Cyber intelligence for anti-money laundering, counter-terrorism financing and know your customer

Kristian Vepsäläinen

Master’s thesis

March 2024

Information and Communication Technology

Master’s Degree Programme in Information Technology, Cyber Security (YAMK)

jamk | Jyväskylän ammattikorkeakoulu
University of Applied Sciences

Kristian Vepsäläinen

Cyber intelligence for anti-money laundering, counter-terrorism financing and know your customer

Jyväskylä: JAMK University of Applied Sciences, March 2024, 47 pages.

Information and Communication Technology. Master's Degree Programme in Information Technology, Cyber Security (YAMK). Master's thesis.

Permission for web publication: Yes

Language of publication: English

Abstract

In today's society, the regulation of the financial sector is constantly increasing, especially about know-your-customer (KYC) and anti-money laundering (AML). The measures related to this regulation required more and more personnel and caused costs for both financial institutions and their customers.

One way to solve this problem was digital open source intelligence (CYBINT). This method has not been used much in the financial sector. Since most companies and people had some digital content online, it was an effective way to approach the problem.

Due to banking and insurance secrecy, it was not possible to use an identifiable, real customer in the experiments, but the tests had to be done anonymously. Both anonymous personal and anonymous business customers were used in the test.

In the tests, a Windows laptop was used, which ran Kali Linux both from a USB stick and from a virtual machine, so that both methods could be tested. The tests used Kali Linux's standard tools and their free versions.

As a result of the tests, both files containing individual observations and broader links between different data were obtained. The Maltego program turned out to be particularly useful because it already produced the links between the data and their type.

Based on the tests, it was found that the information obtained is useful in the use of AML/KYC activities, but the biggest challenge was considered to be the culture of the financial sector and the resulting difficulties in introducing new methods.

Keywords/tags (subjects)

case study, customer identification, Kali Linux, money laundering, open source intellingence

Miscellaneous (Confidential information)

Kuvailulehti

Kristian Vepsäläinen

Cyber intelligence for anti-money laundering, counter-terrorism financing and know your customer

Jyväskylä: Jyväskylän ammattikorkeakoulu. Maaliskuu 2024, 47 sivua.

Tietojenkäsittely ja tietoliikenne. Kyberturvallisuuden maisteriohjelma. Opinnäytetyö YAMK.

Verkkojulkaisulupa myönnetty: kyllä

Julkaisun kieli: englanti

Tiivistelmä

Tämän hetken yhteiskunnassa finanssialan säätely kasvoi kokoajan, erityisesti asiakkaan tuntemiseen (KYC) ja rahanpesun estämiseen (AML) liittyen. Tähän säätelyyn liittyvät toimenpiteet vaativat kokoajan enemmän henkilöstöä ja ne aiheuttivat kuluja sekä finanssilaitoksille että heidän asiakkailleen.

Yksi tapa ratkaista tämä ongelma oli digitaalisten avointen lähteiden tietonhankinta (CYBINT). Tätä menetelmää ei ollut käytetty juurikaan finanssialalla. Koska suurimmalla osalla yrityksistä ja ihmisistä oli jotain digitaalista sisältöä netissä, se oli tehokas tapa lähestyä ongelmaa.

Pankki- ja vakuutussalaisuuden vuoksi kokeiluissa ei voinut käyttää tunnistettavaa, oikeaa asiakasta, vaan testit tuli tehdä anonyymisti. Testissä käytettiin sekä anonyymia henkilö- että anonyymiä yritysasiakasta.

Testeissä Käytettiin Windows-läppäriä, jolla ajettiin Kali Linuxia sekä usb-tikulta että virtuaalikoneesta, jotta molempia tapoja voitiin testata. Testeissä käytettiin Kali Linuxin vakiotyökaluja ja niiden ilmaisversioita.

Testien tuloksena saatiin sekä tiedostoja, joissa oli yksittäisiä havaintoja että laajempia linkityksiä eri tietojen välillä. Erityisen hyödylliseksi osoittautui Maltego-ohjelma, koska se tuotti valmiiksi tietojen väliset linkitykset ja niiden tyypin.

Testien perusteella havaittiin, että saadut tiedot ovat hyödyllisiä AML/KYC-toiminnan käytössä, mutta suurimpana haasteena pidettiin finanssialan kulttuuria ja tästä tulevia hankaluuksia uusien menetelmien käyttöönotossa.

Avainsanat (asiasanat)

avointen lähteiden tiedustelu, asiakkaan tunnistaminen, Kali Linux, rahanpesu, tapaustutkimus

Muut tiedot (salassa pidettävät liitteet)

Contents

Figures

Tables

Acronyms

AI artificial intelligence

AML anti-money laundering

AMLD4 Fourth Anti-Money Laundering Directive

AMLD5 Sixth Anti-Money Laundering Directive

AMLD5 Fifth Anti-Money Laundering Directive

API application programming interface

AS autonomous system

ASN autonomous system number

BI business intelligence

BIOS Basic Input-Output System

CI counter-intelligence

CPU Central Processing Unit

CTF Counter-terrorism financing

CV curriculum vitae

CYBINT cyber intelligence

DLP data leakage protection

DNS Domain Name System

DVD Digital Video Disc or Digital Versatile Disc

ECS Experimental Computer Sciences

EU the European Union

Exif Exchangeable image file format

FIN-FSA Finnish Financial Supervisory Authority

Finansinspektionen Swedish Financial Supervisory Authority

FOCA Fingerprinting Organizations with Collected Archives

GDP gross domestic product

GDPR General Data Protection Regulation

GEOINT geospatial intelligence

GUI graphical user interface

HR human resources

HTML HyperText Markup Language

HTTP Hypertext Transfer Protocol

HTTPS Hypertext Transfer Protocol Secure

HUMINT human intelligence

I2P Invisible Internet Project

I&W indications and warnings

IM instant messaging

IMDP International Movie Database

IMINT imagery intelligence

IP Internet Protocol

IPQS IPQualityScore

IPv4 Internet Protocol version 4

IPv6 Internet Protocol version 6

JSON JavaScript Object Notation

KYC know-your-customer

LITINT literature intelligence

MAC address media access control address

MASINT measurement and signature intelligence

MAT metadata anonymization toolkit

MI Military intelligence

MiFID 2 Markets in Financial Instruments directive 2

ML Machine learning

NATO North Atlantic Treaty Organization

NLP Natural language processing

NPO Nonprofit organization

NSA National Security Agency

OECD Organisation for Economic Co-operation and Development

OPSEC operation security

OS operating system

OSINF Open Source Information

OSINT open source intelligence

PSD2 Payment Services Directive 2

RAM Random Access Memory

RIPE NCC Réseaux IP Européens Network Coordination Centre

SAS Statistical Analysis System

SEI Software and Engineering Institute

SIEM Security Information and Event Management

SIGINT signal intelligence

SMTP Simple Mail Transfer Protocol

SOCMINT social media intelligence

Tails The Amnesic Incognito Live System

Tor The Onion Router

UEFI Unified Extensible Firmware Interface

UK the United Kingdom

USA the United States of America

USB Universal Serial Bus

VAT value-added tax

VM virtual machine

VPN virtual private network

WEBINT web intelligence

XML Extensible Markup Language

1 Introduction

This work proves that cyber intelligence (CYBINT) is a worthwhile process in the financial sector. In the financial sector, my point of view is anti-money laundering (AML) and know-your-customer (KYC) because there are a lot of new and upcoming regulations in those two fields. A good example of those regulations is The European Union's Anti-money-laundering Package 2021. (Binder, 2021)

CYBINT is one new way to solve problems that come from regulations. This means that investigators can use machine learning, artificial intelligence (AI), social engineering etc. (Troia, 2020) methods to AML/KYC. Also, one big issue is to use new, non-traditional data sources like social media, metadata, or dark web (Gibson, 2017) . An old and traditional way to do AML/KYC is to ask things from customers and check things from authority registers like the civil registry or company register.

In the middle of this thesis writing process, Russia attacked Ukraine, and Western countries put sanctions on Russia and Belarus. This makes the topics of this thesis more actual. This crisis shows why we need CYBINT for the financial sector.

The research questions are "Is it possible to use CYBINT methods and tools for AML/KYC purposes?" and "What kind of relevant information CYBINT can offer from an AML/KYC point of view?"

My research is Experimental Computer Sciences (ECS) study. My method for this work is a case study. In the case study, the researcher analyzed defined problems in real situations and used real information as a methodological tool. (Herrera et al., 2016)

So, I do the first literature review and after that, I do a case example where I do a hands-on project on how to use CYBINT from an AML/KYC perspective. References for this thesis have been found to use search engines and also get tips from colleagues.

From an ethical point of view, this work is a case study so there are no big concerns for data handling, etc. On the other hand, there are two case examples where I use real-world examples and there I must do pseudonymization for privacy reasons. This is possible because one key aspect of CYBINT operations is that those are "secret" and the target does not know that he or she is the target(Suriadi et al., 2016; Chauhan and Panda, 2015, p. 166-167; Hassan and Hijazi, 2018, p. 21-22; Pocher, 2020; Troia, 2020, p. 25-26; p. 72-75 Revell et al., 2017, p. 155-156, 159–160).

The Jyväskylä University of Applied Science’s thesis guidelines say that "the main focus is on working life development", so the cases must be as relevant as possible for the work-life perspective. (Jyväskylä University of Applied Science, n.d.) For reliability, this thesis has a good literature review, but the case study is not as reliable as the bigger statistical research.

In the second section of this thesis, I define the terms "CYBINT", "AML" and "KYC". In the third section, I do a general, deeper diving into CYBINT and I focus on CYBINT methods, data sources, tools, and operational security. After that, we have a general view of how CYBINT works and what must be kept in mind at CYBINT projects. In this thesis, my focus is the European Union (EU) regulation and the EU perspective on AML/KYC. In the real world, in many cases, financial institutions have to obey both EU and the United States of America (USA) regulations.

Fourth section I focus on CYBINT on the financial sector. There I first write a few words about legal aspects. Next, I tell how CYBINT has been used in the financial sector and how it should be used. After that, I focus on what kind of CYBINT project we can do in the AML and KYC fields.

In the fifth section, I take on case examples and show concretely how to do the CYBINT project from an AML/KYC perspective. In this section, I use all tools introduced in section four, but I report only relevant parts.

2 Terms and defenitions

2.1 Cyber Intelligence (CYBINT)

There is not one unique and global definition for Cyber intelligence. In this thesis, Cyber intelligence (CYBINT) is intelligence where data sources are in digital format and typically available from online (Bonfanti, 2018; Kandiko, 2018) .

Cyber intelligence (CYBINT) is part of open source intelligence (OSINT). Other part of OSINT are signal intelligence (SIGINT), imagery intelligence (IMINT), measurement and signature intelligence (MASINT), geospatial intelligence (GEOINT), and human intelligence (HUMINT) (Kandiko, 2018; Şahin, n.d.) . Some other organizations also add counter-intelligence (CI) and operation security (OPSEC) under OSINT (Kernan, 2001, p.3) . Hassan and Hijazi write lots of operation security, but in the definitions with the uses, there is no specific term OPSEC (Hassan and Hijazi, 2018, p.21-49) .

On the other hand, many organizations, like North Atlantic Treaty Organization (NATO) (Nato, 2002) , and sources use the term "OSINT" as a synonym of "CYBINT" (Hassan and Hijazi, 2018) . Some writers also use the terms "Online intelligence" (Troia, 2020) and web intelligence (WEBINT) (Tabatabaei and Wells, 2017) . As an opposite this, the "traditional" intelligence based on journals, books, reports, etc. "physical material" can be called name literature intelligence (LITINT) (Tabatabaei and Wells, 2017) . If OSINT is used for military purposes, it is called the name Military intelligence (MI). (Norton, 2011)

OSINT is intelligence that is produced from publicly available information promptly to a selected audience for a specific purpose (Hassan and Hijazi, 2018, p. 2) (Kernan, 2001, p. v) . OSINT itself is a very old area. It has been used since before biblical and ancient times. Before and during World War times, newspapers, pamphlets, interviews, and paper maps were the most important OSINT sources. During the Cold War era, electronic sources like TV and radio broadcasts became important parts of OSINT sources. (Norton, 2011)

CYBINT is part of OSINT which is used for online information (Hassan and Hijazi, 2018, p. 5-6) (Şahin, n.d.) . So, the biggest difference between OSINT and CYBINT is the source of information - at CYBINT it is the Internet and other digital sources like photos, videos, maps, and metadata. On the other hand, at OSINT, investigators can use sociological data collection methods like interviews or observations. This kind of procedure can be called the name non-technical OSINT (Handnagy, n.d., p. 34-39) .

Depending on its scope, CYBINT can be strategic, tactical, or operational. The difference is in the scope, strategic CYBINT focuses long long-term phenomenons like political, social, or economic trends. Tactical CYBINT focuses on the intelligence of cyber threads. Operational CYBINT has interested long-term things and threads to an organization. It produces information that can be used for daily work. Those three classes are more academic, so in real-life CYBINT cases, there are no clear limits between them. Depending on the case there can be parts for all of them (Kandiko, 2018) .

There can be different costs (technical, human resources, skills, tools, etc.) for different types of OSINT. WEBINT is normally quite cheap from a technical point of view but it requires special skills for the investigator. On the other hand, LITINT is technically cheap and doesn’t require special skills but requires lots of human resources and signal intelligence requires lots of technical cost and special skills (Europol, 2002) .

In the Cold War era, 90 percent of all intelligence information was based on military or other "secret" organizations’ information and 10 percent open data. Today, 90 percent of intelligence data comes from open sources and 10 percent secret material. Open-source information is cheaper and faster to use than secret information (Staniforth, 2017) . Nowadays, because of simple and cheap technology, even teenage children can do effectively CYBINT. There are also CYBINT challenges in which everybody can take part (Yari et al., 2020) .

In the 2020’s most common use cases for cyber intelligence are (Pastor-Galindo et al., 2020)

- Cybercrime and organized crime
  - Spot illegal actions
  - Retrieve suspicious traces
  - Monitor malicious groups
- Social opinion and sentiment analysis
  - Marketing
  - Political campaigns
  - human resources (HR) recruiting
  - Journalism
- Cybersecurity and cyberdefence
  - Foot printing
  - Forensics analysis
  - Cyberattack attribution

- Social engineering/phishing attack prevention

For the Cybersecurity context, CYBINT comes very near to data science and uses the same methods as data science like data mining or traditional statistical analysis methods like correlation analysis, regression analysis, or outlier detection. Some systems like Security Information and Event Management (SIEM) use CYBINT techniques to analyze cyber threats. On the other hand, social engineering gives possibilities to do non-ethical and even illegal things like fake news and deepfake videos (Pastor-Galindo et al., 2020) .

2.2 Anti-money laundering (AML)

The first step is to define AML. Money laundering is any action where someone wants to hide amounts of money or other properties' original background and show it as legal. It is very important to notice that there can be many non-financial sector subjects like games or social media (Andersén, 2020, p.17-19) . Simply, Anti-money laundering is the set of controls to use to prevent anti-money laundering (Andersén, 2020, p.29) .

Nowadays, money laundering is a way to collect a profit from other crimes and/or to way to get money to do other crimes. So, it is almost always linked to other crimes. Understanding those linked crimes also helps surveillance and find money laundering activity (Andersén, 2020, p. 17-21) .

Money laundering is big business, its value is ca. 2-5 percent of the whole world's gross domestic product (GDP), so from 1600 to 4000 billion euros. AML processes have become more monitored and regular after 9/11 (Andersén, 2020, p.29) . In Finland in the year 2018, we had 39 220 money laundering reports, and those were 1 295 526 events. Those numbers are increasing. Most reports come from gambling companies (Andersén, 2020, p.35) .

For those reasons, the risk-based approach is the key aspect of AML. The focus is that all financial institutions do the right risk estimations at the right time. The most important thing for a risk-based approach is the get useful right-timed information. Without it, it is impossible to do a reliable, effective, and convincing AML process. (Andersén, 2020, p. 63-68)

In Finland, we have cases where banks have to get big penalties for the lack of AML. The latest case is S-pankki getting about 1,6 million euros fine for Finnish Financial Supervisory Authority (FIN-FSA) (the Financial Supervisory Authority (FIN-FSA), 2021) . In Sweden, Swedish Financial Supervisory Authority

(Finansinspektionen) has given even bigger penalties for breaking AML regulations. (Ahlander and Dickson, 2015)

2.3 Know-your-customer (KYC)

KYC is the principle that financial service providers must know their customers' backgrounds. The financial service provider must know who is their customer and what this customer's typical way to do financial actions is. Different customers' "typical" behavior means different things. Some customers pay bills every day to go abroad while some customers never do that. In this context, a customer can be a person or organization (Andersén, 2020, p.57-62) .

Traditionally the way to collect data from customers has been asking those things from customers and forcing them to bring documents and/or use information that arises when a customer uses financial services. How much data is needed to know your customer? Again, it depends on the case and what are customer's intentions. Customers can be high-risk customers or low-risk customers. The risk status of the customer can change over time. One that thing is out of financial institutions' or customers' control is sanctions and other political restrictions. Those things are guidelines that are forced to follow without exceptions (Andersén, 2020, p.71-110) . The most important regulations for the KYC field are Fourth Anti-Money Laundering Directive (AMLD4), Fifth Anti-Money Laundering Directive (AMLD5), Sixth Anti-Money Laundering Directive (AMLD5), Payment Services Directive 2 (PSD2), Markets in Financial Instruments directive 2 (MiFID 2), and General Data Protection Regulation (GDPR) (Christensen, 2019) .

2.4 Counter-terrorism financing (CTF)

Counter-terrorism financing (CTF) is an action where someone collects money directly or indirectly to make it possible (to someone other) to make terror attacks (Andersén, 2020, p.17-19) . After 9/11 there have come big advances for counter-terrorism financing. The most effective way to prevent terrorism is to prevent terrorists from getting money (Andersén, 2020, p.29) . Often, money comes from legal actions, but it is used for illegal activities. Prevent terrorism financing is multinational action with requires very coordinated international cooperation (Staniforth, 2017) . Many countries like the United Kingdom (UK) is spacially units for this purpose. (Akhga, 2017)

In many cases, terrorist groups use very sophisticated techniques like social engineering (Handnagy, n.d., p.306-307) , dark web (Kalpakis et al., 2017) , blockchains, and cryptocurrencies to hide the origin of the money (Andersén, 2020, p.174-175) . In some cases, even countries and their governments help directly or indirectly terrorist groups to get money for religious, ethical, or political reasons. Typically, money comes from Western countries and it transfers to developing countries. For those reasons counter-terrorism finance activities have a big overlap with know-your-customer, for example so-called ‘Politically Exposed Persons’ (Dean et al., 2013) .

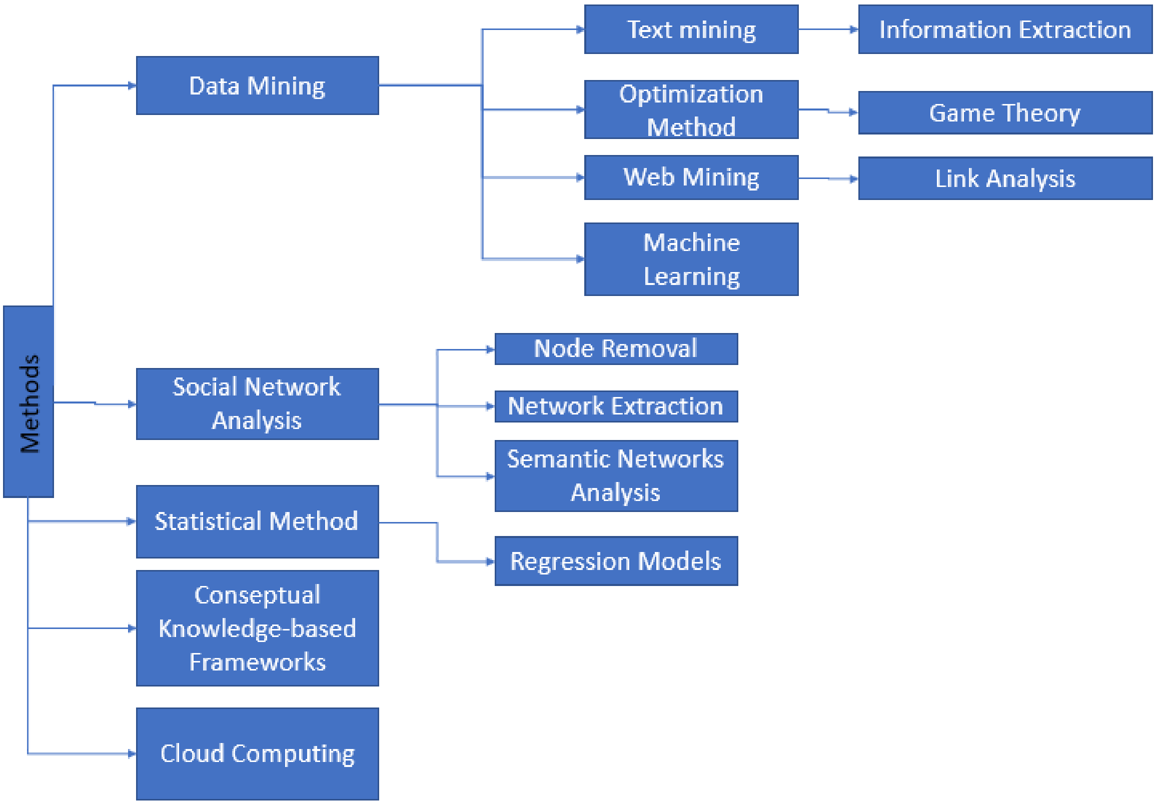


Figure 1: The categorization of CYBINT methods and tools. Adapted from (Tabatabaei and Wells, 2017)

3 CYBINT methods, tools, and data sources

3.1 Methods

There are several different methods in CYBINT. It is important to choose the right method for the right situation. The suitable method depends on what we want to know and what kind of information we have (Kernan, 2001, p.1) . Methods are synthesis for different fields of science like data science, statistics, machine learning, programming, databases, and computer science. But there isn't the all-purpose theory of CYBINT or a list of all possible methods of CYBINT (Ponder-Sutton, 2016, p. 14-15)

.

One important method is the Software and Engineering Institute (SEI) model. That institute is located at the Carnegie Mellon University. That SEI model has 5 steps: (Kandiko, 2018)

1. Terminate the scope of CYBINT and questions where we want answers.
2. Exploring data sources and data collection
3. The "functional analysis" where we want to answer questions "what" and "how"
4. The "strategic analysis" where we want to answer questions "why" and "why"
5. Report findings to target audience and collect feedback

The most important CYBINT methods and tools are categorized in figure 1. Because most of CYBINT information is text-based, one of the key things is to analyze text data. There are many different methods to do that. It is impossible for humans can read everything so, there must be other ways. One much-used method is Natural language processing (NLP) (Gibson et al., 2017, pp. 96-102) .

In the CYBINT field, we are often interested in one specific person. This means that we have been capable of linking information for that person. So, then anonymization and de-anonymization play a big role. Methods with work today to protect people's data, won't work in the future. So this is a race between de- and reidentification. (Ponder-Sutton, 2016, p. 15-16) For many cases, where investigators are exciting links and connections between different people and/or organizations, graphs, and network theory are very useful and widely used methods (Perez and Germon, 2016) .

The intelligence cycle as shown in figure 2 is the typical way to describe the intelligence process. It is a continuous process that with starts direction and ends with feedback. Depending on feedback it can always start again because the result of the intelligence process often brings new intelligence needs. (Gibson, 2017)

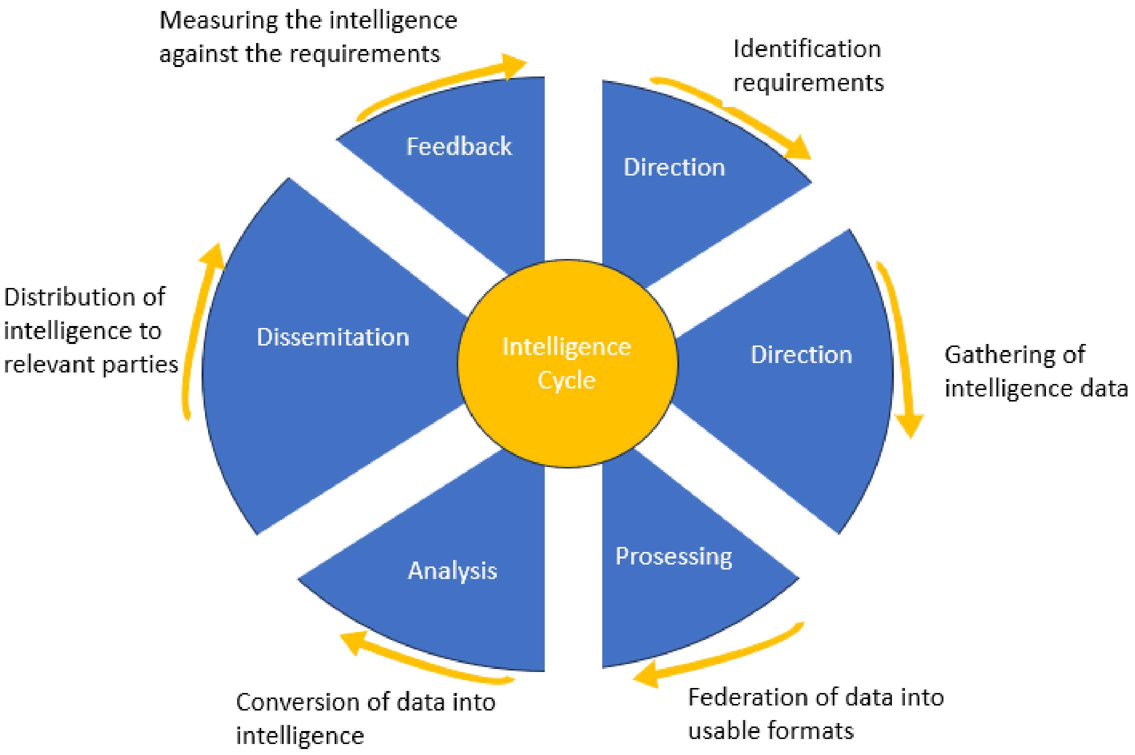

Measuring the intelligence against the requirements
Identification requirements
Feedback
Direction
Distribution of intelligence to relevant parties
Dissemitation
Intelligence Cycle
Direction
Gathering of intelligence data
Analysis
Prosessing
Conversion of data into intelligence
Federation of data into usable formats

Figure 2: The intelligence cycle. Adapted from (Gibson, 2017)

The first step of the intelligence process is "Direction". That part of the process defines intelligence questions so what are things that wanted to find out about using intelligence tools and methods. It also defines the required resources and schedule. In the second step called name "Collection", the investigator makes a plan for data collection including needed levels of intelligence (CYBINT, IMINT, HUMINT, etc.), and collects data. "Processing" is the third step. This step makes translations of raw data from foreign languages, and evaluation of relevance, quality, and reliability of data. The fourth part of the process is "Analysis". This means that the investigator combining data, uses different methods, frameworks, and tools, like mathematical and statistical methodologies, to get new valuable knowledge. (Staniforth, 2017) Typical analysis techniques are lexis analysis, semantic analysis, geospatial analysis, and social media analysis. (Pastor-Galindo et al., 2020)

The second last step is "Dissemination". In this part, the investigator makes reports and other intelligence products depending on the requirements of the intelligence organization or community. A typical example of an intelligence product is an indications and warnings (I&W) bulletin. The last part of the process is "Feedback". In this step results and reports are sent to the customer and the customer gives feedback. Based on the feedback process may start again and there come issues that need more investigation (Gibson, 2017) (Staniforth, 2017) .

In the open source intelligence process, the biggest bottleneck is steps three and four because nowadays the digital world has much information and there is a huge challenge to separate relevant and non-relevant information (Staniforth, 2017) .

At the table 1, there are CYBINT's most common pros and cons (Pastor-Galindo et al., 2020) .

Pros	Cons
Huge amount of available information	Complexity of data management
High capacity of computing	Unstructured information
Big data and machine learning	Misinformation
Complementary types of data	Data sources reliability
Flexible purpose and wide scope	Strong ethical/legal considerations

Table 1: CYBINT pros and cons

3.2 Data sources

Most of today's data is in electronic format and it can accessed via the Internet. About 70 to 90 percent of all used intelligence material is open data. This open data is called the name Open Source Information (OSINF). Technical development in computer science and mathematics fields like AI, data mining, and Machine learning (ML) and computer calculation power have become easier and cheaper to use OSINF (Ghioni et al., 2023) . Cyber intelligence data is also fast to collect. Because the quantity of information on the Web increases exponentially it has to become more and more harder. This requires special search strategies. The verification of the information is critical. Therefore reliability of the data source is critical. (Norton, 2011) .

All electronic devices can be used as a data source in CYBINT. Data can come from the Internet like social media, blogs, Wikis, pages, etc or it can be metadata like Internet Protocol (IP) addresses, network traffic data, logs, etc. (Şahin, n.d.) (Hassan and Hijazi, 2018, p.5) . The one key aspect of CYBINT is to create links between unrelated data sources (Tabatabaei and Wells, 2017) .

For data, we have 4 different types of data: (Gibson, 2017)

1. Open source data: This is raw information and data like statistics
2. Open source information: processed information like journalistic content
3. Open source intelligence: compiled data that addresses a specific query
4. Validated open source intelligence: OSINT which has a high degree of certainty of its truthfulness.

The other way to categorize data is by information type: (Pastor-Galindo et al., 2020)

- Personal information covers the person's identity details like name, email address, and usernames for different services, social networks, curriculum vitae (CV), and education.
- Organizational information is information from the company or team instead of individuals. Typical data are company names and addresses, websites, domain names, files, images, and locations.
- Network information is technical information that covers systems and communication topologies, IP addresses, hostnames, registration information, DNS records, and operating systems.

Data can be structured, semi-structured, or unstructured. Structured data is highly organized such as data held in relational databases and there is a data model that describes each table, column, etc., and the relationships between them. Unstructured data is the opposite of structured data. There are no things like data models or clear variables. Unstructured data is typically used in everyday life like books, web pages, journals, audio, videos, images, etc. Between structured and unstructured data is semi-structured data. This is the data where there is a defined structure and data model but the data itself is unstructured. A typical example is Twitter's or Facebook's application programming interface (API) where users can read documentation that tells them how you get titles or raw text or images etc.. Still, data can be anything (Gibson, 2017, p. 73) .

For the last 10 years, different public organizations have increasingly opened their data. For the CYBINT context, this data has always been too general to be too anatomized but on the other hand, it can help understand the environment and conditions where the CYBINT target lives (Gibson, 2017, p. 77) .

One big question is the quality of the data. At web sources, the source itself sets limitations to data. For example, Twitter has a limit on the length of tweets. This defines what type of data an investigator can get. On the other hand, many web sources save long-term data and many users edit it. So there can be better information than what investigators can get through other methods like interviews. There are a few things that must be kept in mind when choosing a data source: (Munzert et al., 2015, p.7-9)

1. Think what kind of information is needed. It is a different ballgame if the population of the Organisation for Economic Co-operation and Development (OECD) countries is needed to know who owns company X in country Y.
2. Find out data sources on the Web that might provide data to solve the problem. The vague information the harder to find.
3. Check the quality of the data source and think critically. When data has been created? When it has loaded to the Web? Who has done those things? Are there black spots that data don't cover? Can we get better data from some other source or can this data be merged to other data easily?
4. Make a decision. Do we take that data or not? Is there any problem that cannot be solved? How did those problems affect the original investigation plan?

For reasons using CYBINT are wide and varied. On the other hand, there can be a situation in which "official" information is too expensive or too slow to get or the target doesn't want to give it to us. In some cases, investigators don't want the target can know that he or she is the target. The typical situation is that they both (investigator and target) have the same interest like buying the same company or getting the same agreement (Gibson, 2017) .

APIs are one of the most important ways to collect CYBINT-relevant data. Social media platforms offer APIs to collect data and other APIs to collect data. One typical example is Google's Custom Search API. That is a powerful way to get interesting results. APIs are also a great way to automate CYBINT projects. APIs are also handier to use than traditional web scraping (Ponder-Sutton, 2016) . Normally, those APIs provide data in JavaScript Object Notation (JSON),Extensible Markup Language (XML), or ATOM format (Perez and Germon, 2016) .

3.2.1 Social media

Most people have some kind of online presence and profile (Ponder-Sutton, 2016, p. 16) . Social media data like posts, likes, pictures, etc. is the treasury for CYBINT investigators. There is even own term for CYBINT in the Social media context. That term is in social media intelligence (SOCMINT). (Chauhan and Panda, 2015, p. 38-39) For CYBINT purposes, the best way to use social media services is to use their APIs. The three most important social media platforms are LinkedIn, Twitter, and Facebook (Gibson, 2017, pp. 77-80) .

Social media is very important from CYBINT's point of view because there is a lot of useful information like social relationships between peoples and/or companies (Perez and Germon, 2016) . Some social media information is open and free and doesn't require any authentication and some information does (Chauhan and Panda, 2015, p. 39) .

Investigators can use social media in two ways - using graphical search tools or using APIs. Depending on the mission and investigators' experience, it is the case-by-case solution that is handier. In some cases, investigators can also use search engines to find usual information from social media (Perez and Germon, 2016) (Chauhan and Panda, 2015, p. 44-51) . There are also dedicated social media search engines like Social Searcher or SocialMention. Those are often also analytic capabilities (Chauhan and Panda, 2015, p. 82-85) .

For this thesis context, the most relevant social media platform is LinkedIn because it offers information to companies and their employees (Chauhan and Panda, 2015, p. 41-43) . That kind of information is very useful for KYC and AML (Troia, 2020, p. 17) .

3.2.2 Metadata

Metadata (also sometimes microdata, socially structured data, or 'rich snippets') is information about the content of data in specified and well-defined format (Gibson, 2017, p. 76) . This information can be anything and it depends on the form or original file and what is relevant for that file for example a video file, the length of the video, or information from the camera are relevant metadata. If the original file is a Word document then that kind of information would be insane (Chauhan and Panda, 2015, p. 152-153) .

One simple example of metadata is tags of HyperText Markup Language (HTML) web pages. Metadata is one way to find out the value of information from a CYBINT point of view (Gibson, 2017, p. 76) . It helps to decide what information is relevant and what is not, and it helps classify, categorize, and manage files and information (Chauhan and Panda, 2015, p. 152-153) .

Typical metadata information for a file or page includes the author's name, file size, location, creation timestamp, and comments. Usually, metadata is stored in the file, but some file types have their own specific metadata file type like images metadata format Exchangeable image file format (Exif) (Hassan and Hijazi, 2018, p. 46) . There is a specific tool called name Jeffrey's Exif Viewer where you can analyze Exif-type metadata (Chauhan and Panda, 2015, p. 152-153) . For Exif pilot tool can be change, view, edit, and remove EXIF-information from pictures (Hassan and Hijazi, 2018, p. 46) . There is also a search engine where can be found Exif-information (Chauhan and Panda, 2015, p. 154-155) .

The most important metadata tool is Fingerprinting Organizations with Collected Archives (FOCA). FOCA means seal in Spanish. It is a general-use metadata tool that can analyze metadata also from non-image files like Word or PDF files (Chauhan and Panda, 2015, p. 158-159) .

Typical internet or social media users don't care about metadata so it is often very useful for CYBINT purposes. In many cases, the "real" information is less valuable than the metadata. The way to prevent leaking unwanted metadata is called name data leakage protection (DLP). The most effective way to prevent metadata leaks is to remove all metadata. One way to do it is the use metadata anonymization toolkit (MAT) with is Python library for metadata anonymization. (Chauhan and Panda, 2015, p. 161-162) .

3.2.3 Dark web

The web where everybody can go by using search engines like Google or Yahoo can be named Clearweb (Chauhan and Panda, 2015, p. 188) or Surface web (Kalpakis et al., 2017, p. 112) . On this, The Deep web (also known as The Hidden web or Invisible web) is a general term for information that cannot be found using general-purpose search engines. For example, the bank's website with a login page is easily found by using search engines but users cannot find detailed information from one customer’s bank account, so that login page is on the public internet but every customer’s web bank page is in the deep web (Kalpakis et al., 2017) . The deep web content can be used by normal internet browsers and Hypertext Transfer Protocol (HTTP)/Hypertext Transfer Protocol Secure (HTTPS) protocols (Hassan and Hijazi, 2018, p. 95) .

The dark web is the part of the Deep web where is impossible to go by using standard Web browsers like Firefox or Explorer. To get The Dark web requires specific browsers, software, or configurations. The main principle for the dark web is anonymity and privacy for the user and data perspective (Kalpakis et al., 2017) .

The three most popular Dark web networks are The Onion Router (Tor), Invisible Internet Project (I2P), and Freenet (Kalpakis et al., 2017) (Hassan and Hijazi, 2018, p. 102) . The most popular of those three is Tor (Hassan and Hijazi, 2018, p. 104) . All of those require different applications/browsers to access them (Chauhan and Panda, 2015, Chapter 9) .

Many different people use the dark web. On the other hand, there is lots of criminal activity but also whistleblowers, researchers, journalists, and human rights activists (Kalpakis et al., 2017, p. 113) (Chauhan and Panda, 2015, p. 189-190) . Like the "normal" web, there are some functionalities in the Dark web. There is a newsgroup, online stores, blogs, search engines, forums, etc (Hassan and Hijazi, 2018, p. 102) (Chauhan and Panda, 2015, p. 189-190) (Kalpakis et al., 2017, p.127-128) .

Tor is the most widely used dark web. It uses an onion router and the web address ends ".onion". It is written using C language. It was originally the US Navy project. Nowadays, it is under a non-profit organization. The idea of an onion router is the use multi-layer encryption (at least three) and every node removes only the upper layer. The routes between nodes choose randomly so it is quite slow but it is good for what it does (Chauhan and Panda, 2015, p. 182-184) (Kalpakis et al., 2017, p.114-116) (Hassan and Hijazi, 2018, Chapter 3) . There is no secret backdoor in Tor, but several different attacks against Tor have been done - both theoretically and practically (Bidwell, 2022) .

From the safety point of view the biggest risk is is end note because it always knows the destination. National Security Agency (NSA) runs its tor nodes and this is a security risk, especially if NSA’s node is the user’s end node. (OccupyTheWeb, 2018, p. 142-143)

Some countries have access to the Tor network monitored or blocked. In those cases, a user needs to use a feature called "bridge" to access the Tor network. The bridge is a specific server where the user takes a connection and then uses it to go to the Tor network. (Drake, 2024) .

The I2P is designed for independent darknet. It supports both server and client applications. It also offers an anonymous way to use Surface Web (Kalpakis et al., 2017, p. 116-117) . It is released in 2003. Traffic is encrypted using four layers before the destination. It is written using Java language.The address ends ".i2p" (Hassan and Hijazi, 2018, p. 117) . I2P uses garlic routing. It is an expansion to onion routing. The idea of garlic routing is that it encrypts messages to a single packet. This makes attacker harder to do network analysis (Kalpakis et al., 2017, p.117) .

There are a few differences between I2P and Tor. Firstly, they are designed for different thread models. The main idea behind Tor was to allow users to use the surface net and do anonymously. I2P was created primarily as an anonymity network that allows fully anonymous communication between two parties. Tor has a much better foundation than I2P. Because Tor has written using C and I2P uses Java, Tor runs much faster and uses less RAM than I2P (Hassan and Hijazi, 2018, p. 122-123) .

Unlike Tor and I2P, Freenet is only dark web. There is no possibility to use the Surface web via Freenet and stay anonymous. The other big difference between Freenet and Tor/I2P is that the files with are uploaded to Freenet stay on Freenet even the loader goes offline. This is because the files are stored in small chunks at different nodes. Due to the anonymity of Freenet, the original publisher remains unknown (Kalpakis et al., 2017, p. 177-118) .

3.2.4 Other data sources

For companies value-added tax (VAT) information in all European Union countries finds from VAT Search (https://vat-search.eu). (Hassan and Hijazi, 2018, p. 274)

Réseaux IP Européens Network Coordination Centre (RIPE NCC) is a non-profit organization that coordinates regional services that support the Internet's infrastructure via technical coordination. From CYBINT's perspective, their most important service is the () and services related to Internet Protocol version 4 (IPv4), Internet Protocol version 6 (IPv6), and autonomous system number (ASN). (RIPE NCC - RIPE Network Coordination Centre, n.d.-a)

The autonomous system (AS) is a collection of IP networks with run network operators (one or more) with a single defined routing policy. ASN is an identifier number with an identified specific routing policy. (RIPE NCC - RIPE Network Coordination Centre, 2019)

The Ripe Database is a database with contains registration information for networks with RIPE NCC service regions. It contains information for Internet number resources and routing policies. (RIPE NCC - RIPE Network Coordination Centre, n.d.-b)

3.3 Tools

Many business intelligence (BI) tools, like Statistical Analysis System (SAS) (https://www.sas.com/en_in/home.html), Pandas (https://pandas.pydata.org/), or R (https://www.r-project.org/), can also use as CYBINT purposes. The most important programming language is Python which has written many important CYBINT tools or packages (Chauhan and Panda, 2015, Chapter 13) . On the other hand, there are also specified CYBINT tools like Maltego, the Harvester, or Spiderfoot. Many everyday tools like search engines are often valuable also CYBINT context (Ponder-Sutton, 2016) (Kadar, n.d.) .

In many cases, geospatial information is needed. Typical places to collect that information are different map services like Google Maps, Wikimapia, or Bing Maps (Pastor-Galindo et al., 2020) .

Because the internet has continuously changed, web pages are created, deleted, and edited, it is important to find out the structure and content of the specific page at a specific moment, there is a service Wayback Machine with archives copies of the web pages regularly. (Pastor-Galindo et al., 2020)

3.3.1 Operating Systems

From a technical point of view, it is possible to use any modern operating system for CYBINT purposes. No matter, what is the operating system - that can be Windows, Linux as well as MacOS. In reality, there is a big difference between different operating systems. The best operating system for CYBINT purposes is Kali Linux because that system is ready-installed with several OSINT tools (The Kali Team, 2022) . If the investigator stays extremely hostile environment like a war zone, then the best solution is to use Linux distributions with maximized privacy like The Amnesic Incognito Live System (Tails) (Drake, 2023b) operating system (OS) or Kodachi. Those are designed to give users as good privacy as possible (Sharma, 2017) (Ponder-Sutton, 2016) (Hassan and Hijazi, 2018, p. 76, 93)

. Tails was developed by the Tails project and it started in 2009. The biggest with the Tails is that it does not protect against hardware keyloggers and other sophisticated attacks. Therefore, Tails live distributions should be run only by trusted computers. (Drake, 2024)

Tails and Kodachi routes all network traffic through the Tor network (Sharma, 2017) and there are other privacy tools like secure instant messaging (IM) chat and encrypted email client. Those are also portable OS so it is possible to use it directly from a Digital Video Disc or Digital Versatile Disc (DVD) (Drake, 2023b) or a Universal Serial Bus (USB) stick. (Hassan and Hijazi, 2018, p. 109) .

For signal intelligence purposes, there is also a special Linux distribution for signal intelligence. Its name is SigintOS. (SigintOS, n.d.)

The effective way to maximize security is to use the Qubes OS operating system. Qubes OS is based on Fedora (Drake, 2023a) and it is signed to a security-oriented Linux system (Qubes OS, n.d.) . Qubes OS, as the name says, uses qubes to isolate different functions for each other. Every qube is virtual machine (VM) with runs specific application. Therefore, if a hacker attacks one application it cannot break into another (Mansfield-Devine, 2010) . Those virtual machines are lightweight Xen virtual machines. There is one qube for admin purposes. This qube is named dom0. That dom0 qube can master other cubes and monitor Random Access Memory (RAM) and Central Processing Unit (CPU) usage. (Drake, 2023a) .

When a user installs Qubes OS, there is a virtual machine with contains a Linux system called the name Whonix. Whonix is a Linux distribution with is forced to internet anonymity and security. (Whoinx, n.d.) It is also possible to install Kali Linux on the Qubes OS. In that case, Kali Linux is one virtual machine on Qubes OS. (Joshua, 2020)

In this thesis, I focus on Kali Linux and its capabilities for CYBINT/OSINT.

3.3.2 Search engines

If a user wants to find something from the Internet he or she probably uses the search engines like Google or Yahoo. Those are also very handy tools for CYBINT investigators. Every search engine has many features with are useful for CYBINT perspective (Hassan and Hijazi, 2018, p. 127-128) . Different search engines use different algorithms so the same method used in two different search engines gives two different results. For this reason, it is important to use many different search engines to find valuable information (Chauhan and Panda, 2015, p. 72-73) .

There are semantic "general" search engines and also search engines for special purposes like finding metadata, people, companies, or technical information like IPs, user names, etc (Chauhan and Panda, 2015, Chapter 4) .

The first critical thing is to choose the right keyword. This can be very tricky especially if the field of the target(s) is not the same as the CYBINT investigator's comfort zone. To solve this problem has been published keyword list by themes (Crowder, 2016, Introduction) . The other service is where CYBINT investigators can find interesting search terms in search engines' own statistics for the most trending terms for relevant geographical areas. One example of those services is Google Trends (Hassan and Hijazi, 2018, p. 129) .

The second search technique is the advanced search operators. This means specific keywords that searchers can add to our search so that you can get more specific results. Using those operators you can search for specific web pages, the searcher can limit the format of the results (like PDF files) or the user can require that a keyword must be in the title. Also, Boolean algebra is one class of advanced search operators. Most advanced search operators have general so those can be using all different search engines, but there are also search engine-specific operators (Chauhan and Panda, 2015, p. 97-106) .

3.3.3 Maltego

Maltego is the tool created by Pateva. It comes as a part of Kali Linux (lalitmohantiwari7700, 2021) . Maltego is an information-gathering and visualization tool. It is almost "standard" for CYBINT projects. The general idea of Maltego is that it visualizes connections and links between different data sets. It uses icons and symbols for different kinds of information and relationships (Chauhan and Panda, 2015, p. 142-144, 208–214) .

There are two different versions of Maltego commercial and community. The commercial version needs a license key. The community version needs only registration. The community version has some limitations like a limited amount of data extraction, lack of user support, etc. At Maltego there is graphical user interface (GUI) which makes it easier to use. Maltego uses client–server architecture. Maltego has written using Java (Chauhan and Panda, 2015, p. 142-144) .

Next, lets define a few important Maltego terms: (Chauhan and Panda, 2015, p. 143-146)

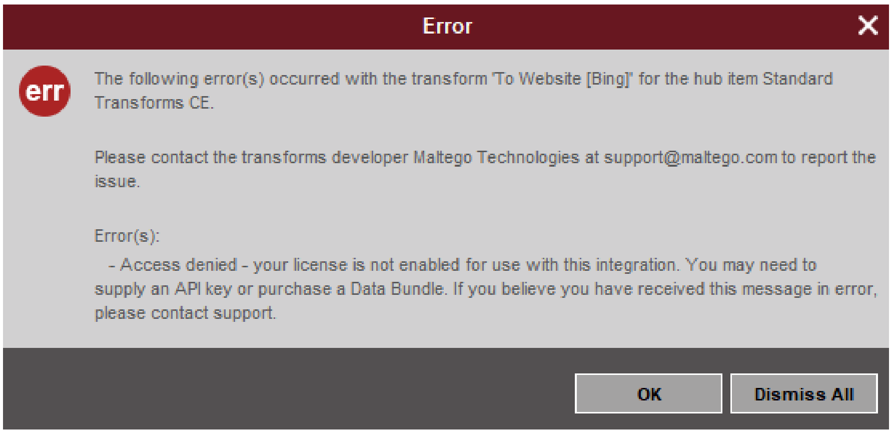
- **Entity** is the term for a piece of input data with has loaded into the Maltego. There can be one or groups of entities in Maltego. They are shown as icons in Maltego.
- **Transform** is the piece of code with takes an entity as an input and then extracts a new form of entity based on relationships.
- **Machine** is a set of transforms that are linked together.
- **Investigate** This is an option with offers basic functionalities like cut, copy, paste, link/entity, etc.
- **Mange** is the option with offers entity and transform management.
- **Organize** is a way to handle graphs, like layout, type, etc.
- **Collaboration** is the way to working together with other users

IPQualityScore (IPQS) is the tool to investigate and verify IP (Maltego Team, 2020a) and email (Maltego Team, 2020b) addresses for fraud. The tool uses over 25 different data points to analyze scores in real time. (IPQualityScore, n.d.) The tool gives scores from zero to one hundred. If the score is greater or equal to 75 then it is considered suspicious, but not necessarily linked to fraudulent activity. Often it is a proxy, VPN, or Tor connection. If the score is greater or equal to 85 then there is a high risk for fraud (Maltego Team, 2020a) .

The tool also gives tags with tell more specifically the type of fraud. Those tell if the IP is a proxy, VPN, or Bot. It also gives frequent abusive behavior for the previous 24-48 hours. Possible levels are "none", "low", "medium", and "high". It can also give IP's geographical location, ASN and Domain Name System (DNS). That information can be relevant when the investigator does fraud analysis. (Maltego Team, 2020a) .

In the case of email addresses, the tool is not suitable for a large number of addresses because Simple Mail Transfer Protocol (SMTP) servers block requests. In many cases, this also prevents Maltego from telling if the email address exists or not. (Maltego Team, 2020b)

In this thesis, one major limitation for Maltego is licensing. In this thesis has used only the free version, because huge cost of the license. This leads to limitations, especially in cases where is used



Error

The following error(s) occurred with the transform 'To Website [Bing]' for the hub item Standard Transforms CE.

Please contact the transforms developer Maltego Technologies at support@maltego.com to report the issue.

Error(s):
 - Access denied - your license is not enabled for use with this integration. You may need to supply an API key or purchase a Data Bundle. If you believe you have received this message in error, please contact support.

Figure 3: Typical license error in Maltego.

personal information instead of company or technical data. In the figure 3 is the typical error for Maltego license limitations. It comes when the investigator uses the entity type "Person".

3.3.4 SpiderFoot

SpiderFoot is a CYBINT platform with the principle idea of making information chains so when investigators find one piece of information it automatically leads to the next part of information until all information has been collected. SpiderFoot supports "shotgun" approaching, so there can be collect much information and then later decide what information is relevant and what not. There is also a premium version of SpiderFoot, called name SpiderFoot HX. HX is cloud-based and contains properties that do data gathering faster like multi-threat scanning, fast scanning, better visualization, etc (Troia, 2020, p.84-94) .

3.3.5 The Harvester

The Harvester is an open-source tool where investigators can find email addresses and employee names from public pages such as LinkedIn or public search engines like Google. The tool was created by Christian Martorella and it is written in Python. It is a kernel-based tool, and it saves results in HTML or XML format (Chauhan and Panda, 2015, p. 100-102) .

The Harvester usually takes as input the company name or domain name but it is also possible to use keywords. The command for this tool is the harvester and at Table 2 there is possible parameters: (Chauhan and Panda, 2015, p. 100-102)

parameter	description
-d	company name or domain name
-l	max number of results
-b	specifying data source
-s	start with a defined result number (the default value is 0)
-v	get virtual hosts
-f	save results
-n	perform DNS resolve the query for all discovered ranges
-c	perform DNS bruteforce for all domain names
-t	perform DNS TLD expansion discovery
-e	use specific DNS server
-h	use the Shoban database to discover hosts

Table 2: Parameters for The Harvester tool

3.3.6 Ashok

Ashok is a free and open-source tool for information gathering. The key idea of Ashok is to scan web pages and web applications for information catering purposes. It is one of the easiest and most used tools for reconnaissance. Ashok is available on Linux and it is written using Python. The tools are written by Ankitdobhal. The main user group for the tool is penetration testers. There is command line interference on Ashok. Ashok has the very same logic as Metaspoilt (mohdshariq, 2021) (Zusyaku, 2021) .

Using Ashok is simple. The user gives it the target domain (web address). It contains several features like http-headers extractor, dns-lookup, whois-lookup, nslookup, subdomain-finder, nmap scanning, GitHub,githubrecon, cms-detector, link extractor, banner-grabbing, subnet-lookup, GeoIP-lookup. There are also specific features for WordPress, Drupal, Joomla, and Magento CMS. (mohdshariq, 2021)

3.3.7 Recon-ng

Recon-ng is an open-source tool for reconnaissance and information gathering. It is written in Python and it required Python 2.7.x to run. The main developer is Tim Tomes. The tools focus on web-based reconnaissance with independent modules. The framework for Recon-ng is quite the same as Metasploit so if the user has experience with Metasploit user has a very easy-to-use Recon-ng and vice versa. Recon-ng is a tool with can be used in console (Chauhan and Panda, 2015, p. 131-138) (Kings, n.d.) .

Available commands for Recon-ng are listed in the table 3. (Chauhan and Panda, 2015, p. 133-134)

Recon-ng tool has five different type of modules: (Chauhan and Panda, 2015, p. 135-136)

command	description
add	Add records to database
back	Exit current context
del	Delete record from database
exit	Exists the framework
help	Displays the menu
keys	Management the API keys
load	Load module
pdb	Starts Python debugging session
query	Queries the database
record	Record command to a resource file
reload	Reload all modules
resource	Executes commands from resource file
Search	Search modules
set	Sets module options
shell	Execute shell commands
show	Show items from framework
spool	Spools to output file
unset	unset modules options
use	load specific module
workspaces	manages workspaces

Table 3: Commands for the Recon-ng tool

1. Discovery
2. Exploitation
3. Import
4. Recon
5. Reporting

3.4 Operations security

One of the most important parts of OSINT operations is OPSEC. Be safe for targets' surveillance or the third parties tracking (Hassan and Hijazi, 2018, p. xxi) . If CYBINT investigators don't secure his/her background, they can tell many sensitive things about their background or their own company's or employer's motives. Depending on the target, exposing that information can be very dangerous (Hassan and Hijazi, 2018, p. 21) .

The first step in Operations security checks is that users must have installed security software that they are using in everyday life like anti-virus and firewall software (Hassan and Hijazi, 2018, p. 32-33) . The second step is the protected computer. It is highly recommended that the investigator lock his/her PC using a USB drive, set a password for Basic Input-Output System (BIOS)/Unified Extensible

Firmware Interface (UEFI), and disable unnecessary ports/protocols and services (Hassan and Hijazi, 2018, p. 35-38) .

The second step for OPSEC is to use virtual private network (VPN). The reason for this is that it gives the user to anonymous IP address so it hides the user's real location. It's really important to choose a VPN service provider with offers DNS leak protection (Hassan and Hijazi, 2018, p. 65-66) (Layton, 2016) .

After VPN, the next thing is to secure online browsing the investigator leaves only a minimum digital fingerprint (Hassan and Hijazi, 2018, p.59-64) . Another important way to hide own location is to use proxies (Munzert et al., 2015, p. 123-124) (Layton, 2016) . To maximize their anonymity, investigators can use multiple proxies and make proxy chain (Mukonyi, 2021) . The best way to confirm the trusted proxy servers is to use payment proxies because free proxies are always risky they sell IP addresses. (OccupyTheWeb, 2018, p. 148) .

At Kali Linux, there is also a tool called the name "Who Am I" which can change the investigator's own computer's DNS and media access control address (MAC address) and other computer's identification information. (aftab1x0, 2021)

When the investigator has secured the physical computer and network traffic, next must secure their internet browser. The easiest way is to use browsers, which are developed from a maximum security point of view like the Tor browser or Epic browser (Hassan and Hijazi, 2018, p.59-63) (Layton, 2016) . For mainstream browsers, the best is Firefox because it is open source and there are lots of privacy settings (Hassan and Hijazi, 2018, p.59-63) . The first step for private browsing is to use the "private browsing"-window. At the settings, it is important to put tracking protection on and put the "never remember history"-settings.

The next logical step is to think about the security of using web services. Many services want users' personal information and/or payments. Therefore investigator needs anonymous webmail addresses. Services like Protonmail or onetime email services handling this (Hassan and Hijazi, 2018, p. 84-85) . For many CYBINT-relevant services users have to pay that they use those services. This is a situation where the investigator needs a way to pay an anonymous payment method. There are two main ways - there can be use cryptocurrencies or prepaid gift cards (Hassan and Hijazi, 2018, p. 79-80) .

The last important part of OPSEC is to fake online identity. Many times investigator needs it to register different services like social media (Hassan and Hijazi, 2018, p. 92-93)   (Troia, 2020, Chapter 17) .

4 CYBINT and financial sector

4.1 Legal aspects

The main reason to hole work comes from regulations. In 2007-2009 the world had a financial crisis. One of the main reasons for that was the lack of regulation and supervision of banks. Their politicians create more regulations and that continues. (Slovik, 2012)

In Finland, there is the act for preventing money laundering and terrorism financing (Valtioneuvosto, 2017) and a Decree of the Government procedures for getting to know the customer and risk factors in preventing money laundering and terrorist financing. (Valtioneuvosto, 2021)

For those regulations, there comes a compulsion to collect and analyze data for anti-money laundering purposes. This means that companies must do risk assessments and risk classifications. (European Banking Authority, 2021)

4.2 Using CYBINT at financial sector

Traditionally, CYBINT has been used to hunt cyber criminals in the financial sector. (Troia, 2020, Chapter 2) I see that one way to use CYBINT in the financial sector is the risk-based approach. From this point of view, we use CYBINT tools and methods to do better risk estimations and to classify threads. Those threads are customer-based. Nowadays, on the Internet and different online services, we can get lots of useful information for potential customers we can confirm information what they tell us and/or collect new information. Both cases are very useful when we do customer risk estimation.

Financial analysis has used public information like company registers or news. For public data, there has been indicator recognition or company analysis. (Europol, 2002)

Banks and other financial institutions have typically a lot of sensitive data in their systems, so the best results from an anti-money laundering point of view are if investigators can do combinations and use both CYBINT data and the company's data. (Day et al., 2017)

4.3 CYBINT and AML

Cyber intelligence on the new and less-used way to battle against money laundering. In many cases Cyber intelligence is to way to collect information with other methods can use. For example, if cyber intelligence finds something interesting documents, NLP can use those and find something relevant. In the best cases, investigators can even build thread models based on cyber intelligence. (Neotas, 2017) (Watters, 2013)

Because financial sector institutions, like banks, have lots of their sensitive data, the investigator gets the best results if there is the possibility to combine sensitive non-osint and public osint data. (Day et al., 2017)

In the financial sector, the concept of using CYBINT to prevent white-collar crimes is to use a framework where there are four stages of investigation:

1. Investigator to technology systems
2. Investigator to investigator systems
3. Investigator to information systems
4. Investigator to application systems

So, the main focus on to use both interdependent and differentiated factors with focusing on system dynamics. (Tabatabaei and Wells, 2017)

4.4 CYBINT and KYC

CYBINT from a KYC perspective can provide a streamlined and scalable way to screen customers. It can be an automated process that produces watchlists or gives proof of unwanted behavior like illegal activity or brand risks. It can also give risk factors that the company knows what kind of information must as from the customer. (Babel Street, n.d., p. 21)

The 4^{th} EU Money Laundering Directive requires that banks must use risk classification profiles in their KYC processes. (ACTICO, n.d.) In many cases, CYBINT investigation produces too much information and therefore, after the CYBINT process, nowadays many organization uses machine learning or artificial intelligence solutions to find relevant information from a KYC point-of-view (Chau and van Dijck Nemcsik, 2021, p. 189-191; Kuchmai and Shelest, n.d.).

5 Case example

5.1 Goal

My goal for this section is to show how to do the CYBINT project practically. I try to get helpful information for AML/KYC purposes. I called customer A because of banking privacy. Customer A is a private person.

In the second part, I do the same thing for the company with the alias "Company B".

5.2 Operations security

The first step for every CYBINT project is operational security. In this case, I use Kali Linux which can start from a USB stick. Kali has been installed on a USB stick in public libraries because there is a possibility to use the internet anonymously.

Before the project has started, there must be a dummy email address because Maltego's free version, Maltego CE, requires registration. (Maltego, n.d.) . In this case, the investigator uses Proton mail with dummy information.

This project has used ProtonVPN as a VPN service in Kali Linux. (Kali Linux, n.d.) . Another option was to use proxies but it is unnecessary to use both at the same time. (Nakutavičiūtė, n.d.) (Kurniadi, n.d.)

After that, there has also been used a "Who am I" tool for hiding computer information for counter-investigators. (aftab1x0, 2021)

5.3 Data sources

Because investigators want to find data with are not easy to find and because financial institutions can use most of the authority registers in Finland, the main data sources for this case are social media and the deep web.

Social media is engaging because there is lots of information and one part of that is the target's own posted information. Deep web information is also important because traditionally those kinds of sources are not used for KYC/AML purposes.

5.4 Tools, Methods, and Results (Person)

5.4.1 Maltego

The first tool to use in this project is Maltego. The first step is to start the new project. After that investigator adds Customer A's name for the personal item. After that, the investigator gets information shown in figure 4.

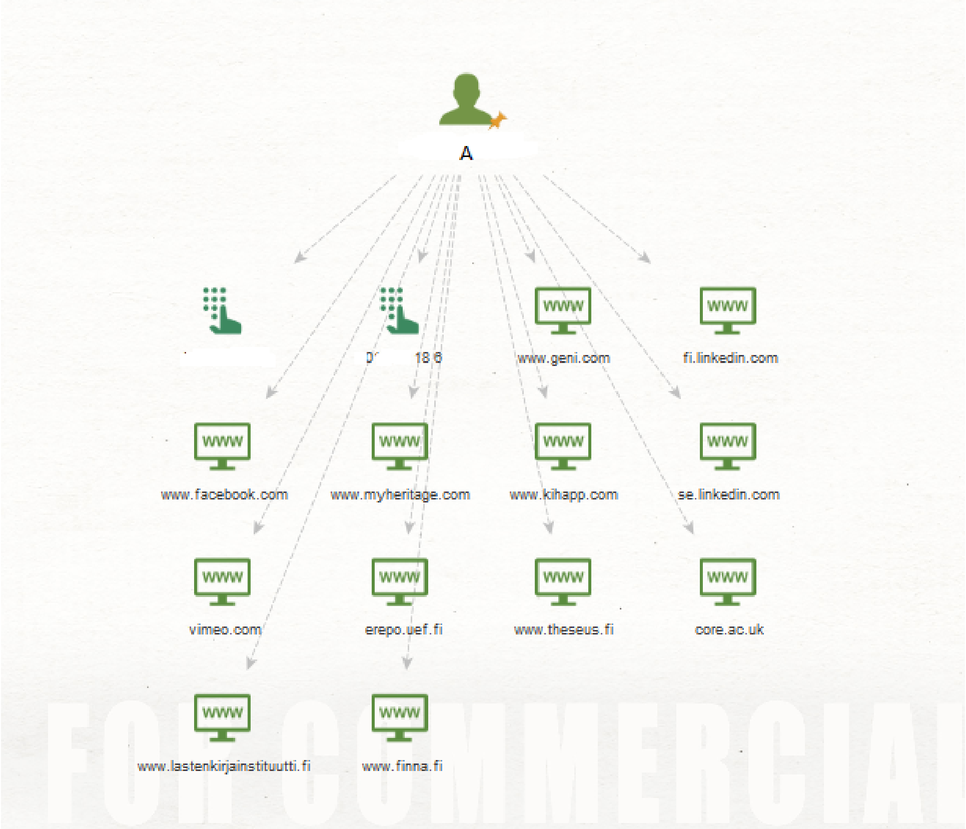


Figure 4: Fist step results for Maltego

With this information, the investigator noticed a few interesting aspects. First is that Customer A's personal information has geni.com and myherritage.com services so it is easy to find out customer A's relatives. Their investigator can easily check that is it there any section listed persons.

Another notable piece of information from A is that he/she has a university degree from the University of Eastern Finland. This can be noticed because there is a connection to erepo.uef.fi where the thesis of that university is. Also, finna.fi and thesus.fi services confirm that information.

One very usual site to get information is LinkedIn. In figure 4, there is website knot "fi.linkedin.com" and "se.linkedin.com". So, their investigator gets the target's occupational information. This is very important information from a KYC/AML perspective.

When I continue to dig deeper, I get information shown in figure 5. There is a lot of information about Customer A's activity. There have been lots of tips for customers' occupations. There are different kinds of articles with A has written. Another tip for occupation is the International Movie Database (IMDP) which tells that A has worked Finnish movie/TV sector.

Also, there is information from IP addresses with tells that A has been in a long time in Eastern Finland. For hobbies, I found out that A has done martial arts sport. This is also information with can use to find more connections to other persons.

One notable thing that I did not find. It is anything outside Finland and anything about Cryptocurrency. Therefore, it is very likely that A has no notable connections abroad. Also, there is a high probability that A does not own cryptocurrencies.

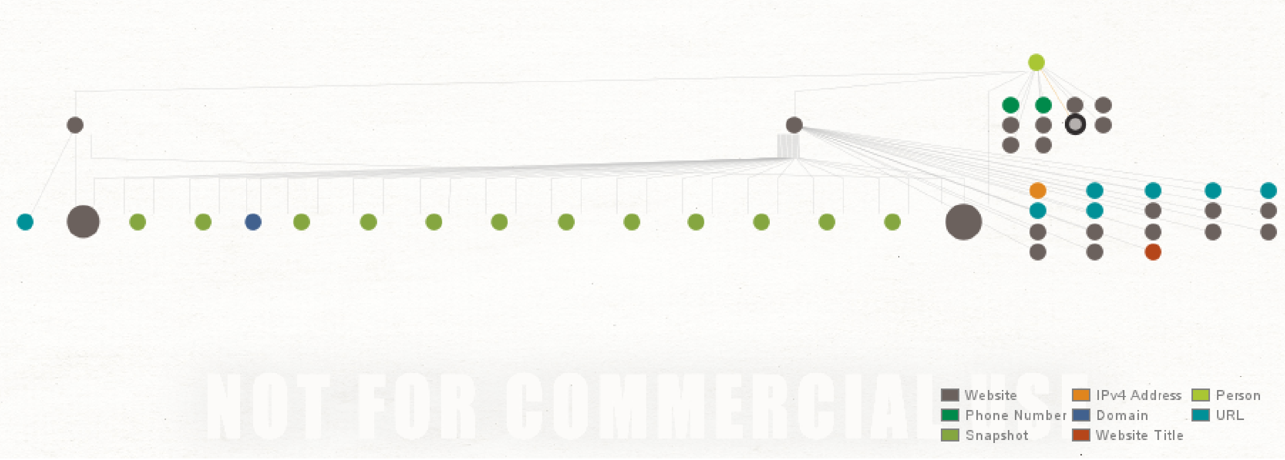


Figure 5: Final results for Maltego

5.5 Tools, Methods, and Results (Company)

For a company, the typical situation is that the investigator knows the company's name and web page (or web page is easy to find).

5.5.1 Maltego

At first, the investigator uses the company's name to start the investigation. The first investigator tries the entity type "company". Soon, the investigator finds that it won't work because the target company, called the name Company B, is a Finnish company. Then investigator changes the tactic and uses the start entity as a "phase" type entity. After that, there are lots of interesting data shown in figure 6.

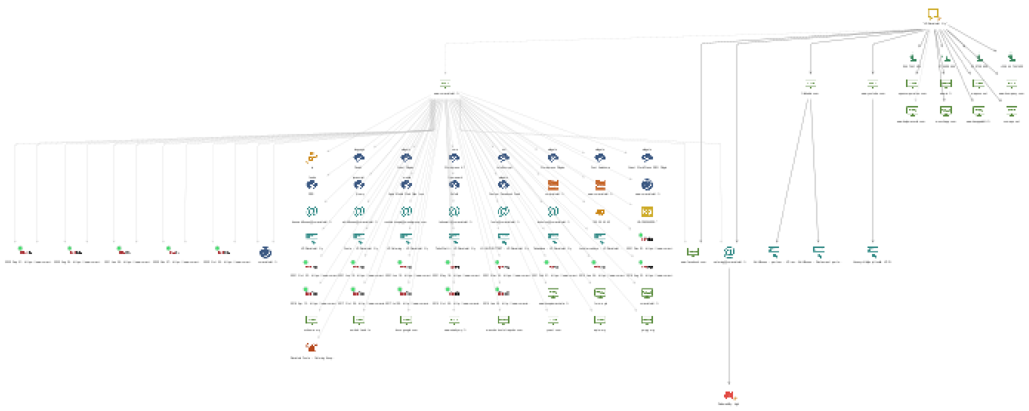


Figure 6: Final results for Maltego

The first thing that the investigator finds is a web page where can get phone numbers and addresses for different offices of the company. Then there are several email addresses and phone numbers for companies. Also, there find the owner’s names and emails.

For the internet, the Web Archive finds old versions of the web page. They may find out that the company has different offices earlier but they have quite those. This can tell the situation of the company.

There is also information for YouTube where owners are entrepreneurship society’s video. There is also a link to a page where there is an article on how COVID-19 affects the company.

The last thing with founds there is to information from different company databases where is possible to get information for companies financial situation. Those are normally paid databases so in this case I don’t dig those more.

5.5.2 SpiderFoot

The Spiderfoot gives information related to ASN numbers. For investigating this data, the RIPE Database is essential.

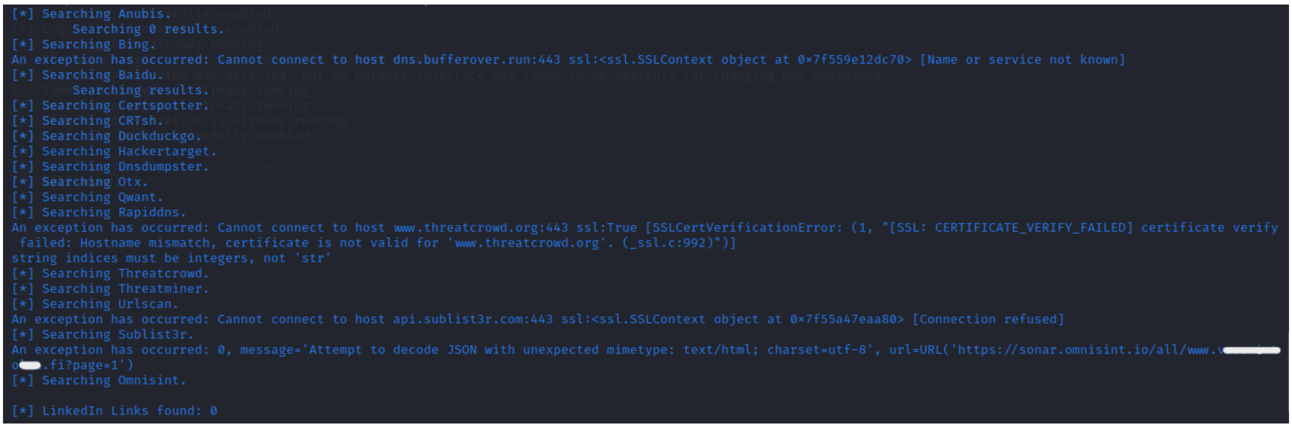
Spiderfoot gives lots of interesting data when input was email addresses and web pages that were found to using Maltego. The next table has the most important findings.

Type	Name	Description	Why important
Text	name	Code with is the name in AS system.	Tells who is email’s internet service . provider. This can be used to check if there is connections for section countries
IP Address	cidir, IP addr	IP Address for cloud service provider	Tells name (and country) for cloud service provider
Text	country code	ISO 3166-1 Alpha-2 code	Country code of cloud service provider tells connections of the spesific country
Text	deception	Deception of the incident	This tell reliability of the page when has done and what
Timestamp	first seen, last seen	Dates and times for different incidents	This tell reliability of the page when has done and what

Table 4: Findings with SpiderFoot

5.5.3 The Harvester and Ashok

The findings of using The Harvester were disappointing. It found only three email addresses with were known earlier. In the figure 7 is The Harvester tool for running.



```
[*] Searching Anubis.
        Searching 0 results.
[*] Searching Bing.
An exception has occurred: Cannot connect to host dns.bufferover.run:443 ssl:<ssl.SSLContext object at 0×7f559e12dc70> [Name or service not known]
[*] Searching Baidu.
        Searching results.
[*] Searching Certspotter.
[*] Searching CRTsh.
[*] Searching Duckduckgo.
[*] Searching Hackertarget.
[*] Searching Dnsdumpster.
[*] Searching Otx.
[*] Searching Qwant.
[*] Searching Rapiddns.
An exception has occurred: Cannot connect to host www.threatcrowd.org:443 ssl:True [SSLCertVerificationError: (1, "[SSL: CERTIFICATE_VERIFY_FAILED] certificate verify
 failed: Hostname mismatch, certificate is not valid for 'www.threatcrowd.org'. (_ssl.c:992)")]
string indices must be integers, not 'str'
[*] Searching Threatcrowd.
[*] Searching Threatminer.
[*] Searching Urlscan.
An exception has occurred: Cannot connect to host api.sublist3r.com:443 ssl:<ssl.SSLContext object at 0×7f55a47eaa80> [Connection refused]
[*] Searching Sublist3r.
An exception has occurred: 0, message='Attempt to decode JSON with unexpected mimetype: text/html; charset=utf-8', url=URL('https://sonar.omnisint.io/all/www.[redacted]
o[redacted].fi?page=1')
[*] Searching Omnisint.

[*] LinkedIn Links found: 0
```

Figure 7: The Harvester tool on running

Ashok was also disappointing. It also found the same kind of basic information that The Harvester and Spiderfoot found.

5.6 Tools, Methods, and Results (Nonprofit organization)

There is over 100 000 Nonprofit organization (NPO) in Finland. They are a very important part of Finnish society, but they have big difficulties to get financial services. (Oikeusministeriö, 2024) In many cases, those problems are KYC and AML related. (Alma Talent, 2024)

5.6.1 Maltego

At first, the investigator uses the NPO's web address to start the investigation. Results are in the figure 8. The results give lots of personal information. Those tell board members, most of the staff, and their contact information. Also, it gives much technical information like IPQS and servers. Using this information it is possible to know where city individual board members live.

5.6.2 SpiderFoot

SpiderFoot has lots of findings but those are only technical, with tells lack of maintaining web pages not relevant from an AML/KYC perspective. The only relevant personal information was links to a list of old sports records.

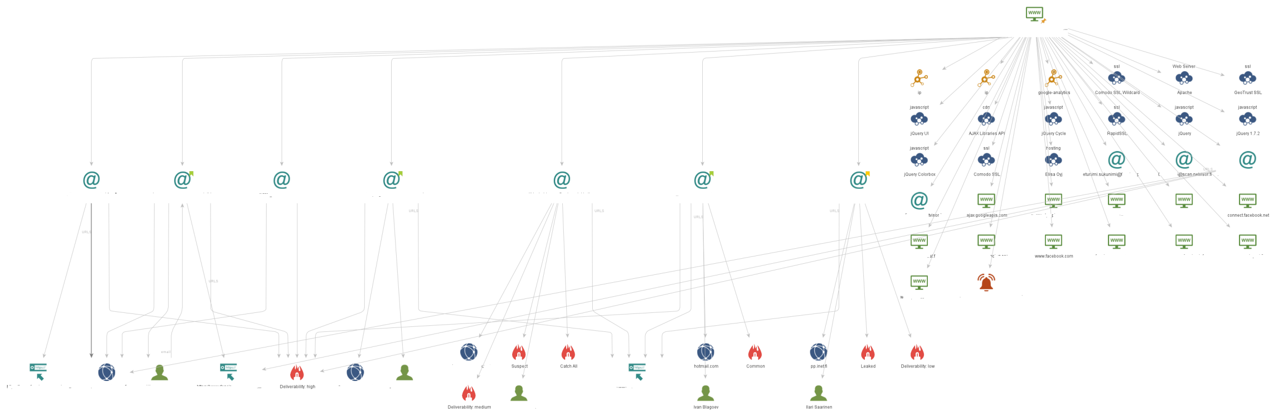


Figure 8: The Maltego for non-profit organization

6 Summary

There are plenty of different CYBINT tools on Kali Linux. Some of them like SpiderFoot and The Harvester give quite specific information for Target. On the other hand, there are also "general" CYBINT tools like Maltego which give almost all relevant information with great visualization.

One major thing in the future would be using Maltego's licensed version. It gives much more information and addins to use. From a cyber security perspective, one powerful feature is verifying and identifying email addresses. It also detects possible email and VPN-related frauds.

From the AML/KYC perspective, CYBINT methods and tools could give very valuable information, especially to company customers. I think that more than technical limitations, there are cultural limitations to using CYBINT in the financial sector.

I think that this work is ethical because there are no situations where I tell the targets of the CYBINT analysis. For reliability, in this thesis, there are only two cases one for a person and one for a company. So the results are not generalized. The most of used sources say that every CYBINT/OSINT case is unique so it is unclear how generalized results are possible to get.

The next research would test those things in bigger cases and do more cases. Also, there are possibilities to use new data sources like geospatial data or automated social media mining. There are possibilities to apply CYBINT to new fields like journalism or competitor analysis.

References

ACTICO. (n.d.). *4th eu money laundering directive – a practical guide*. https://www.actico.com/blog-en/4th-eu-money-laundering-directive-a-practical-guide/

aftab1x0. (2021). *Kali-whoami – stay anonymous on kali linux*. https://www.geeksforgeeks.org/kali-whoami-stay-anonymous-on-kali-linux/

Ahlander, J., & Dickson, D. (2015). Update 1-nordea, handelsbanken fined over money laundering breaches [Accessed: 8.12.2021].

Akhga, B. (2017). Osint as an integral part of the national security apparatus. In B. Akhgar, P. S. Bayerl, & F. Sampson (Eds.), *Open source intelligence investigation from strategy to implementation* (pp. 3–10). Springer.

Alma Talent. (2024). *Asiakkaat kohtaavat vaikeuksia pankkipalveluissa – lainsäädännön muuttaminen voisi auttaa*. https://www.almatalent.fi/juridiikan-ja-talouden-uutiset/asiakkaat-kohtaavat-vaikeuksia-pankkipalveluissa-lainsaadannon-muuttaminen-voisi-auttaa/

Andersén, A. (2020). *Rahanpesun estäminen*. Alma Talent.

Babel Street. (n.d.). Open source intelligence (osint) use cases: How publicly available information (pai) can help manage hidden commercial risks [Accessed: 6.2.2022].

Bidwell, J. (2022). Secure your vpn: Tor and other vpn alternatives. *Linux Format*, (3), 38–39.

Binder, E. (2021). *Anti-money-laundering package 2021 - strengthening the framework*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/662624/EPRS_BRI(2021)662624_EN.pdf

Bonfanti, M. E. (2018). Cyber intelligence: In pursuit of a better understanding for an emerging practice. *Cyber, Intelligence, and Security*, *2*(1), 105–121.

Chau, D., & van Dijck Nemcsik, M. (2021). *Anti-money laundering transaction monitoring system implementation*. Wiley.

Chauhan, S., & Panda, N. K. (2015). *Hacking web intelligence: Open source intelligence and web reconnaissance concepts and techniques*. Elsevier.

Christensen, C. (2019). European kyc regulations and their impact on the compliance function.

Crowder, N. (2016). *Open source intelligence glossary: Guide to keywords, phrases for improved internet research results*. Crowder Publications.

Day, T., Gibson, H., & Ramwell, S. (2017). Fusion of osint and non-osint data. In B. Akhgar, P. S. Bayerl, & F. Sampson (Eds.), *Open source intelligence investigation from strategy to implementation* (pp. 133–152). Springer.

Dean, A., Thompson, E., & Keatinge, T. (2013). Draining the ocean to catch one type of fish: Evaluating the effectiveness of the global counter-terrorism financing regime. *Perspectives on Terrorism*, (4), 62–78.

Drake, N. (2023a). Qubes 4.1.2. *Linux Format*, (11), 22.

Drake, N. (2023b). Tails 5.18. *Linux Format*, (12), 21.

Drake, N. (2024). Tall tails. *Linux Format*, (1), 50–53.

European Banking Authority. (2021). *Guidelines on ml/tf risk factors (revised)*. https://www.eba.europa.eu/regulation-and-policy/anti-money-laundering-and-e-money/revised-guidelines-on-ml-tf-risk-factors

Europol. (2002). *Analyysiopas*. https://op.europa.eu/en/publication-detail/-/publication/c5abefbc-f9ed-472a-875e-2fe378ba4ba8/language-fi/format-PDFX

Ghioni, R., Taddeo, M., & Floridi, L. (2023). Open source intelligence and ai: A systematic review of the gelsi literature. *AI & Society*. https://link.springer.com/article/10.1007/s00146-023-01628-x.

Gibson, H. (2017). Acquisition and preparation of data for osint investigations. In B. Akhgar, P. S. Bayerl, & F. Sampson (Eds.), *Open source intelligence investigation from strategy to implementation* (pp. 69–93). Springer.

Gibson, H., Ramwell, S., & Day, T. (2017). Analysis, interpretation and validation of open source data. In B. Akhgar, P. S. Bayerl, & F. Sampson (Eds.), *Open source intelligence investigation from strategy to implementation* (pp. 95–110). Springer.

Handnagy, C. (n.d.). *Social engineering: The science of human hacking*. Wiley.

Hassan, N. A., & Hijazi, R. (2018). *Open source intelligence methods and tools: A practical guide to online intelligence*. Apress.

Herrera, S., Cruz, C., Ramízer, Y., & Cisternas, L. (2016). Conceptual process design for boric acid: A case study for engineering education. *Computer Aided Chemical Engineering*, (38), 1437–1442. https://www.sciencedirect.com/topics/computer-science/case-study-method.

IPQualityScore. (n.d.). *Fraud detection solutions - bot detection*. https://www.ipqualityscore.com/

Joshua. (2020). *Qubes-whonix™ overview*. https://medium.com/@port1399/qubesos-4-0-3-with-kali-linux-2020-2-f78db2e48e8a

Jyväskylä University of Applied Science. (n.d.). *Opinnäytetyö - thesis: 2.2 research-based development assignment*. https://oppimateriaalit.jamk.fi/opinnaytetyo/en/thesis-implementation-methods-and-results/research-based-development-assignment/

Kadar, T. (n.d.). *The best tools for open source intelligence (osint)*. https://seon.io/resources/the-best-tools-for-osint/

Kali Linux. (n.d.). *Protonvpn*. https://protonvpn.com/support/official-linux-vpn-kali/

Kalpakis, G., Tsikrika, T., Cunningham, N., Illiou, C., Vrochidis, S., Middleton, J., & Kompatsiaris, I. (2017). Osint and the dark web. In B. Akhgar, P. S. Bayerl, & F. Sampson (Eds.), *Open source intelligence investigation from strategy to implementation* (pp. 111–132). Springer.

Kandiko, U. L. (2018). Cyber intelligence: Reinventing the wheel. *Triarius - Prevention and Security Bulletin on Terrorism and the News Threats*, *2*(34), 27–32.

Kernan, W. F. (2001). *Nato osint handbook*. Nato.

Kings, I. (n.d.). *Osint with recon-ng*. https://www.prismacsi.com/en/osint-with-recon-ng/

Kuchmai, O., & Shelest, T. (n.d.). Using open source intelligence (osint) as one of the effective and legitimate ways to avoid threats to the corporation. *Scientific and Practical Cyber Security Journal (SPCSJ)*.

Kurniadi, D. D. (n.d.). The difference between using proxy server and vpn. *Sisforma - Journal of Information Systems*.

lalitmohantiwari7700. (2021). *Maltego tool in kali linux*. https://www.geeksforgeeks.org/maltego-tool-in-kali-linux/

Layton, R. (2016). Relative cyberattack attribution. In R. Layton & P. A. Watters (Eds.), *Automating open source intelligence: Algorithms for osint* (pp. 48–71). Elsevier.

Maltego. (n.d.). *Register a maltego ce account*. www.maltego.com/ce-regisration

Maltego Team. (2020a). *Identify suspicious ip addresses with ipqualityscore transforms in maltego*. https://www.maltego.com/blog/identify-potentially-fraudulent-ip-addresses-with-ipqualityscore-transforms-in-maltego/

Maltego Team. (2020b). *Verifying and investigating email addresses with ipqualityscore transforms in maltego*. https://www.maltego.com/blog/verifying-and-investigating-email-addresses-with-ipqualityscore-transforms-in-maltego/

Mansfield-Devine, S. (2010). Security through isolation. *Computer Fraud Security*, *2010*(5), 8–11.

mohdshariq. (2021). *Ashok – osint recon tool in kali linux*. https://www.geeksforgeeks.org/ashok-osint-recon-tool-in-kali-linux/

Mukonyi, J. (2021). *How to install and setup proxy-chains in linux*. https://www.hackerxone.com/2021/08/17/install-and-setup-proxy-chains-in-linux/

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2015). *Automated data collection with r: A practical guide to web scraping and text mining*. Wiley.

Nakutavičiūtė, J. (n.d.). *Proxy vs vpn: What are the main differences?* https://nordvpn.com/fi/blog/vpn-vs-proxy/

Nato. (2002). *Intelligence exploitation of the internet*.

Neotas. (2017). *Using open source intelligence to battle fin crime*. https://www.neotas.com/using-open-source-intelligence-to-battle-fin-crime/

Norton, R. A. (2011). Guide to open source intelligence: A growing window into the world. *Journal of U.S. Intelligence Studies*, *18*(2), 65–67.

OccupyTheWeb. (2018). *Linux basics for hackers : Getting started with networking, scripting, and security in kali*. No Starch Press, Inc.

Oikeusministeriö. (2024). *Selvitys: Järjestölähtöisiä pankkipalveluja tarvitaan*. https://oikeusministerio.fi/-/selvitys-jarjestolahtoisia-pankkipalveluja-tarvitaan

Pastor-Galindo, J., Nespoli, P., Mármol, F. G., & Pérez, G. M. (2020). The not yet exploited goldmine of osint: Opportunities, open challenges and future trends. *Special section on emerging approach*, *8*, 10282–10304.

Perez, C., & Germon, R. (2016). Graph creation and analysis for linking actors: Application to social data. In R. Layton & P. A. Watters (Eds.), *Automating open source intelligence: Algorithms for osint* (pp. 114–141). Elsevier.

Pocher, N. (2020). The open legal challenges of pursuing aml/cft accountability within privacy-enhanced iom ecosystems. *Proceedings of the 3rd Distributed Ledger Technology Workshop Co-located with ITASEC 2020*.

Ponder-Sutton, A. M. (2016). The automating of open source intelligence. In R. Layton & P. A. Watters (Eds.), *Automating open source intelligence: Algorithms for osint* (pp. 12–31). Elsevier.

Qubes OS. (n.d.). *What is qubes os?* https://www.qubes-os.org/intro/

Revell, Q., Smith, T., & Stacey, R. (2017). Tools for osint-based investigations. In B. Akhgar, P. S. Bayerl, & F. Sampson (Eds.), *Open source intelligence investigation from strategy to implementation* (pp. 153–165). Springer.

RIPE NCC - RIPE Network Coordination Centre. (n.d.-a). *Get to know us*. https://www.ripe.net

RIPE NCC - RIPE Network Coordination Centre. (n.d.-b). *Ripe database*. https://www.ripe.net/manage-ips-and-asns/db

RIPE NCC - RIPE Network Coordination Centre. (2019). *What is an as number?* https://www.ripe.net/manage-ips-and-asns/as-numbers

Şahin, Ö. (n.d.). *The new discipline of intelligence world and its critical component: Cybint and humint* (tech. rep.). Air War College, Turkish War Colleges.

Sharma, M. (2017). Roundup: Privacy distros. *Linux Format*, (4), 22–27.

SigintOS. (n.d.). *Sigintos - about*. https://www.sigintos.com/about/

Slovik, P. (2012). *Systemically important banks and capital regulation challenges* (tech. rep.). OECD.

Staniforth, A. (2017). Open source intelligence and the protection of national security. In B. Akhgar, P. S. Bayerl, & F. Sampson (Eds.), *Open source intelligence investigation from strategy to implementation* (pp. 11–19). Springer.

Suriadi, S., Foo, E., & Smith, J. (2016). Enhancing privacy to defeat open source intelligence. In R. Layton & P. A. Watters (Eds.), *Automating open source intelligence: Algorithms for osint* (pp. 72–89). Elsevier.

Tabatabaei, F., & Wells, D. (2017). Osint in the context of cyber-security. In B. Akhgar, P. S. Bayerl, & F. Sampson (Eds.), *Open source intelligence investigation from strategy to implementation* (pp. 213–232). Springer.

the Financial Supervisory Authority (FIN-FSA). (2021). Toimituskirja finanssivalvonnan johtokunnan päätöksestä: Seuraamusmaksun määrääminen [Accessed: 14.9.2021].

The Kali Team. (2022). *Kali linux: Tools*. https://www.kali.org/tools/

Troia, V. (2020). *Hunting cyber criminals: A hacker's guide to online intelligence gathering tools and techniques*. Wiley.

Valtioneuvosto. (2017). Laki rahanpesun ja terrorismin rahoittamisen estämisestä [Accessed: 12.7.2022].

Valtioneuvosto. (2021). Valtioneuvoston asetus menettelyistä asiakkaan tuntemiseksi ja riskitekijöistä rahanpesun ja terrorismin rahoittamisen estämisessä [Accessed: 12.7.2022].

Watters, P. A. (2013). Modelling the effect of deception on investigations using open source intelligence (osint). *Journal of Money Laundering Control*.

Whoinx. (n.d.). *Qubes-whonix™ overview*. https://www.whonix.org/wiki/Qubes

Yari, S., Mäses, S., & Maennel, O. A method for teaching open source intelligence (osint) using personalised cloud-based exercises. In: 15th International Conference on Cyber Warfare and Security. Tallinn University of Technology (TalTech). Tallinn, Estonia, 2020.

Zusyaku. (2021). *Ashok : Osint recon tool for hackers*. https://zusyaku-id.blogspot.com/2021/10/ashok-osint-recon-tool-for-hackers.html