" metadata field.
Servers	We were unable to detect servers.
Software, application	One of the most used software for document creation procces is Microsoft Office 2013. We detected the usage of this software among many users and timestamps information confirms that this software is actively used. Furthermore, we detected that many users use other PDF reader than Adobe. Also for PDF creation, several organization workers use Multi-functional printer for generating PDF documents. Software information that we were able to extract from PDF documents gives information about software portfolio of that organization.
Printer information	We detected one printer name from MS Office documents. The name of the printer indicates that it is directly connected to the computer. We determined that many Multi-functional printers were used for generating PDF documents. Most popular were "Canon" printers.
Personnel information, roles	We were able to detect 116 unique usernames that exist in metadata, also possible website content managers.
Relationships between persons, companies, organizations.	Connections between other organizations, and between persons and other organizations or companies are detected. We found many other companies names in metadata fields which means we received

	knowledge about who the cooperative partners for that organization are. We identified 8 unique organization or company names.
Operating systems	The majority of the detected OS were Windows 7 64 (Ultimate 64x), also some occurrences of Windows Vista Ultimate Home Edition x64. The number of Windows 7 information in metadata in comparison with timestamp and author information lets us state with certainty that the organization uses Windows 7 operating system.

Due to the lack of MS Office documents existing on Ministry B website, we did not detect any server information from the metadata. However, high occurrences of PDF documents on the website gives us a pretty good overview of the software portfolio and their versions that are used in that organization. This information matched with organization workers can give an overview of user software running on their computers. Attackers can select exploits and prepare malware based on that information. Theoretically, it is possible to detect software's update cycles in organizations. We observed that one of the document author names was actively in metadata. Comparing the same author name with software version and putting this information into timeline, one can understand the possible update times. Although were not able to detect update cycles, that amount of PDF documents gave good prerequisites.

We observed that the PDF documents contain in its metadata fields (in "producer", "creator tool" fields) accurate information about software and its versions which creates it. Also, those fields tends to include operating system information. Some of the examples:

- `PDF-XChange 4.0.193.0 (Windows Seven Ultimate x64 (Build 7600))`
- `Acrobat PDFWriter 4.0 for Windows NT`
- `PDF-XChange (PDFTools4.exe v4.0.0212.0000) (Windows)`

Software information not only exposes software names and version information, it also exposes vulnerabilities that the computer where the document was created may have. For example, in the above list of software "PDF-XChange 4.0.193.0" version shows that it is outdated and has vulnerabilities. Comparing version information, author name, and timestamps we can detect an outdated computer which aids attackers in selecting attacks or exploits for the infected computer. We detected that users had been using that vulnerable "PDF-XChange" application for three years.

4.3.3 Analysis of Ministry C

The final stage of the fingerprinting method is analysis of extracted information which were gathered from previous stages. Table 14 presents conclusive results of the Ministry C.

Table 14: Summary of metadata analysis about Ministry C

Findings	**Description of findings**
Detected services	**Internal:** Detected file sharing services, most likely a

	file server, working on Windows operating system.
Doman namespace	Detected two possible internal domain names.
Servers	We were able to detect 5 different print servers. In addition, we detected a potential internal e-mail server.
Software, application	From local file path information we detected an E-mail client application. Most frequently used software for that organization was Microsoft Office 2010. For PDF documents creation, Microsoft 2010 Word and Adobe InDesign are most used applications. Since we analyzed a large amount of PDF documents we gained good visibility of software portfolio used for creating electronic documents.
Printer information	This organization uses multiple manufacturer's printers and it seems that printers are not consolidated to one vendor. Most of the printers are connected to one certain print server which is probably one of the main print servers. Printer names are not associated with physical locations in this organization. One printer name indicates a different department.
Personnel information, roles	Detected 160 unique author names which can potentially be usernames for local computers. We detected two possible Web content managers whose names occur most frequently in metadata.
Relationships between persons, companies, organizations.	Detected connections between several cooperation partners. In addition, we detected links with a company that probably provides translating services for our observed organization.

Operating system	Most used operating system is Windows and most likely the version is 7 64x. In addition, we detected some occurrences of Macintosh. We detected many occurrences of Macintosh computers having Adobe commercial products, such as InDesign and Illustrator installed in their systems. It can be assumed that these computers belong to designers or editors.

Since the observed website contained many PDF documents, we got a good overview of the various software used in their organization. Some of the software names contained information in which the software's website address was presented. For example:

```
• PDF Printer / www.bullzip.com / FPG / Freeware Edition (max 10 users)
```

This data gives additional information for attackers for preparing exploits or sending fake download links to "update" the PDF software. Furthermore, the software that is mentioned above is allowed to be used only by 10 users in a corporation, so if the number of users is higher then the organization is violating the software usage policies. We detected 5 different users for that software. However, we analyzed documents which were available in public websites; therefore, since most of the documents are created and stored on internal networks then software usage's policy violation might be possible. This information can be used for blackmailing the organization or cause reputational damage.

The majority of the MS documents were in binary format which means we were not able to gather much custom information. Path information which were extracted, contained mainly information about the network share name and not full path address to server. For example:

```
• S:\mpo\2_büroo\Teemad\Õigusaktid, juhendid\Rakendusaktid\2017\VORMID\
• P:\Viisastatistika\
```

Analysis presented in Section 4.3 tables above showed that organization fingerprinting depends highly on the document formats which are hosted in the observed website, and the amounts. Most meaningful metadata for fingerprinting basis were extracted form MS OOXML documents. From the Ministry A presentation document we were able to detect 17 different embedded spreadsheet documents which had preserved metadata. One of the documents could potentially expose significant amount of compromising information. In addition, PDF documents which are often known as a metadata strict document format, contained compromising data about users, local paths, printers, applications, and their versions.

From the observed websites we detected documents that were originating from other companies and contained compromising metadata. We noticed this situation in one of the observed organizations's website. This shows that when metadata is not properly moved it can advance to unknown places and the original document author cannot control that flow.

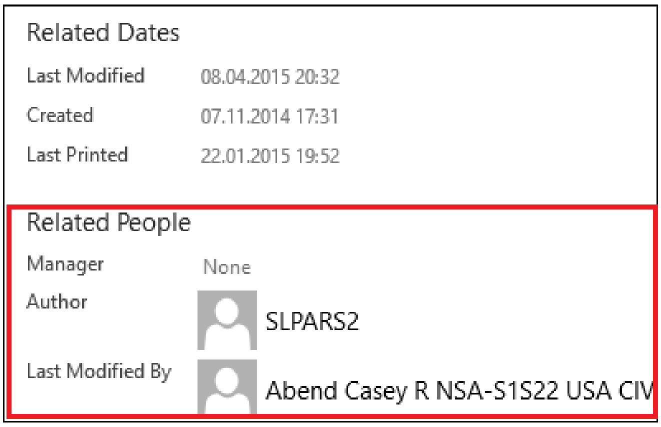
We determined links between documents authors and organizations. Many documents of other companies were uploaded to observed websites which raises many questions. This information might possibly expose partner companies' names, giving some additional attack vectors to attackers. For example, sending spoofed e-mails under the partner's name to the targeted organization.

The information which documents' metadata gives about an organization is significant. Documents' metadata not only describes the content of the document but also the environment where it was created. The information exposes much many details about the organization's internal assets, detects vulnerable targets, shows ways to approach users through links and relationships with other people and organizations. The results shown in this section along with the raw metadata in Appendixes are presented to the responsible CIO-s of observed ministries.

4.4 Attack vectors

When planning to attack someone, it has to be known who the target is and where they are located. Before attacking, cyber criminals often conduct OSINT for gathering information about the target and its assets – it is called reconnaissance stage. In the same stage, attackers could use information hosted on the target's websites' metadata. As presented in the tables of the previous chapter, metadata provides several pieces of information that can be take into account when building attacks. In the following section we describe some of the attack vectors metadata can provide based on the tables shown in previous chapter.

Document author information which metadata exposes helps attackers to learn a possible target name who to attack in order to gain access to corporate networks. Furthermore, exposing usernames through metadata makes attackers' lives much easier gaining hidden working personnel names from metadata. Some organizations do not publish their workers names and information on websites, however, the information is still there in the documents metadata. For example, in Figure 35 there is shown the metadata of a spreadsheet which was downloaded from nsa.gov website. The metadata exposes NSA personnel name which, for instance, is a very good starting point for any attacker to start searching intelligence about the exposed name from social media.


Related Dates

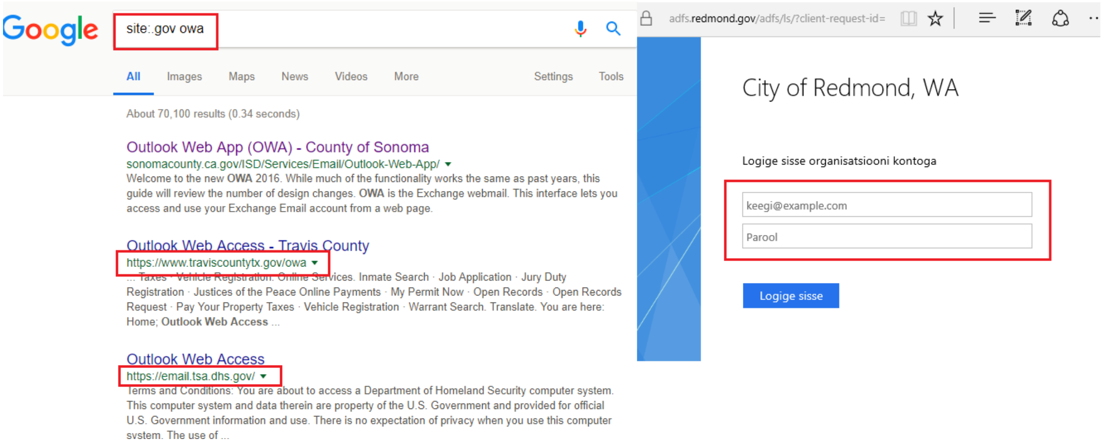
Last Modified	08.04.2015 20:32
Created	07.11.2014 17:31
Last Printed	22.01.2015 19:52

Related People

Manager	None
Author	SLPARS2
Last Modified By	Abend Casey R NSA-S1S22 USA CIV

Figure 35: Metadata disclosed a secret personnel name

Names give options for potential spear-phishing attacks as well as brute-forces attacks on the web services. In Figure 36, results from a simple google search are shown that expose possible webmail login pages. Comparing extracted usernames to names which are leaked already on the internet (for instance Yahoo or Myspace data breach), the attacker could have a possible password for attempting to log on to the targeted organization's services. What is more, usernames can be used for a denial of service attack, for example, locking user accounts by generating failed login attempts on the services available on the internet. For instance, users cannot use webmail for a certain time since their accounts are locked.



site:.gov owa

About 70,100 results (0.34 seconds)

Outlook Web App (OWA) - County of Sonoma
sonomacounty.ca.gov/ISD/Services/Email/Outlook-Web-App/
Welcome to the new **OWA** 2016. While much of the functionality works the same as past years, this guide will review the number of design changes. **OWA** is the Exchange webmail. This interface lets you access and use your Exchange Email account from a web page.

Outlook Web Access - Travis County
https://www.traviscountytx.gov/owa
... Taxes · Vehicle Registration. Online Services. Inmate Search · Job Application · Jury Duty Registration · Justices of the Peace Online Payments · My Permit Now · Open Records · Open Records Request · Pay Your Property Taxes · Vehicle Registration · Warrant Search. Translate. You are here: Home; **Outlook Web Access** ...

Outlook Web Access
https://email.tsa.dhs.gov/
Terms and Conditions: You are about to access a Department of Homeland Security computer system. This computer system and data therein are property of the U.S. Government and provided for official U.S. Government information and use. There is no expectation of privacy when you use this computer system. The use of ...

adfs.redmond.gov/adfs/ls/?client-request-id=

City of Redmond, WA

Logige sisse organisatsiooni kontoga

keegi@example.com

Parool

Figure 36: Webmail logins, place for brute-forcing or denial of service attacks

Internal information, such as domain names, are important for preparing attacks like phishing e-mails or for malware preparation. For example, if an attacker wants to be sure that the victim is located on the targeted organization’s networks, he could prepare malware in a way that it checks system where it runs (checking domain name) and executing if it is on the target system. In addition, internal domain name is good for preparing fake login pages and lure targeted organization personnel to visit those pages.

Server information exposing targets inside the internal network means the attacker does not need to do any extra movement. If the attacker gains access to the internal network, he does not need to start scanning the network because he already has knowledge of the file servers’ location. Old documents which are uploaded on target websites often contain information about old servers and assets. This information can also be used by attackers. For instance, old file servers are often running for archiving purposes or for backups, however, their operating systems are vulnerable and not patched which makes them easy targets.

The links between organizations give attackers information about potential partner organizations. Again, this information can be used for sending phishing e-mails. Furthermore, information on partners gives opportunities to conduct watering-hole attacks, infecting partners’ webpages with specific malware and wait for audience form the targeted organization only. Also, infecting or overtaking partners’ infrastructure, which is probably not so secure, and then sending e-mails from the legitimate e-mail server. Those are common APT practices and tactics that are utilized. Sometimes partner companies manage some of their offered services or hardware over Virtual Private Network (VPN). Should the attackers take over the partners’ networks which are most probably not so well protected, they could get backdoor access to target systems.

When checking timestamps on documents and adding software version information, we can determine the software version number in that particular time frame. When the same user is uploading documents in a certain period and we build a timeline of timestamp information and software versions, we could possibly detect software’s update cycles of the organization. Knowing which software runs in target systems helps attackers select exploits or prepare themselves for privilege escalation if the target computer is already infected. The version information shows the organization’s overall update policies, whether software versions are consolidated, if they are vulnerable, and what type of software is running in the systems.

In Chapter 4 we were able to detect possible links between working personnel. We detected in metadata persons’ names who exist most frequently in metadata fields which give indication that those persons might be the website content managers who upload the documents.

Since those persons always check the documents’ content before uploading, their names appear in metadata. Those content managers are valuable targets for attackers, since they have access to webservers and permissions to upload data there. Infecting website content manager computers with malware gives the attacker permissions to upload, for example, PHP shells on target websites and use that webserver as a bridge for data exfiltration.

Those are some of the attack vectors the attackers could execute based on findings in Chapter 4. As we can notice, the metadata provides significant information for attackers and helps understand the target and its systems. The next chapter presents different options on how this issue can be mitigated and how to remove metadata from documents.

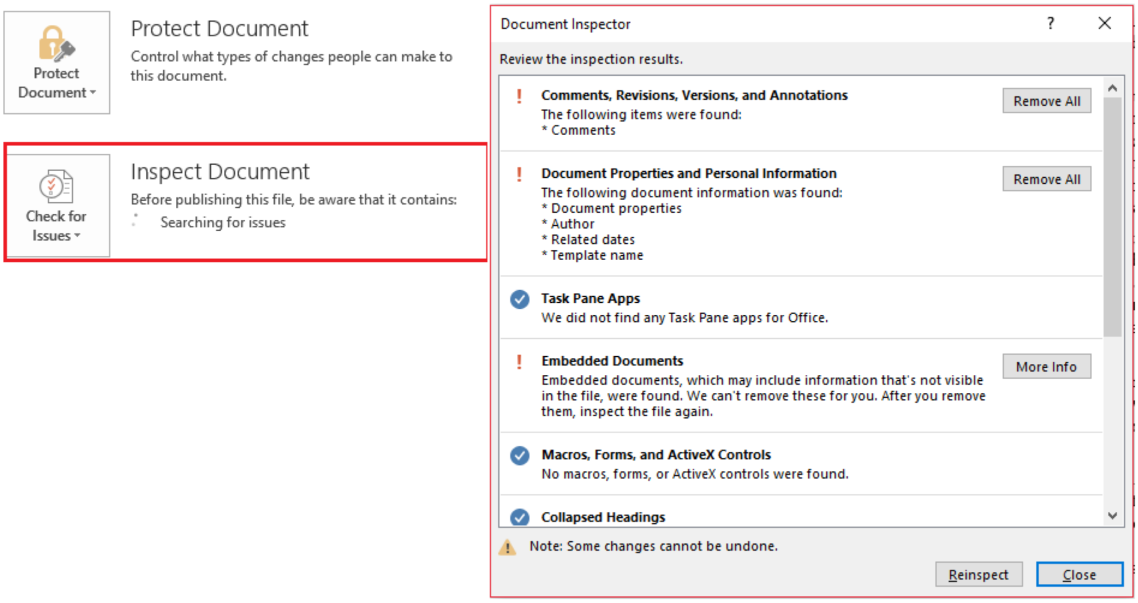
5 Mitigation

In the previous chapters we discussed the existence of metadata in documents and the issues it might bring. The current chapter focuses on mitigating this problem and discusses the methods for removing metadata.

To mitigate harm and risks metadata can bring, it has to be removed from documents. Removing metadata from document is not a very difficult activity. Microsoft Office has built in metadata removal tool “Inspector” which removes most of the hidden information from documents. Acrobat reader also has metadata sanitize functionalities to remove metadata from PDF documents. Even though those softwares are normally accessible for users, metadata still exists in documents. Most people are not fully aware of the metadata existence, so they can unwittingly send confidential information to outside of their organization or publish it on the Internet where everyone has access to it.

In paper [9], there discussed that for handling metadata issues in organizations, a policy regarding metadata has to be in place. That policy should include several aspects such as education of the users and raising awareness. By our observation, from Chapter 4 it can be clearly understood that almost every document which was analyzed contained metadata that could be used on fingerprinting basis. We can assume that the document creators are not actually aware that the documents they create contain some extra information which in some cases can be sensitive and dangerous. Introducing metadata’s nature and its possible contents is one way to start removing the information before publishing.

Document “Inspector”, a built-in MS Office tool for removing metadata and inspecting it, gives feedback about which extra data is located inside document and a choice option whether to keep the information or remove it. For instance, analyzing this thesis paper with document inspector (navigating to File menu and choosing Inspect Document) it detects comments, author information, embeddings, and template names. Figure 37 presents document inspector’s output of this thesis paper.



Protect Document
Control what types of changes people can make to this document.

Inspect Document
Before publishing this file, be aware that it contains:
Searching for issues

Review the inspection results.

! Comments, Revisions, Versions, and Annotations
The following items were found:
* Comments

! Document Properties and Personal Information
The following document information was found:
* Document properties
* Author
* Related dates
* Template name

✓ Task Pane Apps
We did not find any Task Pane apps for Office.

! Embedded Documents
Embedded documents, which may include information that's not visible in the file, were found. We can't remove these for you. After you remove them, inspect the file again.

✓ Macros, Forms, and ActiveX Controls
No macros, forms, or ActiveX controls were found.

✓ Collapsed Headings

Note: Some changes cannot be undone.

Figure 37: Inspecting the current thesis paper with document inspector

Since OOXML documents might contain custom information in its data files, those tools are not always capable of removing all of the sensitive information. One possible way to make sure that metadata does not leak information, is to make and use a special template. In the template one can insert prefilled information showing up in metadata or just erase all and leave it blank. If a document is created, those fields are empty and custom information most likely does not exist.

To maximize metadata removal, the document can be cleaned with document “inspector” and then saved into another format: to PDF or Rich Text Format (RTF). However, those solutions need manual interaction by users. If the user forgets to remove metadata or convert the file to other format, it is still exposed. Since much of the document sharing is done by e-mails, FTP servers, USB drives, or by cloud services, and it can be intense and in huge amounts, then this solution might be not the most efficient. To mitigate those issues, several commercial solutions are available for protecting documents. Microsoft has a solution named RMS [30] which helps to protect electronic assets and it is capable of protecting data on all important devices. There are also several toolkits for anonymizing documents. For example, the PayneGroups Metadata Assistant [31] which can be configured to remove metadata from Word and PDF files on demand or automatically (i. e. when files are sent by email to other parties).

If document authors see the metadata and are aware of its existence in documents, they will most probably remove it. Simson L. Garfinkel [26] proposed an idea to mitigate metadata information leakages: *”A better approach would be to modify tools so that the underlying data model is in line with what’s presented in the user interface—that is, by making it harder for users to produce documents with hidden information*”. If a document is edited, one can see the metadata in its interface and is aware of its existence which might mitigate the awareness problems.

This chapter discussed metadata removing approaches. Removing metadata from documents is not a very difficult activity. The question is about the awareness and missing metadata policies or procedures. The next chapter presents the conclusion of this thesis paper.

6 Conclusions

In this thesis, metadata of documents hosted on public websites were analyzed, in order to extract information for fingerprinting basis and detect information that aids in carrying out cyber attacks against observed organizations.

We observed three Estonian governmental organizations' documents that were uploaded on their websites. We gathered these documents into our local system and conducted metadata extraction and analysis. In Chapter 3 we introduced some novel methods and techniques how to carry out this study. In Chapter 4, we presented the results of the conducted research.

Our analysis demonstrated that metadata existed in almost all the documents that we managed to download and analyze. We processed 2643 documents, from which only 12 documents had metadata properly removed. All the other documents contained metadata fields which describe the environment where the document was created or modified. Metadata extracted from the observed organizations' websites gave us insights to internal networks, internal servers, software and their versions, physical locations of people, internal services, usernames, vulnerable computers and software, and other information. This information was obtained simply by analyzing documents' metadata, no other activities or network scans were done. Metadata is intrusive and can be used as the pre-stage for conducting cyber attacks. The outcome of the metadata extraction and analysis validates our hypothesis that the documents uploaded on the Estonian governmental websites contain compromising information that could be used for cyber attacks. Since the compromising information is available for everyone on public websites, it is nearly impossible to determine whether that information is used for conducting devastating attacks or not.

The high number of metadata existence in documents reflects the need for awareness about metadata vulnerabilities. People are not aware of the metadata existence in electronic documents and might unwillingly leak information about themselves or the organization. Metadata removal processes are not difficult and the Office applications have metadata removal tools built-in; however, those tools are not often used due to the lack of awareness.

This is the first study in Estonia that aims to clean up Estonian governments' webpages from documents' metadata. We presented the results of our analysis to the responsible CIO's to raise awareness about metadata security issues. Talking about metadata and demonstrating its capabilities are a first step in mitigating metadata exposures.

7 References

[1] The world's most valuable resource is no longer oil, but data. *Issued by Economist,* 2017. [Online] https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource. [Accessed: 02.04.2018].

[2] Jeffrey R. Jones. Document Metadata and Computer Forensics. *James Madison University Infosec Techreport Department of Computer Science JMU-INFOSEC-TR-2006-003,* 2006.

[3] Eleni Gessiou, Elias Athanasopoulos, and Sotiris Ioannidis, Institute of Computer Science. Digging up Social Structures from Documents on the Web. *2012 IEEE Global Communications Conference publications,* 2012. [Online] https://pdfs.semanticscholar.org/29d6/8faebd46ef4962e953dbaf440719c8925e38.pdf. [Accessed: 02.04.2018].

[4] Andrea Vance. The Snowden files: What did we learn? *Issued by Stuff,* 2014 [Online] http://www.stuff.co.nz/national/politics/10503457/The-Snowden-files-What-did-we-learn. [Accessed: 03.03.2018].

[5] Orcale White Paper. The Risks of Metadata and Hidden Information, *Issued by Oracle,* 2007 [Online] http://www.oracle.com/technetwork/testcontent/stellent-wp-metadatarisks-0307-134347.pdf. [Accessed: 02.01.2018].

[6] Jenn Riley. Understanding metadata: what is metadata, and what is it for? *A primer publication of the National Information Standards Organization (NISO),* 2017.

[7] NSA spying scandal: what we have learned, 2013 [Online] https://www.theguardian.com/world/2013/jun/10/nsa-spying-scandal-what-we-have-learned. [Accessed: 03.02.2018].

[8] Tuomas Aura, Thomas A. Kuhn, Michael Roe. Scanning Electronic Documents for Personally Identifiable Information. *Proceedings of the 5th ACM workshop on Privacy in electronic society,* 2006, pp 41-50.

[9] Randall Farrar. Metadata: The Hidden Disaster That's Right in Front of You. *A publication of the Senior Lawyers Section of the New York State Bar Association,* 2014. pp 29-33.

[10] Charkes F.Luce, Jr. What's the Matter With Metadata? *The Colorado Lawyer Vol. 36, No. 11,* 2007.

[11] Matt Loney. 'Dodgy-dossier syndrome' rife in the workplace, 2003. [Online] http://www.zdnet.com/article/dodgy-dossier-syndrome-rife-in-the-workplace/. [Accessed: 12.02.18].

[12] Venable's $20-Million-Plus-Sanctions Trade Secrets Win for Government Contractor: What It Means For You. *Issued by Venable* 2015. [Online] https://www.venable.com/venables-20-million-plus-sanctions-trade-secrets-win-for-government-contractor-what-it-means-for-you-10-16-2015/. [Accessed: 02.02.18].

[13] Ionut Arghire. Microsoft Office Has over One Billion Users, 2012. [Online] http://news.softpedia.com/news/Microsoft-s-Office-Has-Over-One-Billion-Users-280426.shtml [Accessed: 01.02.18].

[14] OLE Background. *Issued by Microsoft,* 2016 [Online] https://docs.microsoft.com/en-us/cpp/mfc/ole-background. [Accessed: 10.03.18].

[15] Open XML Formats and file name extensions. *Issued by Microsoft.* [Online] https://support.office.com/en-us/article/Open-XML-Formats-and-file-name-extensions-5200D93C-3449-4380-8E11-31EF14555B18. [Accessed: 01.03.18].

[16] Standard ECMA-376 Office Open XML file Formats. *ECMA International - European association for standardizing information and communication systems,* 2006 [Online] http://www.ecma-international.org/publications/standards/Ecma-376.htm [Accessed: 10.05.18].

[17] File format reference for Office 2013. *Issued by Microsoft,* 2016. [Online] https://technet.microsoft.com/en-us/library/dd797428.aspx. [Accessed: 07.03.18].

[18] Phil Harvey. ExifTool by Phil Harvey. [Online] https://www.sno.phy.queensu.ca/~phil/exiftool/. [Accessed: 23.02.18].

[19] Custom XML. *Issued by Microsoft.* [Online] https://download.microsoft.com/download/5/1/6/5160beb0-ed51-4fa5-9e73-334114e7fb24/CustomXML.pptx. [Accessed: 06.02.18].

[20] E. Taft, J. Pravetz, S. Zilles, L. Masinter. The application/pdf Media Type. *Issued by IETF Network Working Group RFC 3778 Standard,* 2004. [Online] https://tools.ietf.org/html/rfc3778. [Accessed: 29.01.18].

[21] Adobe Systems Incorporated. PDF Reference manual sixth edition. *Issued by Adobe,* 2006. [Online] https://www.adobe.com/content/dam/acom/en/devnet/acrobat/pdfs/pdf_reference_1-7.pdf. [Accessed: 15.01.18].

[22] Larry Pesce. Document Metadata, the Silent Killer. *SANS Institute InfoSec Reading Room,* 2008. [Online] https://www.sans.org/reading-room/whitepapers/privacy/document-metadata-the-silent-killer--32974. [Accessed: 14.02.18].

[23] Chema Alonso, Enrique Rando, Francisco Oca and Antonio Guzmán. Disclosing Private Information from Metadata, hidden info and lost data. *Black Hat Europe 2009 Media Archives,* 2009. [Online] https://www.blackhat.com/presentations/bh-europe-09/Alonso_Rando/Blackhat-Europe-09-Alonso-Rando-Fingerprinting-networks-metadata-whitepaper.pdf. [Accessed: 24.01.18].

[24] Hanno Langweg. OOXML File Analysis of the July 22nd Terrorist Manual. *NISlab Norwegian Information Security laboratory,* 2012. [Online] https://pdfs.semanticscholar.org/4cf2/1053010041fad98cf1d9e915b5dccc7d2334.pdf. [Accessed 03.01.18].

[25] Muhammad Ali Raffay. DATA HIDING AND DETECTION IN OFFICE OPEN XML (OOXML) DOCUMENTS. *Master thesis, University of Ontario Institute of Technology,* 2011. [Online] https://ir.library.dc-uoit.ca/bitstream/10155/146/1/Raffay_Muhammad.pdf. [Accessed: 07.03.18].

[26] Garfinkel Simson L. Leaking Sensitive Information in Complex Document Files - and How to Prevent It. *IEEE Security & Privacy (Volume 12, Issue 1),* 2013. pp 20-27. [Online] https://calhoun.nps.edu/bitstream/handle/10945/42550/Garfinkel_redaction.pdf?-sequence=1&isAllowed=y. [Accessed: 04.01.18]

[27] Lee Allen, Kevin Cardwell. Advanced Penetration Testing for Highly-Secured Environments. *Published by Packt Publishing Ltd,* 2016.

[28] Sudhanshu Chauhan, Nutan Kumar Panda. Hacking Web Intelligence Open Source Intelligence and Web Reconnaissance Concepts and Techniques. *Published by Elsevier Inc,* 2015.

[29] Paterva. What does Maltego do? [Online] https://www.paterva.com/web7/buy/maltego-clients/maltego-ce.php. [Accessed: 03.04.18].

[30] Dan Plastina. The NEW Microsoft Rights Management services Whitepaper. *Issued by Microsoft*, 2013. [Online] https://cloudblogs.microsoft.com/enterprisemobility/2013/07/31/the-new-microsoft-rights-management-services-whitepaper/. [Accessed: 12.03.18].

[31] Metadata Assistant [Online] http://www.thepaynegroup.com/products/metadata/. [Accessed: 08.04.18].

Appendix

The appendix includes all the additional material developed through this thesis. It is divided into 5 parts. They are the following:

I. **Office files represented in EE domain.** This appendix describes methods collecting statically information about existent of Offices files in Estonia domain.

II. **OOXML metadata files core.xml and app.xml.** This appendix presents contents of typical OOXML document files which contains metadata.

III. **Ministry A extracted raw metadata.** This appendix describes extracted raw metadata which were gathered form Ministry A documents.

IV. **Ministry B extracted raw metadata.** This appendix describes extracted raw metadata which were gathered form Ministry B documents.

V. **Ministry C extracted raw metadata.** This appendix describes extracted raw metadata which were gathered form Ministry C documents.

I. Office files represented in EE domain

The following research was conducted to understand which document formats exists mostly in Estonia websites – in .ee top level domain.

Research was conducted 19.02.18 using Google search engine and its features which support queries of specific file type or extension.

Example of querying multiple file types together from Estonia domain is shown below:

site:ee ext:doc | ext:docx | ext:rtf | ext:xls | ext:xlsx | ext:ppt | ext:pptx | ext:odf | ext:odp| ext:ods| ext:pdf | ext: bdoc

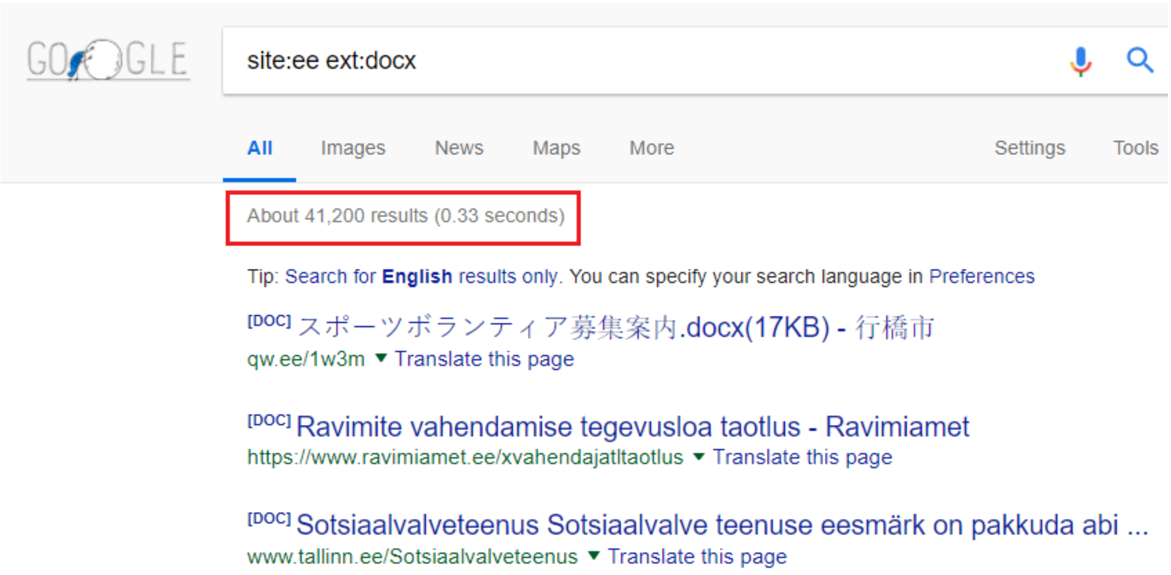
Description of used google search operators:

site:ee – Google advanced search operator, is used for searching from a specific domain or website. In our case it represents querying from EE domain.

ext – Google advanced search operator, represents extension modifier, works same way as file type.

| - Google advanced search operator, representing logical OR.

Since we were interested in which file type is the most popular in Estonia domain, we are building query each file type separately: *site:.ee ext:doc*, after that we are marking down response results and take the next file type (*for example site:.ee ext:docx*). The following illustrates query described before: *site:ee ext:docx* Google finds 41200 positive matches:



site:ee ext:docx

About 41,200 results (0.33 seconds)

Tip: Search for **English** results only. You can specify your search language in Preferences

[DOC] スポーツボランティア募集案内.docx(17KB) - 行橋市
qw.ee/1w3m ▼ Translate this page

[DOC] Ravimite vahendamise tegevusloa taotlus - Ravimiamet
https://www.ravimiamet.ee/xvahendajatltaotlus ▼ Translate this page

[DOC] Sotsiaalvalveteenus Sotsiaalvalve teenuse eesmärk on pakkuda abi ...
www.tallinn.ee/Sotsiaalvalveteenus ▼ Translate this page

Figure 38: Gathering statistical information about documents occurrences in EE domain

Queries conducted:

Table 15: Query responses

Query	File extension	Count
site:ee ext: doc	doc	277000
site:ee ext:docx	docx	41200
site:ee ext:rtf	rtf	28200
site:ee ext:xls	xls	43000
site:ee ext:xlsx	xlsx	13400
site:ee ext:ppt	ppt	18100
site:ee ext:pptx	pptx	6200
site:ee ext:odt	odt	2000
site:ee ext:odp	odp	548
site:ee ext:ods	ods	951
site:ee ext:bdoc	bdoc	1080
site:ee ext:ddoc	ddoc	811
site:ee ext:pdf	pdf	4220000

For getting a better visual overview of document types on Estonian domain, PDF format is not included in the chart:

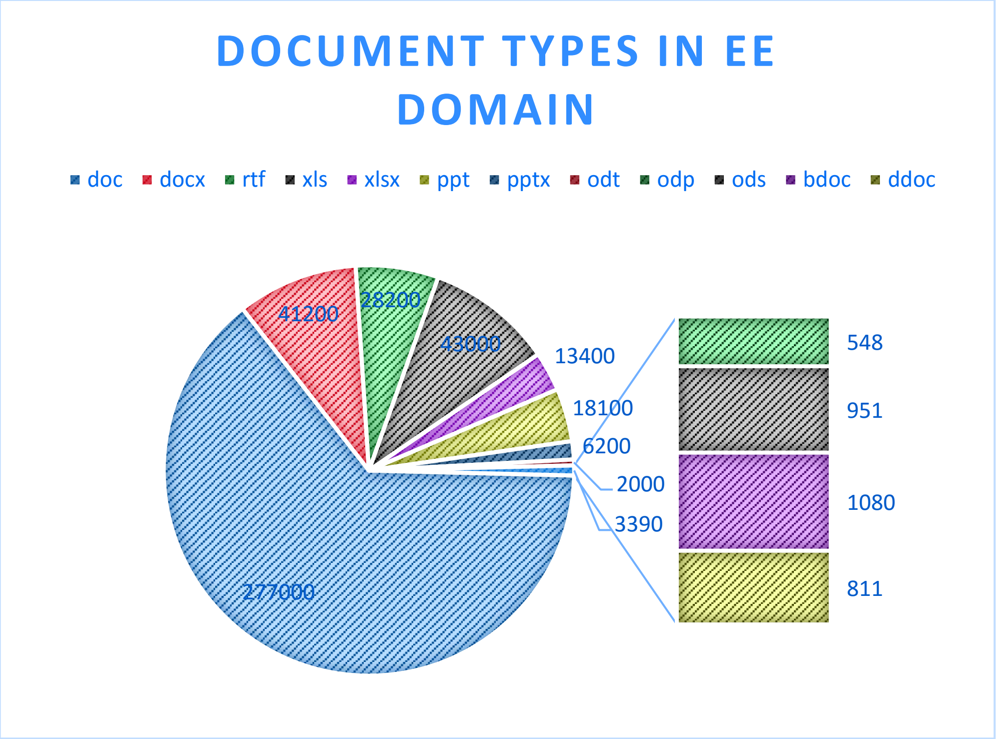


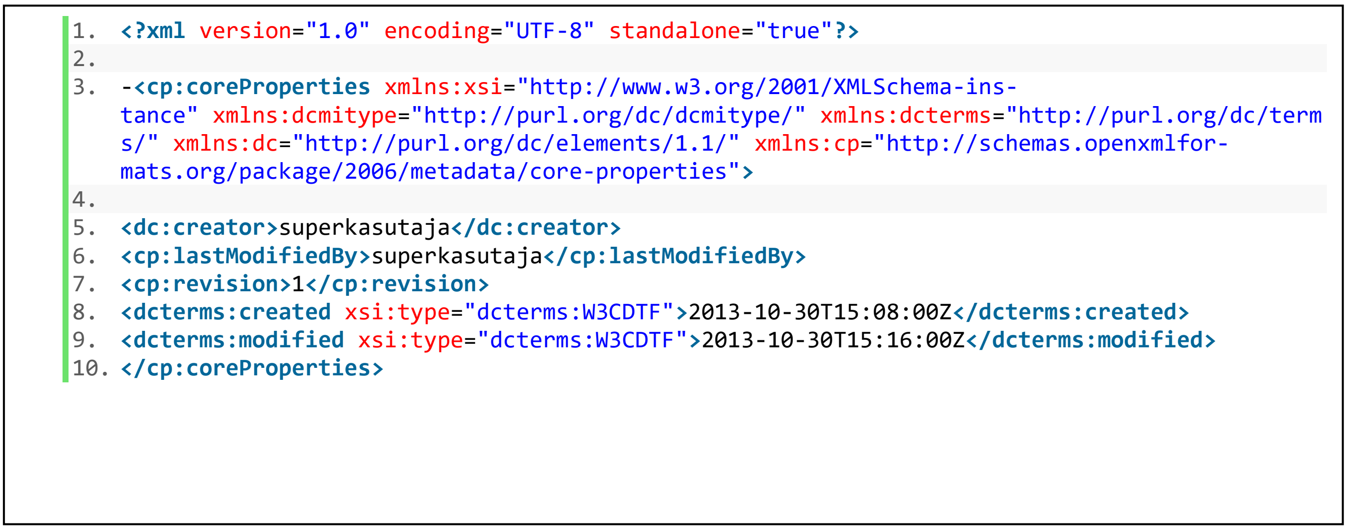
Figure 39: Graph of document occurrences in EE domain, PDF files are excluded

The most popular documents which one can find from websites are PDF documents, followed by different types of Microsoft Office documents.

II. OOXML metadata files core.xml and app.xml

This appendix presents two XML files which are found in every OOXML and which contains metadata. Those files are *core.xml* and *app.xml*.

The Word Documents were downloaded from TTU site: http://cloud.ld.ttu.ee/idu0010/Portals/0/Harjutustunnid/Financial%20Analysis.docx and opened with 7zip program to examine metadata XML files inside OOXML structure. The file name under observation is *Financial Analysis.docx* and core.xml, app.xml are presented below:

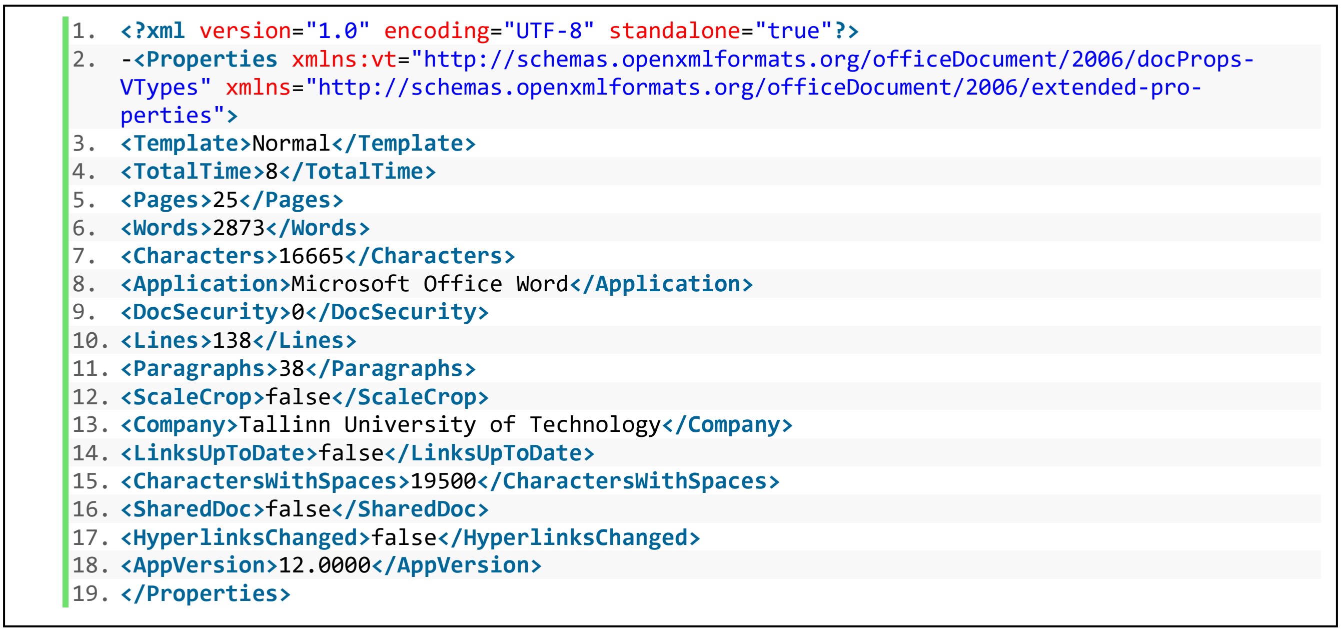


```
<?xml version="1.0" encoding="UTF-8" standalone="true"?>

-<cp:coreProperties xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dcmitype="http://purl.org/dc/dcmitype/" xmlns:dcterms="http://purl.org/dc/terms/" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:cp="http://schemas.openxmlformats.org/package/2006/metadata/core-properties">

<dc:creator>superkasutaja</dc:creator>
<cp:lastModifiedBy>superkasutaja</cp:lastModifiedBy>
<cp:revision>1</cp:revision>
<dcterms:created xsi:type="dcterms:W3CDTF">2013-10-30T15:08:00Z</dcterms:created>
<dcterms:modified xsi:type="dcterms:W3CDTF">2013-10-30T15:16:00Z</dcterms:modified>
</cp:coreProperties>
```

Figure 40: Content of core.xml file



```
<?xml version="1.0" encoding="UTF-8" standalone="true"?>
-<Properties xmlns:vt="http://schemas.openxmlformats.org/officeDocument/2006/docPropsVTypes" xmlns="http://schemas.openxmlformats.org/officeDocument/2006/extended-properties">
<Template>Normal</Template>
<TotalTime>8</TotalTime>
<Pages>25</Pages>
<Words>2873</Words>
<Characters>16665</Characters>
<Application>Microsoft Office Word</Application>
<DocSecurity>0</DocSecurity>
<Lines>138</Lines>
<Paragraphs>38</Paragraphs>
<ScaleCrop>false</ScaleCrop>
<Company>Tallinn University of Technology</Company>
<LinksUpToDate>false</LinksUpToDate>
<CharactersWithSpaces>19500</CharactersWithSpaces>
<SharedDoc>false</SharedDoc>
<HyperlinksChanged>false</HyperlinksChanged>
<AppVersion>12.0000</AppVersion>
</Properties>
```

Figure 41: Content of app.xml

III. Ministry A extracted raw metadata

The following tables present extracted metadata from the documents of Ministry A. Metadata is extracted utilizing fingerprinting method which was described in Chapter 3.

Information which is presented in the tables below are unique values, duplicates are removed.

Table 16: Ministry A local and network paths:

Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons

Table 17: Ministry A Server paths

Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons

Table 18: Ministry A list of URLs found in metadata

Raw metadata is removed for privacy reasons

Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons

Table 19: Ministry A list of software's and applications which were extracted from PDF documents

Creator Tool	Producer
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons

Table 20: Ministry A company and printer information, extracted from Microsoft Office documents

Company	Printer name
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons

Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
	Raw metadata is removed for privacy reasons
	Raw metadata is removed for privacy reasons

IV. Ministry B extracted raw metadata

The following tables present extracted metadata from the documents of Ministry B. Metadata is extracted utilizing fingerprinting method which was described in Chapter 3.

Information which is presented in the tables below are unique values, duplicates are removed.

Table 21: Ministry B Local and network paths

Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons

Table 22: Ministry B list of software's and applications which were extracted from PDF documents

Producer	Creator Tool
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	

Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	

Table 23: Ministry B company and printer information, extracted from Microsoft Office and PDF documents

Company	Printer
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons

V. Ministry C extracted raw metadata

The following tables present extracted metadata from the documents of Ministry C. Metadata is extracted utilizing fingerprinting method which was described in Chapter 3.

Information which is presented in the tables below are unique values, duplicates are removed.

Table 24: Ministry C local and network paths

Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons

Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons

Table 25: Ministry C list of software's and applications which were extracted from PDF documents

Producer	Creator Tool
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons

Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons

Table 26: Ministry C company and printer information, extracted from Microsoft Office documents

Company	Printer
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons

Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	Raw metadata is removed for privacy reasons
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	
Raw metadata is removed for privacy reasons	

License

Non-exclusive licence to reproduce thesis and make thesis public

I, **Karl Mendelman**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

of my thesis

Fingerprinting a Organization Using Metadata of Public Documents,

supervised by Olaf Maennel, Raimundas Matulevicius

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **21.05.2018**