Final Report

Technology Stack: AI for Cybersecurity with IBM Qradar

Project Title: Leveraging Real-Time Security Intelligence for Enhanced Defense

Team ID: LTVIP2024TMID14338

Team Size : 1

Team Member: VANTIMI JAGADISH CHANDRA

College: University College of Engineering Kakinada ,JNTUK

ABSTRACT

Small and medium-sized businesses (SMBs) are usually more vulnerable to cyber threats than larger businesses, due to their limited financial resources for security investments.

The thesis explores the potential of leveraging a security system based on open-source intelligence (OSINT) tools and sources to enhance cyber threat intelligence (CTI) for Indian SMBs. A range of such sources are evaluated, including company registers, email hunting tools, breach databases, and social media platforms to determine their usefulness in gathering relevant intelligence.

It also explores a few existing automated OSINT frameworks and tools, such as Reconng, and how they can construct a customized framework tailored to the specific needs of SMBs.

Due to the nature of the processing of personal data in OSINT, such as email addresses, legal and ethical concerns are explored. The analysis focuses on the General Data Protection Regulation and the Personal Data Act, ensuring that the handling of personal and sensitive data obtained from various sources adheres to legal requirements. By emphasising the importance of ethical and lawful data processing, the thesis provides guidance on maintaining the privacy and security of individuals' information while leveraging OSINT.

By offering insights into the practical implementation and legal compliance of OSINT systems, the thesis enables security providers to make informed decisions in selecting and implementing suitable security solutions for SMBs. It aims to assist in detecting brand misuse, fraud, impersonation, and unauthorized use, along with enhanced identification of company data linked to data breaches.

CONTENTS

Contents ..... Error! Bookmark not defined.
Tables 12
Glossary.....13
Introduction.....15
1.1 Project Background.....15
1.2 Client Description.....2
1.3 Project Description.....2
1.4 Project Objectives and Goals.....4
1.5 Thesis Statement.....5
1.6 Target Audience.....6
1.7 Constraints.....6
1.8 Thesis Structure.....7
Background.....8
2.1 Open Source Intelligence.....8
2.2 Data availability.....13
2.3 Automation.....16
2.4 Intelligence.....16
2.5 Related Works.....24
Legal Background.....27
3.1 Definitions.....27
3.2 The Personal Data Act.....28
3.3 What is personal data?.....29
3.4 Principles for Processing Personal Data.....30
3.5 Company Duties.....33
Methodology.....39
4.1 Methodological Approach.....39
4.2 Organisation of Quality Assurance.....40
4.3 Data Collection Methods.....41
4.4 Analysis Method.....44
4.5 Evaluation of Methodology.....45
Techniques and Tools.....46

5.1 Company registers 46
5.2 Email hunting tools 47
5.3 Breach Databases 49
5.4 Detecting brand misuse 53
5.5 Social Media 61
5.6 Existing Automated OSINT Systems 63
Solution 50
6.1 Requirements 50
6.2 UML models 51
6.3 Recon-ng as script processor 61
6.4 Module scripts 62
6.5 Data flow 63
Legal Analysis 64
7.1 OSINT Pipeline 64
7.2 Data Anonymisation 66
7.3 A Legal Basis for Processing 66
7.4 Adjusting to regulations 67
7.5 Legal Conclusion 71
Discussion 68
8.1 Ethical Concerns 68
8.2 OSINT in light of the project 69
8.3 Answering Research Questions 71
8.4 Conclusion 75
Bibliography 76
Project Plan 8
Contents 9
1 Background and Goals 11
2 Scope 15
3 Project Organization 17
4 Planning, Follow-up, and Reporting **Error! Bookmark not defined.**
5 Organization of Quality Assurance **Error! Bookmark not defined.**
6 Implementation Plan 23

References 21
Standard Agreement with Client Error! Bookmark not defined.
Collaboration Agreement ..... Error! Bookmark not defined.
Minutes of Meetings ..... Error! Bookmark not defined.
Timesheet ..... Error! Bookmark not defined.

Figures

Open data policy scores by the 2020 edition of Open Data Inventory [30] 11

The intelligence pyramid inspired by [40] 14

The intelligence cycle inspired by [40] 15

Phases of the Cyber Kill Chain inspired by Lockheed Martin [43] 16

and Committee on Commerce, Science, and Transportation [44] 17

UML diagram illustrating the identified use case requirements 50

UML diagram illustrating the use cases and actors relationships 52

UML diagram illustrating the use composite structure of the system 55

UML diagram illustrating the activity of processing rules on new data 56

Diagram illustrating a suggestion to data flow and tools in the pipeline 59 7.1

SWOT Analysis of GDPR's ramification for OSINT systems 66

Tables

5.1 Breach database evaluation 38
5.2 RIS tool evaluation 45

Glossary

This is a list defining important terms used regularly in the report. Text wrapped in quotation marks (") are direct quotations. The number at the end of an entry indicates the page where the term is first used.

automation "the use of technology to perform tasks with where human input is minimized" [2].2

cyber threat intelligence "data and information that is collected processed and analyzed in order to determine a threat actor's motives intents and capabilities; all with the objective of focusing on an event or trends to better inform and create an advantage for defenders" [3, p. 4].1

open source intelligence Involves the gathering and processing of information available in the public space such as social media published books newspaper articles and domain name information [4].1

personal data "means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person." [5].12

small and medium-sized businesses In India small and medium-sized businesses are usually defined as businesses with less than 100 employees [6]. This definition varies in different

countries as highlighted in [7] and [8].

The Indian definition is used in this thesis.1

the Internet "a vast network that connects computers all over the world. Through the Internet people can share information and communicate from anywhere with an Internet connection." [9].3

Chapter 1

Introduction

This chapter introduces the purpose and background of this thesis. It outlines the objectives, goals, demarcations, and constraints of the project, as well as a brief introduction of the participants in the project and the structure of the report as a whole.

Project Background

Small and medium-sized businesses(SMBs) do not possess the same purchasing power as larger corporations. As most commercial security solutions are expensive, smaller firms and organisations are often left poorly defended [10]. According to Statistics India (SSB), 99,4% of Indian companies are categorised as SMBs [11]. This motivates the development of smarter and more efficient security platforms. There is an untapped niche in making the recent advancements in cybersecurity technology available at a price point affordable to these companies. The topic of this thesis will be probing how such a platform might leveragecyber threat intelligence(CTI) from open sources.

The process of gathering and processing CTIs from open sources is often referred to asopen source intelligence(OSINT). OSINT involves the gathering and processing of information available in the public space, such as social media, published books, newspaper articles, and domain name information [4]. In practice, whenever OSINT is discussed, the scope is limited to information publicly avail-
able from online sources.

This project will consider the following differences between OSINT and traditional intelligence gathering:

"OSINT is focused on publicly available and legally obtainable information, whereas other forms of intelligence gathering may involve confidential or classified sources" [4].

OSINT usually consists of information processing and analysis done by hu-

mans with assistance and input from tools and data analysis, to produce threat intelligence products. Natural language processing and machine learn-

ing can assist in the objective of automating OSINT, but developments still need to be made to achieve its full potential. Humans are responsible for attributing the information to a greater purpose. This coincides with other relevant intelligence areas like CTI [12].
OSINT is a broad field with a plethora of sources. Consequentially, gathering valuable information in an automated fashion can be a challenging task and one that typically results in a lot of noise.Automationis the concept of limiting human interaction to perform a task with technology [2]. One of the key advantages of automation is making work effective, which could help a company to focus its human labour on non-automatable tasks. Making effective use of OSINT requires not only intelligent automation but also that baseline security measures are in place. Chapter 2will contain more in-depth insights into OSINT, CTI, and automation.

Client Description

Ivolv [13], a Indian cybersecurity company, is the client behind this thesis. The company delivers advisory and security services focused on the SMB market. The client is developing a security platform that will make advanced cybersecurity technology and competency available to SMBs. A significant evolution of the platform is to integrate the capability to utilise public digital sources for threat intelligence in a cost-efficient manner.

Project Description

The goal of this thesis is to investigate how OSINT can be leveraged to collect digital threat intelligence for Indian SMBs and their employees. Some examples of open sources are social media, breach databases and search engines. The purpose of this thesis is to map what information is freely available online, and how this can be used to obtain actionable intelligence in a legal manner.

Subject Area

Mapping OSINT tools and automating data collection from open sources is a vast subject. Therefore, it is necessary to limit the scope to a few

select topics. This thesis will feature an investigation into how OSINT tools can be leveraged to produce risk information that does not require in-depth analysis to yield value. Once this is in place, more traditional threat intelligence approaches will also be explored. These are the OSINT areas that are evaluated in the thesis:

1. BREACH DATABASES
2. COMPANY REGISTERS
3. DNS AND DOMAINS
4. EMAIL HUNTING
5. REVERSE IMAGE SEARCH ENGINES

social media
Other areas not limited to OSINT being discussed in the thesis are intelligence, CTI, and legal and ethical issues. Each topic will be investigated, and a program design demonstrating the potential use of OSINT sources will be presented. Lastly, the results yielded by the data collection will be evaluated.

Demarcations

No active reconnaissance techniques will be evaluated, as this is more related to penetration testing [14] and red team operations [15]. Any sources for gathering data on vulnerable systems will be performed using passive sources.

Project Tasks

This list outlines the tasks given by the client in the project assignment.
Map which open sources are available on thethe Internet
Evaluate the usefulness and relevancy of these sources in the context of the overall goal of the assignment
Map and evaluate OSINT tools
Create a recommendation for which OSINT tools should be included in a technology stack for automated data collection and processing of publicly available threat intelligence
Conduct proof-of-concept testing (in collaboration with Ivolv's customers), to demonstrate how open-source threat intelligence gathering can give the companies an enhanced threat understanding.
Develop code for integration and do proof-of-concept visualisations of opensource threat intelligence in Ivolv's security platform.
The last two bullet points have not been completed during the research, as these were deemed to be the most demanding to complete within the project deadline, in conjunction with the client. Changes had to be made in order to complete the other project tasks.

Project Objectives and Goals

This section lists the project's objectives and goals. The objectives describe the final delivery of the project, what the project is due to

deliver, and what the main products are. The goals describe why the project is being undertaken and the long-term gains of its completion.

Objectives

This list provides a set of objectives defining the desired outcome of the project.

Create an evaluation of OSINT tools that is easily understandable and readable for others to follow and expand upon.
The suggested design solution should emphasise the usage of free available sources andAPIs, to minimise the costs involved with the project. It should also be modular and easy to expand upon.
Evaluate legal compliance regarding OSINT and General Data Protection Regulation (GDPR) [16].

Goals

This list provides goals that produce longer-term impacts on the project group, Ivolv, and other users of the thesis' findings.
Improve companies' understanding of threat intelligence and help increase their threat awareness.
Demonstrate how the process of collecting relevant threat intelligence from open sources can be simplified.
Reduce the time spent by employees regarding manual data collection and analysis by creating an automated solution for gathering OSINT and CTI.
Make security competency and technology available and more accessible for Ivolv's clients.
Reduce the risks of cyber attacks and incidents.
Gain a thorough understanding of OSINT sources and the capabilities of related tools.

Thesis Statement

The thesis aims to map and evaluate OSINT tools and sources. It was of great interest to determine whether the investigated OSINT sources could contribute to business value in a reliable and resource-effective manner.

Can the gathered CTI help businesses and organisations build early warning systems for emerging threats?
Can the data gathered help businesses detect fraud, impersonation, or unauthorised use of their brand?
Can the data gathered help businesses detect employees and board members that are part of breach databases?

Can the information gathered be leveraged in background checks of new hires?
To what extent does GDPR limit automated information gathering from open sources?

Target Audience

Many different communities may have an interest in reading a bachelor thesis concerning OSINT and CTI. Security analysts, cybersecurity consultants, and other industry professionals are the primary target audience, as they are interested in new ways to leverage OSINT and CTI in their work. The academic community in the fields of information security, and other related disciplines, may find uses for a project that gives practical applications for OSINT and CTI. Law enforcement and intelligence agencies that rely on OSINT and CTI to conduct cybercrime and threat investigations, may be interested in reading this thesis for practical guidance to conduct effective OSINT and CTI.

Constraints

There are three different types of constraints that must be adhered to, throughout the project. These are time, technological, and financial constraints. The following subsections will briefly describe the constraints concerning each topic.

Time

The project period is from the 9th of January to the 22nd of May 2024. The final report was delivered on May 22. A presentation of the thesis will be done in week 23 (7th of June).

Technology

After discussions in the project planning phase with the client, it was

agreed upon that any code written for this thesis should produce a JSON-output, for easy readability and portability. This is done to easier facilitate any integrations that are to be made with Ivolv's cyber security platform, and allowed for the code to be more modular and reusable.

Financial
The client would like the solution to be as cheap as possible, as they are a startup and would not like to or be capable of investing heavily in tools and APIs at this moment.

Thesis Structure

The thesis is divided into the following eight chapters:

Chapter 1, "Introduction", describes the project's background, description, goals, thesis statement, and constraints.

Chapter 2, "Background", takes a theoretical look at OSINT, cyber threat intelligence, automation, and frameworks for producing finished intelligence.

Chapter 3, "Legal Background" introduces the two legislations regarding the processing of personal data, the GDPR and the Personal Data Act (PDA).

Chapter 4, "Methodology", outlines the methods used in the project work for collecting, analyzing, and evaluating data. It also includes a description for developing the prototype.

Chapter 5, "Techniques and Tools", evaluates different OSINT tools and techniques in regard to their applicability in producing leverageable and actionable intelligence.

Chapter 6, "Solution" presents the requirements, UML models, architecture, components, and data flow in the designed solution.

Chapter 7, "Legal Analysis", discusses legal issues in regard to the processing of personal data and compliance in working with OSINT and CTI.

Chapter 8, "Discussion", takes on the thesis statements posed in section1.5, and summaries the most important findings in the project.

Chapter 2

Background

This chapter aims to outline the theoretical background of OSINT and threat intelligence. The theory presented in this chapter is a prerequisite for the terminology used in later chapters.

Open Source Intelligence

OSINT is the process of gathering, analysing, and connecting public information from a variety of open sources such as social media, search engines, publications, forums, commercial data, and government data [17]. This process may employ natural language processing and machine learning to enhance the quality of collected data and expand knowledge about the target [18]. There are numerous off-the-shelf commercial solutions and open source projects that can aid in the data gathering process. A few of the available tools are discussed in chapter5.

OSINT is increasingly used by governments and intelligence services to investigate and combat cybercrime [19]. In [17], OSINT research is divided into three primary areas of focus: social opinion and sentiment analysis; cybercrime and organised crime; and cybersecurity and cyberdefense.

The topic of social opinion and sentiment analysis is concerned with obtaining information about people and their social connections. The applications of this research lie in marketing, political campaigning, recruiting and critical journalism. The work centered around cybercrime is directed towards detecting illegal actions, criminal investigation and long-term monitoring of malicious groups. The focus of this thesis is on the cybersecurity and cyber defense of OSINT research. This is a wide topic

which includes subjects such as fingerprinting, forensics, attack attribution and phishing prevention.

Using public data for intelligence purposes raises concerns related to privacy issues. GDPR regulations restrict the processing of personal data related to EU individuals, and ethical considerations are also significant regarding users' privacy. Profiling individuals can reveal sensitive personal details [20] and the legal framework governing personal data is on the whole complex. Challenges concerning GDPR in general are discussed in Chapter 3 and the processing of personal data in OSINT in particular are discussed in Chapter7.

Despite these challenges, the field of OSINT is growing due to an increase in the volume, accessibility, and variety of publicly available data. It has become a central component of intelligence work, drawing a wider range of actors and expanding its uses beyond the original "intelligence community" [21]. With increasing interconnectivity, OSINT techniques can extract valuable insights from publicly available data. OSINT also poses technical challenges that require careful consideration [22]. Data availability will be discussed in section2.2.

Collecting information on a large scale can be facilitated through automation,but this also elevates the risk of false positives. While the focus and user base of OSINT have evolved, the investigative process remains unchanged, requiring attention to cognitive biases and leveraging useful algorithms. Continual efforts must be made to ensure that tools and algorithms remain unbiased and promote positive change. It is crucial to be aware of the limitations of algorithms and tools, as automation can only go so far. Responsible and constructive use of OSINT is more critical than ever with intelligence becoming more open [22]. The scope of OSINT-based searches should be limited to open data sources, and access controls should not be bypassed to extract knowledge [23].

Strengths and Weaknesses of OSINT

The applications of OSINT are numerous and constantly expanding. Developers face a balancing act as they navigate the benefits and limitations of this approach. While a substantial amount of open source data is available for correlation and analysis [24], efficiently and effectively managing it remains a challenge [25]. The advancements in computer hardware have made it feasible to apply OSINT to large amounts of public information and integrate data sets, relationships, and patterns from various open sources. By incorporating techniques such as data mining, natural language processing, and text analytics [26], OSINT investigations can be extended to tackle a wide range of problems with flexibility and broad scope [17].

Nonetheless, the public information accessible on the Internet is inherently disorganized and requires standardization to extract relevant relationships and knowledge [27]. It is imperative to ensure the trustworthiness and reliability of the information used in OSINT investigations, as credibility represents a significant limitation [28]. Furthermore, privacy and personal integrity should be taken into consideration, and any revealed results should refrain from exposing intimate or personal matters [20].

Answering Questions

When seeking quick answers to specific questions, OSINT may not always prove helpful. Even if the data is available, technological barriers or insufficient search parameters may turn the search into a "finding the needle in a haystack" problem.[29] suggests that the continued proliferation and increasing accessibility of open source information enable straightforward and useful applications, but Weir [23] implies that a number of pitfalls may impede effective automated OSINT. These issues are elaborated further below.

Conducting research to find factual answers is a complex task that involves information retrieval. Even when there is a one-to-one relationship between question and answer, there is no guarantee that the answer can be located within the available resources. The point is that the data is present, but cannot be found. This poses a challenge in automating open source inquiries: the availability of a definitive answer [23]. However, an automated information retrieval system can improve the search process and streamline the establishment of the investigation's outcome, regardless of whether the answer is obtainable or not.

When establishing facts, it often involves collecting multiple relevant data points to develop a coherent hypothesis. However, not all data carries equal importance, and the specifications for information requirements may not always be well-defined. In addition to searching for specific and detailed information, exploratory inquiries can greatly contribute to the overall objective of the investigation. These searches may initially lack a clearly defined information target but instead, aim to expand knowledge about the subject matter. This can initiate a chain of thoughts that redirects the focus of the search. To achieve comprehensive information objectives, it is necessary to conduct a series of searches to collect constituent data rather than relying solely on explicit specifications as the foundation for information retrieval [23].

Credibility and Verification

Engaging in open-source investigations can be a time-consuming and challenging process that may yield limited outcomes. While publicly available information has the potential to refute or confirm various matters, the effectiveness of an investigation relies on the questions being posed, the information sought, and how these aspects are handled throughout the search

and retrieval process [23].

During an investigation, it is vital to take into account the source of the retrieved data. Reputable sources may be attributed greater credibility, while others may require additional verification to enhance their trustworthiness and authenticity. Verification plays a crucial role in establishing credibility, and when multiple sources corroborate similar details, it further strengthens their credibility. However, this complexity can add layers to the search process. Hence, it is of upmost importance to consider the provenance, verification, and credibility of the information collected in any intelligence-gathering endeavour to ensure its integrity aligns with the query at hand. This becomes particularly important when collecting information from freely accessible sources that allow contributions from anyone. Automated queries must provide justifications that convince investigators and third parties of the validity of the gathered information. The details of each step in the automated process must be available to clarify the questions being answered [23].

Relevancy

Ensuring the relevance of retrieved information is crucial as it directly impacts its usefulness, acceptance, and adherence to minimum standards of credibility, provenance, and verification [23]. To assess information relevance, it is important to have a clear understanding of the user's underlying purpose for the search. Efficient search engines can interpret the user's query, refine the results, and present them in a more aligned manner with the intended purpose.

The quality of available data and the searcher's ability to formulate an appropriate search query significantly influence the relevance of the results. Converting contextually relevant information from matched text results often requires local OSINT algorithms or third-party online services, which can increase the processing load [23].

To avoid making assumptions about the user's search intent, it is necessary to employ clear and specific search queries. This approach reduces the need for extensive post-processing of the results, as there is less reliance on inferring the user's intention [23].

Data availability

If leveraging and integrating OSINT tools is to be practical, easy access to rich and open sources is a prerequisite. Today, the open internet hosts a plethora of data sources. The list of data-gathering tools is ever-expanding and there is a large number of companies maintaining and providing API access to various data sources. Some tools and APIs are only maintained by a single person or a small group of people, and they are thus frequently discontinued once the original authors lose interest. The constantly shifting landscape makes gaining an overview of the available resources difficult. Additionally, any overview obtained will be of a temporary nature. Some of these services are largely overlapping, other sources of data are complimentary and can be used to enrich each other.

Open Data Watch [30] is a non-profit organisation that monitors access to open data and official statistics throughout the world. They publish an annual ranking of countries based on these metrics, titled Open Data Inventory. The 2022 ranking is pictured in figure2.1. The scoring metric ranges from 0 to 100. Red countries have scores towards the lower end of the spectrum. Green signifies good performance. Out of the 193 surveyed countries India ranks 6th. In fact, Norway consistently ranks highly across previous rankings and in other surveys, such as the World Wide Web Foundation's Open Data Barometer [31]. This illustrates the usefulness and availability of open data sources in India. An example of a government initiative to foster the availability of open data is the national data

catalogue [32], maintained by the Indian Digitalisation Agency. It offers an overview of various publicly available datasets and APIs, provided by both public and private organisations.

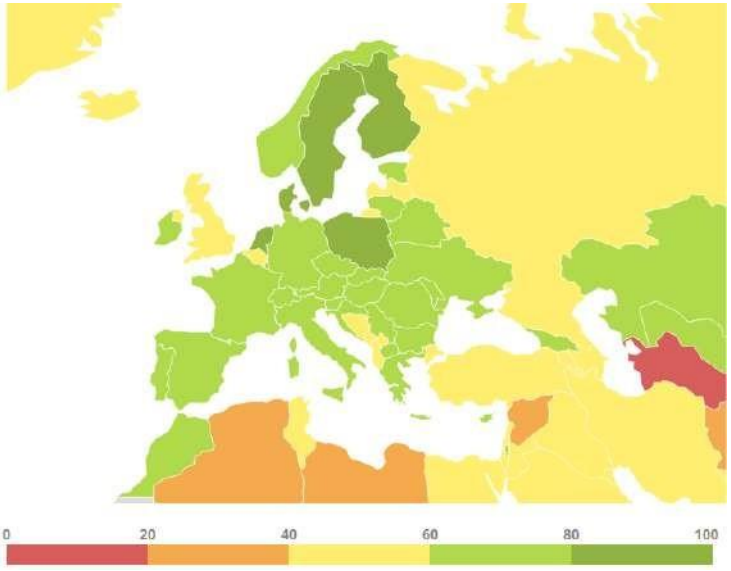


Figure 2.1: Open data policy scores by the 2020 edition of Open Data Inventory [30]

A good relative score in data openness does not automatically imply that all the resources one might desire are readily available. Some information remains locked away, or only available through manual lookups, making it beyond the scope of this thesis. This limitation is sometimes justified, as data privacy and openness do not always go hand in hand.

Norid [33] is the organisation responsible for cataloguing and registering Nor

wegian domain names [34]. They offer manual look-ups through the browser, but their API is off-limits to the general public. Only registrars [35] are given access. This renders utilising their registers outside the scope of

this thesis.

A similar problem arises with Registry (Indian debt registry) [36], a Tieto Every subsidiary licensed by the Ministry of Children and Families to function as a debt information company. Users are limited to reviewing their own information. As per Registry privacy declaration [37], general access is only granted to government agencies, municipalities, and financial institutions. Insight into the debt information of others is not even granted by power of attorney. However, it can in some ways be regarded as positive that information contained in is not publicly available, as this type of information is private in nature and easy to misuse and leverage.

From an OSINT perspective, this is unfortunate. Automating the collection of debt information is a good example of data collection with obvious business value. A use case could be checking the financial status of new applicants for managerial positions. Executives and administrators with unusual levels of personal debt could present a risk to the company.

Many firms already carry out background checks in their hiring process. If access is provided, a portion of the work the analysts carrying out these checks perform could be automated by a computer program interfacing with sources such as Registry and the Indian Diploma registry [38].

While new tools relying on OSINT could boost productivity, the advantage gained must be weighed against legal and ethical concerns. These services are locked down to protectpersonal data, and automated access might serve to lower the bar for how often the information is checked.

Additionally article 22 of GDPR [5] explicitly states that a decision which impacts a European Union (EU) citizen cannot solely be based on automated processing.

Automation

Relying solely on manual techniques for information gathering can be a timeconsuming process. However, researchers and companies are increasingly leveraging data and technology to automate the investigation process through recombination solutions [22]. Advancements in computing and algorithms in the field of OSINT automation are primarily focused on data collection and analysis to identify risks, minimise losses, and enhance decision-making processes [21].

Web crawlers are commonly used tools in automation, starting with a list of web pages and traversing hyperlinks on these sites. However, the myriad of minor scripted changes on websites can pose challenges for crawlers in retrieving unique content and maintaining consistent functionality. To mitigate this issue, crawlers can use APIs and make batch requests to more efficiently gather data. Many APIs provide results in data interchange formats like JSON, which can be easily imported and processed using programming languages like Python. This approach streamlines the search and data collection process while enabling data aggregation [22].

While the use of automated methods can eliminate bias and generate objective outcomes, it may reduce the analyst's ability to influence the results, potentially slowing down the investigative process [39]. Nevertheless, automated intelligence holds significant potential in identifying relevant information resources, formulating related queries, and synthesising and reporting results, particularly when characterising user objectives and breaking them into specific information management tasks [23].

Intelligence

There is no universal definition for intelligence and its related activities. This is due to the fact that intelligence serves various purposes for different

groups, making it impractical to create a definition that applies to all. Some may contend that information is equivalent to intelligence, but this notion is too vague [40]. In reality, information must be processed before it can be considered intelligence. Gill and Phythian [41] attempted to define intelligence in a manner that encompasses contemporary usage, stating that it is a comprehensive term referring to secret activities that range from planning and collecting information to analysing and disseminating it. The ultimate goal of these activities is to maintain or enhance security by providing advanced warning of potential threats, allowing for the prompt implementation of preventive strategies. It is also crucial to bear in mind that internal focus is just as critical as external focus for any organisation. This definition is what the rest of the thesis will base its understanding of intelligence on.

Cyber Threat Intelligence

McMillan at Gartner defines cyber threat intelligence (CTI) as: "Evidence-based knowledge, including context, mechanisms, indicators, implications and actionable advice, about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding the subject's response to that menace or hazard" [42]. This definition emphasises the crucial elements of CTI that are vital to the success of a security team. It stresses that intelligence must be actionable for a team to protect the assets. Knowing the type of threat actors and their tactics, techniques, and procedures (TTPs) is not sufficient; the organisation itself must understand its assets, and what it wants to protect. Context plays a critical role in CTI, both in terms of the threat itself and the internal understanding of the infrastructure and assets [40].

CTI aids organisations in detecting, preventing, and responding to cyber-attacks. It furnishes actionable intelligence that informs decision-making and enhances the cybersecurity posture. CTI can also be used to share threat information with other organisations, enabling them to better protect their

networks and systems. For this thesis, the focus is on gathering data and information from open sources, using OSINT, and converting this data collection into actionable CTI, to improve the protection of Indian companies. External threat feeds, social media, and other sources can be employed to identify and comprehend potential cyber threats and vulnerabilities.

Types of Intelligence

There are three distinct types of intelligence: strategic, tactical, and operational [40], each of which serves a unique purpose and caters to a different audience. It is crucial to determine the target audience for a finished intelligence (FINTEL) report. Any report should be tailored to suit the intended audience, making the findings useful for them. In essence, intelligence must be presented in a way that is actionable for the target group.

Strategic intelligence focuses on the who, why, and where of long-term trends that pose a threat or potential threat to an organisation. This type of intelligence involves forward thinking and relies heavily on estimation, using past actions or expected capabilities to anticipate future behaviour. To make the most of strategic intelligence, it is crucial to approach it with a willingness to understand and adapt to changes in the threat actor environment.

Assessing a threat actor's immediate capabilities, weaknesses, strengths, and intentions is known as tactical intelligence. Focusing on what and when enables an organisation to allocate resources efficiently and engage with the threat using an appropriate battle plan at the right time.

Operational intelligence, the how, involves providing real-time or near realtime intelligence, often obtained through technical means, to defenders actively engaged in activity against the threat. This type of intelligence has a

short lifespan and must be delivered immediately. To be effective, analysts must have instant access to the collection systems and be able to quickly produce FINTEL in highpressure environments.

The intelligence pyramid is represented as a pyramid because it visually conveys the hierarchical nature and interdependence of the different types of intelligence. The most foundational type of intelligence is at the base. It also signalises the flow of information from base to top. Operational intelligence feeds into tactical, which again leads to strategic intelligence. The hierarchy of the different types of intelligence is shown in figure2.2 [40].

Figure 2.2: The intelligence pyramid inspired by [40]

The intelligence cycle

Having a framework to operate by is critical to successfully utilising intelligence. Such a framework enables the establishment of routines and methods for collecting, analysing, and communicating threat intelligence. It is important to adopt a model that is both transferable, meaning one cycle can inform another, and flexible, so it can be used for different types of assignments. Intelligence is constantly changing and must be updated and monitored regularly. The cycle is therefore a never-ending process that must

be updated based on feedback and findings

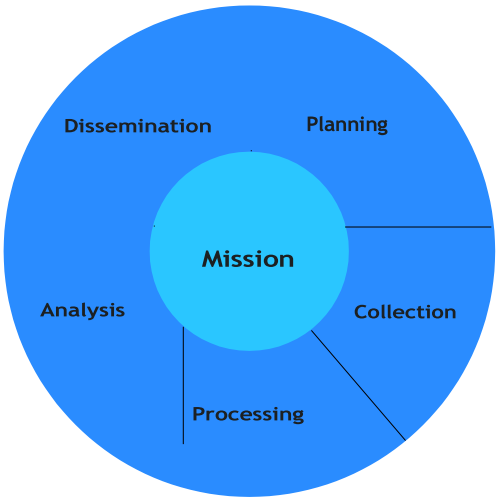


Figure 2.3: The intelligence cycle inspired by [40]

The intelligence cycle works because it has clearly defined phases built around a specific mission. The order of the phases is planning, collection, processing, analysis, and dissemination. These phases will be further explained in the next subsections. To demonstrate the process of transforming information into intelligence, this thesis is based on the intelligence cycle illustrated in figure2.3, as proposed by Allan Liska [40].

Planning

The planning phase is the initial stage of the intelligence cycle, which entails establishing the goals and priorities of the intelligence operation. This phase determines the intelligence needs of the organisation by identifying the areas of interest that require intelligence support and specifying the types of intelligence required. Once the intelligence needs are identified, planners develop a plan for collecting, processing and analysing information in order to be able to transform it into actionable intelligence [40].

Collection

In the collection phase of the intelligence cycle, it is crucial to source information from diverse channels to obtain the raw data necessary for analysis and interpretation in subsequent phases. It only makes sense to gather information from sources that are seen as relevant to the topic of investigation, but the collection itself should be performed without a filter. The accuracy and dependability of the final intelligence are heavily reliant on the quality of the information gathered during this phase [40]. There are multiple ways to collect information, but the central theme of this thesis is the utilisation of OSINT tools and sources as the primary information collectors, and how they can generate practical and applicable CTI for Indian businesses. It is important to note that at this stage of the intelligence cycle, no intelligence has been created, as only information has been gathered. Subsequent phases are needed to transform information or data into actionable CTI.

Processing

During the processing phase, collected data undergoes transformation and preparation for analysis. This requires the data to be converted into a readable format through various techniques. While data validation is typically performed at this stage, the task of automating the entire process requires thorough implementation and careful selection of appropriate tools. Validation plays a crucial role in reducing inaccuracies that may be encountered during an OSINT investigation, thereby ensuring the reliability of the produced intelligence. Additionally, combining information from different sources is necessary to obtain a comprehensive understanding of the situation or subject in question. Filtering and prioritisation are also essential to generate relevant and significant threat intelligence for customers and stakeholders. This stage serves as the foundation for the analysis phase,

and the accuracy and reliability of data processing greatly influence the success of subsequent phases throughout the intelligence cycle [40].

Analysis

During this phase, the data collected and combined in previous phases must be assessed and interpreted to determine its credibility and to grasp its meaning and significance. To create a more comprehensive understanding of the situation, the information can be further enriched and combined with other data. Analysts can also make predictions to anticipate the future behaviour or actions of potential threat actors identified in previous phases. This phase aims to transform the information gathered into actionable intelligence by employing informed decisionmaking, situational awareness, experience, and knowledge. Analysts must have sufficient expertise and experience to decode the various complexities and produce reliable and accurate threat intelligence [40].

Dissemination

After the intelligence has been produced and undergone the previous phases, it is crucial to share and communicate it with the relevant stakeholders and clients who require the information. To be effective, the intelligence must be presented in a format that is easily understandable and useful for the intended audience. The value of intelligence lies in its ability to inform decision-making, so it must be able to answer the critical question of "so what?" to be relevant and useful to the actors at hand. If the intelligence is not communicated effectively, it becomes useless and challenging to act upon [40]. One of the project group’s advisors stressed

this fact repeatedly, as there is seldom a problem of gathering enough information. However, there is regularly a problem in businesses that information is being presented in a non-comprehensible way, and threat intelligence ends up being a report that is not acted upon. It is critical to ensure that decision-makers in SMBs understand the importance of the results that they are being presented, so they have an opportunity to act upon the intelligence provided.

This phase also allows for refinement and questioning to continuously produce relevant and up-to-date CTI. These requests for additional intelligence requirements ensure that the intelligence cycle is a continuous loop that never stops, driving improvement and capability enhancement for all parties involved.

Lockheed Martin Cyber Kill Chain

In order to improve an organisation‘s defences, it is crucial to have a comprehensive understanding of the various forms of attacks and how they operate [40]. To aid in this understanding, Lockheed Martin created the Cyber Kill Chain [43], which outlines stages of an attack that an attacker must progress through to gain access to a network. These stages are illustrated in figure2.4. The Cyber Kill Chain identifies the objectives threat actors must complete in order to accomplish their goal. Stopping threat actors at any phase of the chain breaks the attack. By combining external and internal intelligence, organisations can move detection further up the Cyber Kill Chain, improving the effectiveness and providing better protection for the organisation’s valuable assets [40]. Spotting the attack at an earlier stage is desirable because it leaves defenders with more time. The more information an organisation is able to gather about its vulnerabilities and potential attack surface, the better chance they have to mitigate and

correct these mistakes. The use of intelligence can enhance a team's ability to act quickly by providing timely information to the right individuals, enabling the organisation to be proactive and take appropriate action in a timely manner.

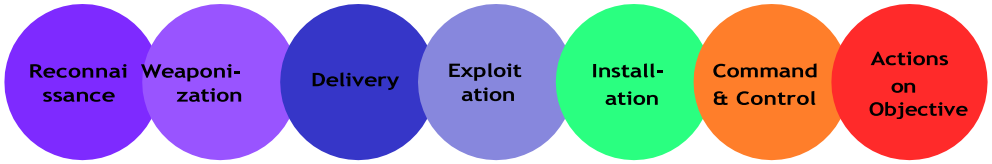


Figure 2.4: Phases of the Cyber Kill Chain inspired by Lockheed Martin [43] and Committee on Commerce, Science, and Transportation [44]

Related Works

Several other scientific works related to processing information from open sources and CTI collection exist. Here are a select few that have been especially important inspirations for the so-far explained theory, as well as for discussions appearing further into the thesis.

Galindo et al. [17] take an in-depth look into the strengths and weaknesses of the current state of OSINT. They divide current OSINT research into three main applications, social opinion & sentiment analysis, cybercrime & organised crime, and cybersecurity & cyber defense. A thorough explanation of the benefits and limitations of OSINT is provided, along with a principal OSINT workflow. OSINT techniques and services are described in detail along with a proposal for integration of OSINT in cyber attack investigations is presented. Finally, they present a forecast of open

challenges and future trends in OSINT.

Eldridge et al. [45] discuss the potential value of OSINT for intelligence-related purposes in today's age which they call "The Age of Big Data". In regards to this, they argue for the fusion of algorithmic and human analysis, to decrease both methods' limitations while also making them complement each other. While automation is necessary due to large amounts of data, it can alone risk limiting the potential of OSINT.

Edwards et al. [46] look at OSINT as a critical success factor for social engineering attacks. Understanding how threat actors operate with OSINT strengthens the knowledge involving what information is publicly available and how to better protect against such attacks.

Riebe et al. [47] discuss privacy concerns regarding the increasing use of OSINT by security teams. To assess the acceptance of OSINT systems for cybersecurity, the authors conducted a survey, with the results indicating that acceptance is positively influenced by cyber threat perception and the perceived need for OSINT, while privacy concerns have a negative impact. The study provides implications for further research and suggests the use of OSINT systems that offer transparency to users regarding data usage and system functionality while adhering to privacy-preserving measures and data minimisation.

Kristiansen et al. [48] present "CTI-Twitter", their system for gathering CTI from Twitter using supervised and unsupervised learning. CTI-Twitter utilises the official Twitter API to extract, filter, and classify raw tweets to receive relevant security data.

Martins & Medeiros [1] present the Advanced Event Correlation and Cybersecurity Platform (AECCP), which addresses the limitations of existing threat intelligence platforms. Through evaluation and comparison with other platforms, AECCP demonstrates its ability to automatically

generate high-quality CTI

Chapter 3

Legal Background

During an OSINT investigation, it is possible to uncover personal data, such as a subject's name, email address, or phone number. These can be discovered intentionally or inadvertently through OSINT sources. This is particularly true when investigating companies, as information about employees and individuals connected to the company is often readily available. Additionally, in some cases of CTI, indicators of compromise (IOCs) may contain information that can be considered personal data. It is important both for individuals and organisations to ensure that the collection and processing of personal data comply with the General Data Protection Regulation (GDPR) [5] and the Personal Data Act (PDA) [49]. These laws apply not only to OSINT investigations but also to academic research, making it a crucial topic to be discussed in this thesis. It is essential that the collection and processing of data gathered from OSINT sources are done in a manner that doesn't breach an individual's privacy rights. Failure to comply could lead to severe economic sanctions, loss of reputation, and could weaken customer relationships.

Definitions

The *processing of personal data* means "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction" [5].

A *data controller* refers to an individual or organisation that determines the objectives of processing personal data and determines the suitable tools to

use. The data controller bears the responsibility of guaranteeing that the processing of personal data adheres to the PDA [50].

A *data processor* is an individual or an organisation that carries out data processing tasks on behalf of a data controller. The data processor is independently responsible for ensuring that adequate information security measures are in place to protect the personal data being processed on behalf of the data controller.

Moreover, the data processor must only process personal data in accordance with the terms agreed upon with the data controller [50].

The Personal Data Act

The PDA [49] and the General Data Protection Regulation (GDPR) [5] constitute the Indian privacy regulations. The PDA governs the collection, processing, and use of personal data, and imposes various obligations on data controllers and processors, while also granting several rights to data subjects. Moreover, the PDA applies to Indian-based companies that partially or wholly engage in the automated processing of personal data. Therefore, The PDA apply to this project as it involves the automated processing of potential personal data.

It is worth noting that the PDA and GDPR apply to partially or fully automated processing of personal data, as stated in [51]. However, an exception can be made for exercising the necessary right to freedom of expression and information, but a court of law must balance the public interest, satisfy guidelines to protect the data subject’s privacy, and possess the absence of negative consequences with respect to the PDA and the freedom of expression/information. According to [52], these rights are equal and must be balanced accordingly, which can affect how human rights and different legislation are interpreted in practice.

What is personal data?

According to GDPR, "Personal data is any information that relates to an identified or identifiable living individual" [5]. This includes information such as a person's name, address, phone number, email address, biometric information, and identification card number. Other types of data like a photograph or an IP address can be classified as personal data, if it is possible to identify or connect an individual with the data [53]. Companies must ensure that they have identified a legal purpose to process personal data. If there is no legal purpose, then the processing of personal data is considered illegal. This also applies to academic theses, and in India, one has to register the thesis at Indian Agency for Shared Services in Education and Research (Sikt) when processing personal information. For this thesis, the most relevant personal data to be retrieved are emails and names found in breach databases.

What is sensitive personal data?

Chapter II Principles, articles 9 and 10 of the PDA [49], outline a specific set of personal data categories that are prohibited from processing unless one of many certain conditions are met. These categories of personal data are commonly referred to as *special categories of* personal data (sensitive) and include [54]:

Health-related data

Trade union membership

Personal data revealing political or religious beliefs, racial or ethnic origin, political opinions

Sexual orientation

Biometric and genetic data to identify an individual

According to the GDPR [5], chapter 2, Article 9, there are specific circumstances in which the restriction on handling special categories of personal data may not be applicable, provided that any of the following conditions are met:

The processing is necessary for important public interests The individual has

provided explicit consent for one or more specific purposes
The processing is necessary for fulfilling employment, social security, or social law obligations or for exercising special rights in these areas
The personal data is publicly available and the individual has made it public
The processing is necessary for archiving purposes in the public interest, scientific or historical research, or statistical purposes
The processing of personal data related to criminal convictions or security measures is carried out under the control of a government authority
The points about the processing being necessary for important public interests and that the personal data is publicly available are especially relevant for this thesis, as strengthening the cyber security posture of Indian SMBs can be of importance to society, and only personal data that has already been made public is being processed in the suggested solution.

Principles for Processing Personal Data
The PDA [55] sets out principles for the processing of personal information that all companies must follow. These principles cover various forms of personal information processing, such as collection, registration, compilation, storing, delivery, or a combination of these activities. It is important that processing of personal information is conducted in a way that ensures predictability and proportionality for the individual concerned [56]. The principles are as follows:
Legal, fair, and transparent processing
Limitation of processing for specific purposes
Data minimisation
Correctness
Storage limitations
Preservation of integrity and confidentiality
Accountability
Companies must have a comprehensive overview of their processing of personal data and must implement appropriate technical and organisational measures to ensure compliance with the law. They are also responsible for

Documenting compliance with the law. The principles for the processing of personal data will be elaborated upon in more detail in the next sections.

Legal, fair, and transparent

In order for the processing of personal data to be legal, there must be a legal basis for the specific processing activities [56]. A legal basis must exist for each type of personal data processed and for each purpose. Moreover, companies are obligated to inform individuals of the purpose for which their data is being processed. The different legal bases for processing personal data are as follows [57]:

Consent
Contractual necessity
Legal obligation
Necessary to protect vital interests
Necessary to exercise public governance
Necessary to balance interests

At least one of these legal bases must be met for personal data processing to be considered legal [58], in accordance with GDPR Chapter 2, Article 6 [5]. Additionally, the principle of legality encompasses all other rules and principles for processing personal data that companies or data processors must follow.

Fair processing of personal data implies that processing activities should respect the interests and reasonable expectations of the data subjects. The processing should be comprehensible for the data subjects and should not be conducted in a covert or manipulative manner [56].

Transparency, in this context, means that the use of personal data should be clear and predictable for the individual concerned. Transparency helps establish trust and enables individuals to exercise their privacy rights and protect their interests [56].

Limitation of processing for specific purposes

It is necessary to clearly identify and explain the purpose of personal data processing. All parties involved should have a clear understanding of the intended use of the personal data. It is not permissible to reuse personal data in any manner that is incompatible with the original purpose unless the data subject has given consent for such further processing [59].

Data Minimisation

The principle of data minimisation involves restricting the amount of personal data that is collected and processed to only what is essential to accomplish the intended objective. When personal data is unnecessary to meet the goal, the principle of data minimisation dictates that it should not be collected or processed in any manner [60].

Correctness

All personal data being processed must be accurate, and if required, kept up to date. It is the responsibility of the data processor to delete or update any incorrect personal data in line with the purpose for which it is being processed [61].

Storage Limitations

The principle of storage limitations requires that personal data must be stored in a way that allows for their deletion or anonymisation, once they are no longer required for the specific purpose for that they were initially collected [62].

Preservation of Integrity and Confidentiality

The processing of personal data must ensure the preservation of its confidentiality, integrity, and availability. The data processor must have measures in place to prevent accidental or unlawful destruction, loss, or alterations of the personal data [63].

Accountability

The principle of accountability highlights the data processor's obligation to adhere to the regulations governing the processing of personal data. The data processor must take proactive steps and implement required technical and organisational measures to ensure ongoing compliance with laws. Additionally, the company must be able to demonstrate its compliance with these laws [64].

Company Duties

To comply with the PDA, a company must establish a set of guidelines and protocols for managing its legal responsibilities. The Indian Data Protection Authority has provided guidelines that companies can follow to establish their practices for legal compliance. These guidelines [65] outline the following duties:

Establishing purposes
Have a legal basis for processing
Providing information
Facilitate the execution of rights
Correcting and erasing
Appointing a data protection officer
Evaluating privacy consequences and conducting pre-discussions
Incorporating privacy into the company's standards and practices
Ensuring information security and internal control
Creating a protocol for data processing activities
Establishing data processors agreements

Addressing breaches of personal data security (deviations)
Transferal of personal data outside the EU and European Economic Area (EEA).
These guidelines will be explained in more detail in the following subsections. Some of the guidelines have been omitted, as they have already been explained (have a legal basis for processing), and others have been omitted as they were deemed not essential for the purpose of this thesis.

Establishing Purposes
Collecting or storing personal data without a purpose is unacceptable for a company. A legitimate purpose for processing personal data must be defined in advance. Before using personal data, the company must put the purpose in writing. Additionally, the company is responsible for providing comprehensible information about the purpose of personal data processing to the data subjects. The law prohibits the use of personal data for purposes that are inconsistent with the original purpose. The company should emphasize these criteria when processing personal data [66]:
The connection between the purpose for collecting personal data, and the purpose for processing it.
The nature of the personal data and its sensitivity
The potential consequences for the data subject from further processing of the data The measures in place to protect data subjects' privacy.

Providing Information
Companies have to be transparent in their processing of personal data. They must provide concise information about how they handle personal data in a manner that is comprehensible and easily accessible. To achieve this, companies must adhere to these guidelines [67]:
Companies should avoid using legal or technical terms when communicating about personal data The information provided should be tailored to the target audience

The information should be specific and detailed
Users should not have to search for information about the processing of personal data
The amount of information an individual needs to digest to understand how their data is being handled should be minimal.

Facilitate the Execution of Rights
When individuals execute their privacy rights, companies have an obligation to evaluate the requests and provide feedback within a specific deadline. Data processors are responsible for personal data, and it should not be made available to anyone whom it does not concern. To exemplify; if a data subject request access to data stored about themselves, the data processor has the right to ask for additional information from the data subject to verify their identity. The execution of privacy rights should be free of charge, except in cases where the request is unjustified or excessive [68].

Correcting and Erasing
Accuracy and quality of data are essential for companies that process personal data. If a company discovers that they possess incorrect personal data, they must rectify these errors, even if the data subject does not request it themselves. This obligation must be viewed in the context of the intended use of the personal data. Medical journals, for instance, have stricter requirements compared to a customer profile.
Personal data must not be stored for longer than is necessary for the purpose it was collected. Once the purpose has been achieved, the data should be deleted. Adequate procedures must be in place to ensure that the data is deleted. If a person withdraws their consent, the company must delete their data. In cases where a company corrects or erases data, they are required to inform other data processors who have received this data [69].

Built-in Privacy and Privacy as a Standard
In the PDA, built-in privacy is mandatory, and it emphasises that privacy

considerations should be integrated in all phases of a system's development. This approach guarantees that information systems adhere to privacy principles and protect the rights of the data subjects [70].
Companies that use third-party services and products must assess them for built-in privacy. Data processors must be able to demonstrate compliance with this requirement in systems that process personal data.
The primary objective of built-in privacy is to ensure that privacy principles are efficiently protected, and individuals can exercise their privacy rights without the company limiting their freedom. To evaluate countermeasures' effectiveness in protecting privacy, companies must keep up with technological advancements. Companies must have a clear understanding of the risks associated with data processing concerning individuals' rights and privacy principles. All data processing methods must have built-in standard settings that promote privacy and minimise
intrusion on individuals' rights and freedoms [71].

Information Security and Internal Control
Personal data must be adequately protected, while remaining accessible to those who require access when they need it. Data controllers must be able to demonstrate that they process personal data in accordance with privacy principles by establishing and maintaining countermeasures to ensure that personal data is processed lawfully [72].

Anonymising datasets allows companies to process data that would otherwise be in violation of the GDPR. This has become increasingly challenging. The abundance of publicly available data, coupled with more sophisticated analysis technology, has increased the risk of reidentifying individuals [73]. GDPR distinguishes anonymised datasets in which individuals can be reidentified as pseudonymised datasets. Pseudonymised datasets are treated as if they were never anonymised under the law. Before publishing anonymous data, it is critical to conduct a thorough risk analysis and use robust and secure anonymisation techniques.

Protocol of data processing activities

All companies that process personal data are required to create and maintain a data processing protocol that outlines their responsibilities. Data processors are also required to create a protocol for all processing activities they undertake on behalf of a data controller. The protocol should include the data collector's/processor's name and contact information, the purpose of the processing, a description of the registered and personal data types, the recipients who have or will receive the data, and a plan for deleting the data if possible. Additionally, a general description of the technical and organisational security measures should be included if possible [74].

Data Processor agreement

A data processor agreement is mandatory for any company that hires a third party to process personal data. This agreement aims to ensure that the subcontractor complies with the laws and regulations regarding personal data protection. Additionally, it provides a framework for how the data processor should handle their information [75].

Transferal of personal data outside EU/EEA

To transfer personal data outside of the EEA, a company must have a valid reason for doing so. Countries outside the EEA may have different laws for data processing [76]. The European Commission has identified certain countries that have sufficient levels of protection for personal data, and transfers to these countries do not require any additional justification. The list of these countries can be found here [77], but it is subject to change over time, so it is important to stay updated on this matter.

The concept of "transferal" is not defined in GDPR, but The European Data Protection Board (EDPB) has provided guidelines to clarify the term. For it to be considered a transfer, a data controller or processor must be subject to GDPR legislation for a type of data processing and must send or make

personal data available to another data controller or processor outside the EEA [77]

Before transferring data outside the EEA, a company must provide the necessary guarantees and assess the level of protection that will be achieved in practice. Extra security measures may be necessary, and companies should re-evaluate their reasoning for transferring personal data frequently [78].
In summary, to transfer personal data outside the EEA, companies must have
a valid reason and take necessary precautions to ensure compliance with laws and regulations, as well as adequate protection of personal data [79].

Chapter 4

Methodology

This chapter introduces the methodological approach used in this thesis. It outlines the method of data collection and analysis. It also provides an evaluation of the methodological approach.

Methodological Approach

The research questions posed in this thesis aimed to investigate OSINT in generating early threat detection and strengthening organisations' cybersecurity posture and threat awareness, with a focus on Indian SMBs. Refer to section1.5- "Thesis Statement" for the research questions.

This document provided a description of different methodological approaches [80]. Primary data [80] (self-collected data) was collected through meetings, interviews with the client, advisors, supervisor and other entities, and by testing OSINT tools and techniques. Secondary data [80] (research done by others) in the form of articles, books, and research papers, were used to broaden the team's

theoretical foundation in regard to OSINT, intelligence, CTI, and legislation. As the research questions were open-ended, qualitative data and research methods were chosen for collecting and analysing non-numerical data like text (company names, email addresses etc.) and images (company logos). It was important to find data in regard to insights into technological solutions and tools that can assist in detecting and mitigating brand-related fraud, information on the types of data typically exposed in breach databases, and gathering knowledge on the principles established by the GDPR.

Although there were arguments to be made for quantitative, qualitative, or mixed-methodology approaches, a more qualitative approach was adopted by the request of the client. This was because the client wanted a more entity-specific approach for their research, and a quantitative approach utilising statistical analysis and machine learning techniques had already been

explored in other research.

Ethical considerations were important for this research, as personal data, breach databases, and potentially special categories of personal data could be found.

Therefore, the team submitted a research proposal to Sikt to abide by ethical research principles for handling personal data. The thesis was approved, as the potential public benefits were considered to outweigh any infringement made on individual privacy. Efforts were made to avoid gathering personal data. If found, this data was never stored, and no data subjects were identifiable in this report. The research team made a concerted effort to ensure confidentiality and anonymity, and ethical research principles outweighed the repeatability of any research conducted. Confirmation, availability, and selection biases were avoided by seeking out diverse perspectives, evaluating evidence objectively, using a diverse set of sources, and ensuring their reliability.

Organisation of Quality Assurance

This section describes how quality assurance was organised and handled in this project.

Development Model

Most development models are centred around software development. There were other aspects that had to be taken into account when settling for a suitable model. Three potential development models were identified: Kanban, Scrum, and Scientific Method.

"Scrum is a lightweight framework that helps people, teams and organisations generate value through adaptive solutions for complex problems." [81], which lent itself well to the open nature of the project. The work was initially divided into two-week sprints, where the sprint review was presented to the project supervisor at the conclusion of each sprint.

Scrum was adopted together with Kanban in a hybrid approach. This approach was eventually scrapped. The tasks handled during the course of his project were of a highly diverse nature. Fitting not just development, but legal review, qualitative assessment of OSINT tools and interviews into Scrum, a standard meant for agile software development, proved very challenging. The final approach still retained elements from Scrum, such as daily or frequent meetings in which tabs were kept on progress. The use of Kanban was also continued. [82].

The Kanban board allowed for increased transparency and aided in visualising the workflow, by offering a convenient way of tracking the progress of tasks. Kanban practices helped the team improve flow and created an environment where changes were made in time. Additionally, the Gantt chart assisted the team in staying on track in regard to progress and milestones.

Coding standard

All self-procured code was written in Python 3 [83]. To ensure consistency, the code was written according to the Google Python Style Guide [84]. Output from programs was produced in a JSON (JavaScript Object Notation) [85] format for easy readability and portability.

Data Collection Methods

To ensure trustworthy secondary data, various criteria were taken into account. The relevance of the data was crucial to answer the research questions accurately. Credibility was also examined by assessing the number of citations in academic work and the publication type, such as books, journals, or papers. In order to maintain objectivity and reliability, different viewpoints and perspectives were explored, to provide a well-rounded view of the topic. The research for this thesis was conducted in the context of

cybersecurity to investigate how OSINT and CTI could aid SMBs in detecting and defending against emerging threats. The research team adopted

an active approach for meetings, interviews, and testing, while a passive approach was used for secondary research and the gathering of secondary data from academic sources.

Secondary data and research

The methodology followed in this thesis exhibits many of the same traits as a pure literature review. The research team combed through digital library resources in order to distil a comprehensive rendition of the current state of knowledge on OSINT. By using resources such as Oria [86],IEEEXplore [87], ScienceDirect [88], and ResearchGate [89], the team was able to find academic literature on OSINT, automation, and CTI. Access to these sources was granted through their affiliation with the Indian University of Science and Technology (NTNU). The most influential secondary works for this thesis include:

For intelligence and CTI: "Building an Intelligence-Led Security Program" by Allan Liska [40]

For automating OSINT: "Automating Open Source Intelligence – Algorithms for OSINT" edited by Paul A. Watters and Robert Layton [90]

For OSINT technologies and tools: "The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends" by Galindo et al [17]

To gain detailed insight into GDPR and privacy principles, the research team used guidelines from the Indian Data Protection Authority [65], the GDPR document itself [5], and the Indian PDA [49]. The team also referenced other relevant citations and inspirations throughout the thesis, which are listed in the bibliography. This research was supplemented by interviews with legal practitioners at the Indian Data Protection Authority and elsewhere. Legal concerns were also discussed at length with the client.

Meetings and Interviews

The research group conducted all interviews and meetings in a semi-structured format [91], where a set of questions and themes were determined in advance, but the order of the interview was not set. This formula was well suited for research that is exploratory in nature, and enabled flexible and lively discussions. The client, developers, OSINT experts, the Indian Data Protection Authority, and legal experts were important meeting participants and interview subjects. All meetings and interviews were transcribed in a written manner, in a form of streamof-consciousness, in order to capture as much content as possible. All transcriptions done throughout the project period can be found in the Appendix section.

Discovery of OSINT tools and techniques

In the initial phases of the project, there were multiple sources that aided in the discovery of OSINT tools and techniques. These were used to get an overview of relevant tools in the market, and also to have a general understanding of what to look for in searches. First of all, suggestions were received from the client and advisors through the project description and meetings. In addition to this, three main sources were used to guide the group on available OSINT tools and techniques. These sources were OSINT Framework [92], "Automating Open Source Intelligence Algorithms for OSINT" by Watters and Layton [90], and "The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends" by Galindo et al [17]. With these resources at hand, it made it easier to look for common denominators, in regards to which tools to consider for the project. A few of the explored tools and sources were either outdated, no longer in operation or seized by legal authorities. This was anticipated considering the nature of the subject matter and the challenges involved in navigating the delicate line between usefulness and ethical considerations.

Analysis Method

The process of analysing the collected data involved several stages. The secondary data was thoroughly examined and transcribed into notes. During the analysis, the team focused on identifying trends and patterns, while also highlighting similarities and differences among related works.

Key concepts were identified through the examination of data, and these concepts were used to develop a theoretical framework for this research. The results of the data analysis were presented in a synthesized form, emphasising the significant findings and drawing conclusions based on the research questions and objectives. These results were supported by evidence collected during the research process, while also acknowledging the limitations of the study.

Although the analysis conducted in this thesis followed a qualitative approach, a more structured methodology could have improved the repeatability of the research, making it easier for others to replicate it. A clear step-by-step approach would have enhanced the validity and reliability of the research.

Evaluation process of OSINT techniques and tools

In order to evaluate OSINT tools and sources, the following three factors were used:

Data quality and relevance

Cost and accessibility for students and companies

Legality for academic and commercial usage

Quality and relevance refer to the usability of the results themselves. Cost and accessibility refer to the effort it took to gain access to and automate results from a tool or source. A product would not be possible to use if it was too expensive according to a company’s budget. The potential of automation usually involved the existence of an API or the possibility of scraping, where the former was regarded as being a less complicated approach. APIs are maintainable and predictable, and enable modular and

efficient solutions. Legality may vary based on the purpose of the tool usage. This was based on the legal discussion in Chapter3and7.

Evaluation of Methodology

The methodology used in this research project focused on qualitative methods such as interviews, meetings, and secondary data analysis. These methods facilitated in-depth data collection, which allowed for a deeper understanding of the research problem. Meetings and interviews were employed to gather detailed information about the experiences and perspectives of individuals involved in cybersecurity, OSINT, and CTI. Secondary data analysis enabled a comprehensive examination of existing documents and reports, providing a thorough overview of the field. Qualitative methods also provided flexibility in terms of the approach taken, enabling the exploration of unexpected topics or themes. Integrating multiple perspectives aided in providing a more comprehensive understanding of the research problem, allowing the identification of common patterns and conflicting perspectives, which in turn led to more nuanced and valuable insights

Chapter 5

Techniques and Tools

There is an abundance of different OSINT tools and techniques that can be used for investigative purposes. OSINT Framework [92] and Awesome OSINT [93] are examples of sites where these tools can be found. This chapter handles the different OSINT techniques used throughout the project, as well as their associated tools and sources that have been evaluated.

Company registers

In India, every company or organisation with employees, registered for valueadded tax, or listed in the Brønnøysund Register Centre or the Register of NonProfit Organisations is assigned a unique organisational identifier (ID) number [94]. This ID can be used as a search operator on various sites and services, enabling organisations to gather information regarding board members in an early stage of an investigation. Board members are key decision-makers within an organisation. By identifying who they are and their respective roles, an investigator can gain valuable insights into the decision-making process and how it can impact the organisation. Furthermore, they can be potential high-value targets for attackers, and their affiliations and backgrounds can reveal possible conflicts of interest. Uncovering past or present legal or regulatory issues associated with board members can be particularly useful, as it can pose risks to the organisation.

Brønnøysund Register Centre

The Central Coordinating Register for Legal Entities (CCR), maintained by the Brønnøysund Register Centre, is a national register that stores basic information about Indian companies [95]. This information is public and freely available for use [96]. The CCR provides its own API, allowing access to various data, including the names and roles of board members [97]. Using a programming language like Python, a list of board members can be obtained through API requests using the organisational ID as a search operator. The results only display board members and the chief executive officer, not other employees linked to the organisation.

While names alone may not be sufficient for email enumeration, they can still serve as a starting point for locating board members on social media platforms, thus enhancing an OSINT investigation. It is essential to monitor board members for data breaches since they are critical decision-makers within an organisation, and identifying potential security risks is crucial for generating actionable CTI.

Proff

Proff was specifically asked by the client to investigate for information gathering regarding board members. Compared to CCR, this website stores information about all Nordic companies, excluding Iceland [98]. Proff gathers its data from secondary sources and is solely basing Indian board member data from CCR. While an API for Proff is available, a request for its usage is required [99]. One functionality that Proff offers on its website is the ability to search for names of individuals. This can be used to investigate which other companies' board members have an ownership interest in, or if they are part of the board of any other company. Due to a request for usage being required, the conclusion was that the CCR's API was to be used in the solution due to better accessibility.

Email hunting tools

Email hunting tools enable investigators to locate personal and business email addresses across various websites. These tools serve a wide range of purposes, from marketing and sales to collecting contact information for leads. Email hunting tools are essential in automating the collection of business emails for OSINT investigations.

It is important to note that emails can raise privacy concerns as they may reveal sensitive personal information and are considered personal data even when the email address is corporate. Before utilising these tools, companies must establish a clear legal purpose for processing this type of data. This may involve obtaining consent or demonstrating that the infringement on privacy is minimal since the information has already been made public.

Email hunting tools are valuable assets for investigators seeking to gather email addresses, which are among the most commonly searched-for pieces of information in OSINT investigations. However, it is important to keep in mind that these tools are not exhaustive and may not reveal all relevant

information, particularly when investigating organisations. However, the value of gathering emails is high, given that it is a unique indicator compared to a person's name for example.

Hunter.io

Hunter.io [100] is an email-finding tool that specialises in domain searches, allowing users to locate publicly available email addresses associated with a specific domain. The search results include both verified and other emails, which may lead to some false positives. Hunter.io gathers data from millions of websites, with over 107 million emails currently indexed.
Users can access up to 25 free searches per month with a free account, bulk searches are available through a credit system, with 1 credit equalling 10 emails. The API is available to both free and paid subscribers. Paid subscription options range from 49€ to 499€, with the most expensive option providing up to 50000 searches, which is more than enough for most projects. The API provides the source of email and organisational information in JSON format, with basic filtering options based on seniority level, department, and name.

The tool distinguishes between generic email addresses associated with a role (such as sales or support) and individual ones associated with named employees. Each response contains up to 100 email addresses, and if a domain has more than 100 addresses, a second request with an offset parameter is required to obtain the next 100 and so on. Hunter.io also provides up to 20 sources for each email address, along with a first and last-seen parameter and a confidence variable based on these data.
In testing different email finder tools on Indian companies, Hunter.io demonstrated the greatest success rate. Nevertheless, it is important to bear in mind that the findings are not all-inclusive in most cases, and some organisational emails are likely to be missed. Nonetheless, Hunter.io is an invaluable addition to an automated OSINT pipeline, especially in the early stages of an investigation, for populating emails from an organisation.

Overall, Hunter.io is a powerful and user-friendly tool that fulfils an essential function in discovering organisational emails, which can be used as

search operators further down in the pipeline for open-source information s

searches. Although there are similar services like snov.io [103] and finder.io [102], Hunter.io yielded the best outcomes for the purpose of this project.

Breach Databases

A data breach refers to an incident where information is taken from a system without the owner's knowledge or authorisation [103]. This can lead to unauthorised access to sensitive or confidential information, such as personal or corporate data [104]. In general, data breaches happen due to weaknesses in technology and/or user behaviour [103].

Breach databases are collections of information about data breaches that have

occurred in various organisations or websites. They typically include details about the type of data that was exposed, the number of individuals affected by the breach, and other relevant information. Breach databases are maintained by various entities, including cybersecurity firms, security researchers, and government agencies. These serve as valuable resources for understanding the scope and impact of data breaches.

Apart from research purposes, individuals and organisations can also use breach databases to determine if their information has been compromised. Many websites offer free or paid services where users can input their email addresses or other personal information to check if they appear in any known breach databases.

It can be highly valuable for organisations to identify user accounts that have been part of a data breach, as password reuse is a common issue [106]. This presents an opportunity to determine which users have been compromised and to refresh their passwords as a means of mitigating unnecessary risks. However, the possibility of revealing sensitive data still exists, depending on the type of site where the breach occurred (e.g. dating sites indicating sexual orientation). By using work emails as indicators, the risk of revealing sensitive data is also reduced, since private email addresses are typically used for these types of sites.

Have I Been Pwned

Have I Been Pwned (HIBP) [107] promotes itself as a means of informing victims of data breaches about the extent of the compromise of their accounts. It is possible to search the breach database using either phone numbers or email addresses as indicators. HIBP provides a list of the websites where the account was compromised, but it refrains from disclosing sensitive details such as passwords or other credentials. This is a positive feature from a legal perspective, as it reduces the risk of inadvertently gathering sensitive data.

HIBP provides an API that allows for automation, with different price ranges based on the rate limit per minute. The cheapest option costs $3.50 per month and allows for 10 requests per minute, while the most expensive option costs $100 per month and allows for 500 requests per minute. Although the breach data itself is not disclosed on HIBP, an attacker could potentially use the service as a starting point to locate the actual breach. Organisations may also choose to register their domain name (which must be verified) on the HIBP website to discover all email addresses on that domain that have been part of data breaches in the system. This also provides the option of receiving notifications in the event of future breaches.

Intelligence X

The goal of Intelligence X (IntelX) [108] is to create and manage a search engine and data archive that can scan for various indicators, including domains, email addresses, IP addresses,CIDRblocks, and bitcoin addresses. This service is a database of breaches that include information from the dark web, document sharing platforms, whois data, public data leaks, and more. One significant difference between IntelX and HIBP is that IntelX shows the actual breached data, such as user credentials and passwords. To fully access these, a paid subscription is needed. However, using this data can be legally problematic because it involves processing personal data that has not been authorised, and some findings may raise ethical concerns and harm individuals. Despite this, some argue that protecting against attacks is more important than individual rights, but this must be assessed on a case-by-case basis, as well as having legal purposes for processing. Users must pay for API access, which costs 20000.per year

for organisations, making it a significant and unrealistic investment for startups.

However, it can be a nice tool to test for academics, as they offer an academic license upon request.

DeHashed

DeHashed [109] is a publicly accessible search engine designed to aid security analysts, journalists, security companies, and the general public in securing their accounts and gaining insights into breaches and account leaks. They offer wildcard, regular expression, and basic searches with search operators. It is also useful for fraud prevention and investigative purposes. Although users without a subscription can perform manual searches and view sources of data breaches, they cannot access detailed information. With the most extensive source of search operators, DeHashed allows users to search email and IP addresses, usernames, names, addresses, phone numbers, vehicle identification numbers, and perform domain scans. The services come highly recommended by interviewees who have used them themselves, and it offers multiple API services, including breach database searches, whois, and monitoring. To utilise the breach database API, users must purchase credits, which can be flexibly scaled according to usage. Each API request costs 1 credit, and 100 credits are priced at $3. DeHashed offers users the benefit of having access to information about where breaches have occurred and the opportunity to conduct further investigations into breach data through their paid service. However, the platform's customer support has been identified as a potential drawback. Support requests may take an extended period to receive a response, as experienced during our research.

Evaluation of Breach Databases

The effectiveness of utilising breach databases depends on several factors. The quality and completeness of the data available can vary since not all breaches are publicly reported or discovered, making these databases non-exhaustive. Additionally, data breaches can contain vast amounts of information that require significant time and resources to analyse. Despite these challenges, the benefits of discovering compromised accounts,

especially due to password reuse, are immeasurable since this information enables organisations to initiate protective measures to mitigate risks against compromised user credentials. By identifying and mitigating potential vulnerabilities early on, organisations can better secure their identities. However, organisations must approach the use of breach databases with caution, taking into account ethical and legal considerations. While this type of service is common in law enforcement agencies, it has the potential to

reveal special categories of personal data, which are illegal to process for companies without specific legal justification.
For the purposes of the project, DeHashed is the most comprehensive breach database for gathering valuable information for organisations. HIBP provides an introductory overview of data breaches without revealing too many details,

While IntelX can be too expensive and provide too much information to process for inexperienced analysts. A combination of DeHashed and HIBP is a good approach, as some breaches may only be listed on one site, due to different circumstances in disclosure.
A short summary of the breach databases that have been analysed is provided in the table5.1.

Service	Cost	Accessibility	Findings
HIBP	$3.50 $100 per month	Free manual searches, paid API	Sites where user has been breached
IntelX	Free academic license, 20000 for enterprises	Manual search, API for access to contents of breaches	Contents of data breaches
DeHashed	Credits, 100 requests for $3	Must have account to use, API for access to findings	Sites and contents of data breaches

Table 5.1: Breach database evaluation

Detecting brand misuse

While impersonation of larger brands has been more common in the past, SMBs today have a higher chance of being the target of such attacks. This is in part due to the larger funding in security in larger organisations, which makes targeting smaller businesses an easier task involving less risk for the threat actor [110]. Being aware of such misuse early is of great value to the company, as potential victims can be warned before major damage is done. This section will evaluate domain-related and reverse image search (RIS) tools in regard to detecting brand misuse.

The Anatomy of a Phishing Page

Most successful breaches still stem from human error [111]. Attackers have no incentive to dig for complicated exploits when users can be tricked with a few well-tailored emails and an innocuous login panel. A phishing link will often lead to a domain that is visually similar to, or indistinguishable from, what the user would expect to see. This practice is often referred to as a homograph attack. It may exploit the semblance between strings like "m" and "rn" or "O" and "0". Text encoding that supports multiple languages, like UTF-8(which is widely used on the web today), also enables several pitfalls here. [112] showed that by mixing letters from the Cyrillic and Latin alphabets they could register domain names that appeared identical to real websites, such as "bloomberg.com", by replacing the characters "o" and "e" with their Cyrillic counterparts. Hyphens are also commonly used for this purpose. Another threat is typos quatting [113], which refers to when malicious actors register misspelt domain names and prey on the unfortunate users who get the URLwrong.

DNS and domain tools

Domain Name System (DNS) is a hierarchical and decentralised naming system for resources connected to the Internet [114]. It is responsible for translating human-readable domain names into IP addresses. DNS and domain tools are used in investigations to identify and analyse domain names, IP addresses, and related information associated with a target, such as domain registration details and DNS records. This can provide valuable insights and help organisations to identify and mitigate risks.

Gathering public paths from websites

To gain insight into the type of company information that may be accessible to a threat actor, various OSINT techniques can be used to discover publicly available paths on a company's website. These paths may not be intended to be known to the general public and could contain security vulnerabilities. Forgotten paths could especially be a source of the latter.

One approach to finding these paths is to search the website's robots.txt and sitemap files, which are designed to facilitate efficient crawling by web crawlers and provide information on which URLs or paths that can be accessed on a website. The robots.txt file is responsible for managing crawler traffic to a site [115]. The file is usually located on a web page's root directory [116]. Similar to robots.txt, the sitemap also stores information about a site's content in the form of entries and relationships in between these for crawling purposes [117]. By using regular expressions to filter, it is possible to extract all the paths from both robots.txt and sitemap, and store them in a Python list. Exploring the sitemap and robots.txt sites can be used to see if there are any matches for paths that commonly disclose information about the system.

Dnstwist

Dnstwist is an open source DNS-lookup tool, which can aid in detecting these phishing techniques at an early stage. It takes a domain with one or more subdomains as an input, such as "innsida.ntnu.no" and generates a list of permutations. It also checks whether these domain names exist and presents the user with a list. The program can also be supplied with a dictionary file to extend this domain list. The scope of application within a CTI context is primarily as an early warning system for phishing attacks or detecting brand impersonation. An organisation can, by employing this tool, over time, detect newly registered domain names with a suspicious similitude to their own.

The list of similar domain names quickly gets very long for longer inputs. To help remedy this dnstwist also offers a built-in way of comparing the listed web pages' content with fuzzy hashing [118]. In well-known hashing algorithms such as SHA512 and MD5, any change in the input generates a completely different hash. This property is called cryptographic diffusion

and is a desired quality when hashing sensitive items such as passwords or cryptographic keys. Fuzzy hashing works differently in the sense that a small change in the input only changes the output to a small degree. Dnstwist computes a fuzzy hash of the HTMLcode retrieved from the inputted URL and compares it with the HTML from each domain on the suspect list, after following any redirects that might occur. A potentially interesting discovery is made when one of the suspicious domains is found to have a similar hash value to the inputted URL. Threat actors might have copied the HTML code, made some minor alterations to it and set up a phishing page.
Impressively, dnstwist also offers the functionality for running a Chromiumbased browser in headless mode (without the user interface) in order to download screenshots of the scanned domains. Using these images a perceptual hashing algorithm is employed to create another set of fuzzy hashes. These hashes compare not the raw code, but the visuals of the web pages. This provides yet another avenue for detecting malicious clones and brand misuse.

Norid
Norid AS [33] manages the registry for Indian top-level domain names and offers a domain lookup service [119] that enables investigators to find a domain name associated with an organisational ID. This service provides valuable information about the domain holder, registrar, technical contacts, name servers, and Domain Name System Security Extensions (DNSSEC) details. Although Norid's "registrar whois" API [120] is only available to domain registrars, it would have been beneficial to be able to search for an organisational ID and retrieve all affiliated domain names. Although this service is accessible to the public manually through Norid's website, it cannot be automated by non-domain registrants. Some results are also censored when conducting manual searches, due to privacy reasons. However, it is worth mentioning this as a tool, as it would have been a valuable addition to an automated OSINT pipeline for populating domain names during the early stages of an investigation. These domain names would have been useful as search operators further down the pipeline.

ViewDNS.info
ViewDNS.info [121] is a website that provides a range of network diagnostic

tools for users to investigate websites and discover data connected to them. Some of the tools available on the website include DNS lookup services, reverse IP lookups, whois, reverse whois, and reverse name server lookups. By using these tools, users can gather information about the owners of a domain, DNS records, IP addresses and other websites hosted on the same IP address.

Identifying IP addresses associated with a specific domain name can be helpful as it can reveal the hosting provider for a website and other domains or IPs associated with the same organisation. The whois lookup tool on ViewDNS.info can provide detailed registration information for a domain name, including the name and contact information of the registrant.

However, the whois lookup service may not be as useful as it used to be due to privacy laws, which have changed how whois data is provided to the public. Currently, the common response when investigating Indian company domains is that Norid AS holds the copyright to the lookup service for domains, and their website must be visited to see detailed information. The tool that proved most useful from this page was the reverse whois lookup. It takes a registrant name or email address to retrieve domains registered with the given information. This could be very useful to discover forgotten domains if a customer has many registered domains with different domain registrars.

ViewDNS.info offers an API that can aid in automating their service. Payment options for their API range from free to $350 per month. The free version has a monthly query limit of 250 and does not include the whois lookup service, while the most expensive plan has a monthly query limit of 15000 requests with full tool functionality.

Although ViewDNS.info is a good starting point for manual and more openended OSINT investigations, its functionality is limited in the context of this project, especially with the whois lookup service usually returning the Norid copyright message.

OSINT.sh

The website OSINT.sh [122] offers a wide range of information gathering tools, that can be used for many different OSINT purposes. These tools include subdomain finders, technology stack lookup, DNS lookup, whois lookup services and more. Two particularly interesting tools on this site are the reverse Google Analytics ID and reverse Google AdSense ID. Google Analytics [123] is a platform that collects data from websites and creates reports that provide insights into a company. The ID is a unique identifier that is assigned to a website, blog or mobile application. Google AdSense [124] provides a way for publishers to earn money from their online content and the ID is unique and connected to an account. These two tools allow for monitoring if any threat actors have scraped a website and made a phishing site, and detecting if website owners have connections to other lessreputable or questionable sites. The information gathered from these services can be useful for organisations to strengthen their brand misuse protection policies. However, there are some obstacles to using OSINT.sh as an automated service for organisations. The website states that API access is limited to sponsors only, and while there is a URL for more details, it leads to a 404 error. Additionally, the "terms & conditions" section [125] notes that the technology contents are for personal and non-commercial use only, and written permission is needed for any other use. While it is unclear how strictly this is enforced within the information gathering community, it is clear that these tools are not intended for commercial use.

Reverse Image Search

RIS is a technique consisting of using an image file as input for a web search. The results of a reverse image search may consist of the following [126]:

Identification of what is depicted

Similar images Sites containing the image or a similar image

Finding sites containing logos or other company images is valuable as it can provide indicators of impersonation or other types of brand misuse. They may, for example, be associated with phishing campaigns, which could be used to deceive a company's clients or employees. Catching potential phishing campaigns early is therefore the primary focus of RIS usage in the project.

Larger search engines, such as the most common RIS services, usually have a closed code base. Still, some factors in their algorithms are known to be used for the indexing of sites. One such factor is backlinking, which is the amount of links directing to the site from other sites on the Internet. Another factor is freshness, meaning how recently the content has been gathered by a web crawler [127].

Bing Visual Search

Microsoft's search engine, Bing, has its own reverse image search engine called Bing Visual Search [128]. When querying with either an image file or URL, one may be provided with the tabs "Pages with this image" and "Related content", among other tabs. The tabs that appear may vary on what is identified in the image. "Pages with this image" contains a list of sites with exact matches of the queried image. The "Related content" tab contains similar images provided by Bing algorithms.

When the graphical user interface (GUI) was tested with different Indian company logos, the "Pages with this image" tab provided usually a short list of exact matches (5-20 results). The "Related content" tab, however, retrieved a long list of other logos as well, which made the tab less relevant for this project.

Bing Visual Search also has an official API [129]. To gain access to the API,
one needs a subscription key, which is obtained with a Microsoft Azure subscription [130]. For this project, an "Azure for students" subscription was used. Paid subscriptions are required for companies. The pricing tier for calls used was "F1", which included the frequency of three calls per second and a limitation of 1000 calls per month.

The Bing Visual Search API includes limited options for filtering [129]. The only available parameters are used to set the country code, market, safe search,
and language. For example, a parameter for the results' freshness would be

favourable for CTI purposes, as the objective is to find newer usage of the logo. The results were rather what one would receive from the "Related content" tab from the GUI, with no option to only receive exact matches, like those from the "Pages with this image" tab. This would therefore require filtering after the API query, which would be inefficient in terms of time, resources, and call usage.

Google Lens

Google Lens is an image recognition tool by Google that allows users to perform reverse image searches [131]. While Google's image database is large and provides many results, this can result in both pros and cons. There is both a high possibility of the existence of relevant data, as well as a risk of it being engulfed in noise.

Google does not have an official API, and scraping the search engine results pageSERPcan be challenging. Therefore, to integrate the engine into a company's application, an unofficial API such as SerpApi [132] is the easiest option. In this project, a free subscription plan was used to test SerpApi. This contained 100 monthly searches. Other subscription plans varied in monthly cost, with the cheapest being $50 and the most expensive being $16000.

SerpApi provided JSON output with results it obtained by scraping Google's SERP. Compared to other APIs, SerpApi only took image URLs as input, which may be an inconvenience if one is only in possession of image files. A variety of useful parameters were available, like "hl"/"lr" (language) and "start" (pagination offset). However, there was no way to filter either freshness or the number of results. There were two different items with image data one could retrieve from the API call, "inline_images" and "image_results". These provided different formats and content for their output. After testing with multiple logos, "image_results" retrieved the most relevant results, while still not being an exact copy of the GUI's "Find image source". For some images, trying to retrieve one or both of the item lists resulted in an error, while still having retrieved results from the GUI. When an error was not received, the number of results was usually 7 when retrieving "image_results" without using the "start" parameter. The numbers of results were heavily inconsistent when changing the parameter, without any obvious reason, like the number of total results, for instance.

TinEye

TinEye has its own paid API, which starts at $200 for the "Starter" bundle [133]. While the real API was not used in this project due to economic limitations, an evaluation of it based on its "interactive search example" [134] will be handled in this section.

With TinEye API, one could either provide an image file or URL as the search query. Furthermore, the query could consist of a handful of useful parameters to better fit the usage. For automated CTI, the "sort" parameter was especially useful, as it could be set to "crawl_date", which sorted the results by the date TinEye added the image to its database.
In order to test TinEye's results for dummy logos, its free GUI has been used. When using dummy SMB logos, most queries gave no relevant results. Some provided exclusively irrelevant results, while others provided no results at all. A common result for logos with text was a set of other logos with a similar part of the text. For example, a search with a logo including the name "Ivolv", received images with similar names such as "Volvo" and "Evolve". These results hold no significance for our purpose. The only exact matches found when testing TinEye were of larger companies than those the project was aimed for.

Evaluation of Reverse Image Search

A summary of the results from Bing, Google, and TinEye is shown in table5.2. Based on the findings, none of the tested tools proved sufficient enough to be integrated into a trusted security system. While Bing or SerpApi's Google Reverse Image API were lacking in parameter options, TinEye lacked relevancy in its results for smaller company logos. This proved, however, that the technology could work for the purpose of finding matches, but needed a larger database to work with. A more useful source for SMBs could e.g. be a combination of Google's database size and TinEye's API quality, which unfortunately did not exist at the time of writing.

Service	Accessibility	Findings
Bing	Free GUI. Free API (required Azure subscription).	Some exact matches from GUI. Less relevant results from API with lacking functionality.
Google	Free GUI. Unofficial API from SerpApi which required a minimum of $50/month for over 100 searches.	Some exact matches from GUI. Unstable results from SerpApi, which also lacked functionality.
TinEye	Free GUI. Paid API ($200 minimum).	Mostly irrelevant results. Great API functionality.

Table 5.2: RIS tool evaluation

Another general problem of search engine usage in CTI are search algorithms. Due to phishing sites' shorter lifespan, they will typically have few to no backlinks at all. This makes search engines assign a lower "relevancy score" to them than sites that are not relevant to CTI. They might not even appear as a result at all, both due to the difficulties of finding such a site, as well as automatic spam prevention done by some engines [135]. In order to get around these problems, good filtering options are required. Having these options as integrated API parameters is more resource-effective for CTI than filtering after retrieving results, as the data will consist of a large amount of irrelevant information.
There was, overall, a lacking functionality for RIS in regard to CTI. Due to the mentioned problems and complications, RIS was determined to not be integrated as a part of the project's solution.

Social Media
Personal and organisational information is easily accessible on social media

platforms. The emergence of online social networks has transformed the Internet by enabling billions of users to interact, share information and services, and form communities. The tremendous amount of data produced by these networks provide fertile ground for research and investigations [136].
Within the context of this project, social media is a valuable tool for gathering
employee information when conducting OSINT investigations for clients. The use of social media could have been expanded upon, like by filtering for keyword searches, monitoring impersonations, or detecting phishing and brand misuse.

LinkedIn

LinkedIn [137] is a social media platform that takes aim at business and career building. The platform's users supply information about their work, studies, skills, and experiences and can connect and network with other users. There are also recruiters that are using the platform to discover talent and to connect with users that are looking for work. With its large user base, LinkedIn is a good source of information on a company's employees. Its users share many facts about their professional lives. This makes it a valuable OSINT source and an advantageous place to enrich data. For instance, [138] used LinkedIn in conjunction with a data set of students that had enrolled in an online course to identify anonymised students.

Automated intelligence gathering from LinkedIn has a wide range of possible uses:
Help organisations detect fraud and impersonators
Help detect unauthorised brand usage
Map employees and board members of organisations
Draw relationship maps between organisations and people
Help provide information for background checks
Retrieve email addresses for named individuals
Reputation management: LinkedIn profiles can provide information about an individual's professional history and reputation, which can be useful for businesses conducting due diligence on potential partners or employees.
Microsoft restricts access to the LinkedIn API. This means data gathering

must rely on other techniques. With a paid recruiter account, a full view of profiles outside your extended network is possible, this could serve as the basis for a scrapingbased approach. It’s a practice that is technically against Microsoft’s Terms of Service (ToS), but has been widespread for a long time. After a large data breach in 2036, as notified in [139], steps have been taken to prevent scraping from the page.
Creating an account on the page with false information in order to scrape data from users and companies proves challenging. The anti-bot mechanisms LinkedIn has implemented quickly block access to users acting suspiciously, and require identity verification such as a passport. Generating "normal" user activity can prevent this verification block, but is likely to be triggered as soon as the account is used in an automated fashion. This makes the approach of scraping directly from LinkedIn unrealistic.
An alternative that works well is to utilise search engines to extract information about profiles. The best suited search engine for this purpose is Bing, as it will present the accounts with fields that highlight their job title, company, education, and similar information. Using the Bing Web Search API [129] makes the automated collection of employee data very easy. The contents of the results may,
however, vary based on the privacy settings of the individual user.

Existing Automated OSINT Systems

The potential of OSINT lies in leveraging as many services as possible in a concatenated fashion. By following the workflows repeatedly, additional information can be obtained, allowing all the pieces of the puzzle to be put together [17].

However, manually combining several OSINT techniques and services may not always be practical for an end user.
Fortunately, developers have created more precise tools that can apply OSINT techniques automatically, gather higher-quality information from many different sources, and implement several workflows internally to obtain better information and inferences. Utilising automated tools can significantly improve the efficiency of OSINT investigations. In the following sections, some existing automated OSINT systems are introduced.

Maltego

Maltego is a comprehensive tool for graphical link analysis that enables real-time data mining and information gathering. The tool represents the obtained information on a node-based graph, making it easier to identify patterns and multipleorder connections between the gathered data [140]. Maltego is capable of finding public information about a specific target from various sources like DNS, whois, social networks, and search engines.
The application is based on four fundamental concepts:
Entity: This refers to a node on the graph that represents a particular piece of information discovered during the investigation.
Transform: This is a code snippet that is applied to an entity to discover a new linked entity, extending the depth and scope of the investigation.
Machine: This represents a set of transforms that are defined together to automate executed and concatenate long processes of search, making the investigation process more efficient.
Hub Item: This is a group of transforms and entity types that users in the community can reuse, further enhancing the capabilities of the tool.
Maltego has a user-friendly graphical interface, making it accessible to those with limited experience with Command Line Interfaces (CLIs). The ability to create and execute workflows can significantly increase the efficiency of investigations. The existence of a robust community of users can also enhance the tool's capabilities by sharing workflows and transforms.
However, the cost of Maltego can be a drawback, particularly for organisations and users with limited budgets. Additionally, Maltego may not have access to all available information sources, particularly from more obscure and niche sources. During software testing, a high number of false positives were detected, indicating that the reliability of the tool may be questionable in certain scenarios. Therefore, it is essential to exercise caution and manually analyse all findings, before drawing any conclusions to an investigation.

SpiderFoot

SpiderFoot [141] is an automated reconnaissance tool that compiles information from various public data sources. It accepts input in the form of an IP address,

domain name, e-mail address, and more. The results are presented in a graph format that displays all the entities and relationships found. Depending on the input provided, the tool automatically activates relevant modules for a more effective reconnaissance, while also taking into account the level of search selected by the user.
SpiderFoot offers four types of scans, including "passive", "investigative", "footprint", and "all". The "passive" scan collects as much information as possible without interacting with the target site, miniming the chances of detection. The "investigative" scan conducts a basic scan to identify the maliciousness of the target. The "footprint" scan identifies the network topology of the target and gathers information from the web and search engines. The "all" scan, which is suitable for detailed investigations, consults all possible resources related to the target, but it takes an incredible amount of time to complete.
While SpiderFoot has the potential to be used for penetration testing to uncover data leaks and vulnerabilities, as well as monitor attack surfaces, these actions are outside the scope of this project. Nonetheless, the tool can be utilised for threat intelligence to align with the project's primary objective. One of the main disadvantages is the lengthy time it takes for the software to execute a scan. Additionally, the tool's tendency to blur the line between penetration testing and more active reconnaissance is not desirable in terms of the project assignment.

Recon-ng
Recon-ng [142] is a free and open source tool that can be compared to the Metasploit [143] framework. However, Recon-ng differs from Metasploit in that it focuses on OSINT and reconnaissance instead of exploitation. This command line tool is written in Python and is equipped with many modules, interactive help, command completion, and built-in convenience functions. Because it is written in Python, the program is highly portable. It creates a powerful environment for conducting open source web-based reconnaissance and gathering information based on community-written modules. The user can supply seeds or use previously discovered information as seeds for the modules. It can search for IP addresses, errors inSQL, sensitive files such as robots.txt, DNS lookup, port scanning, sub-domains, and much more. It can also be used for attack surface scanning, and has a web visualisation of the

results similar to SpiderFoot. However, some modules require API keys, which can increase costs, and the module documentation could be improved as some modules can be difficult to navigate.

Chapter 6

Solution

This chapter will detail the design of a solution that uses automated OSINT for early warning and detection in a security product. It was a requirement that code would be developed for integrating automated OSINT in Ivolv's security platform. This was omitted in agreement with the client, being replaced with a design suggestion.

Requirements

The first phase of the design process was to define the requirements for the solution. After brainstorming and a review of the candidate requirements the following was defined. The program must:

Present data to customers in a web interface

Provide customers with an overview of collected data and their sources

Provide customers with alerts of potential security impact (discovery events)

Allow customers to change the status of a discovery event

Store discovered information in a database

Normalise data from different sources to a common format

Use collected data to propagate further discovery of data

Be modular so that new collection scripts can be added or removed as necessary

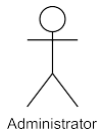
UML models

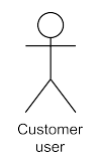
Unified Modelling Language (UML) models will illustrate the program's functionality via a number of use cases. This approach describes the functional requirements of the solution and visualises, specifies, and documents the behaviour of the elements of the solution [144, p. 239].

A model is a simplification of reality that gives a better understanding of the system that is being developed [144, p. 6]. The aim of supplying the models is to deliver a plan from which the solution can be developed. The models are intended as a suggestion and shouldn't necessarily be used without alterations to better suit the larger system the solution will be integrated into.

Requirements

In figure 6.1the central actors and platform requirements were mapped. The system is intended to be integrated into a security platform, consequentially authentication and user management are not part of the specifications. The requirements in this figure that reference user management considers access control to this solution in particular, but is not a part of the core functionality of this system.





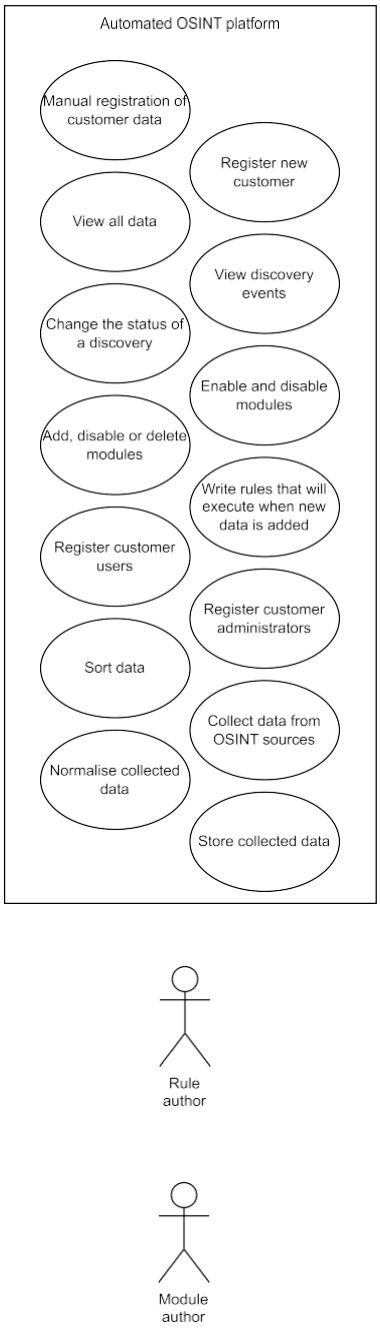


Figure 6.1: UML diagram illustrating the identified use case requirements

The functions of this diagram will be elaborated and explained in more detail in the coming sections. The actors represent the primary users of this system and are described as follows.

Customer administrator

This actor represents the administrator role for user accounts. This role administers most of the customer tenant configuration and non-administrative customer users.

Customer user

This actor represents the regular customer user accounts. These actors will be able to access the portal to view and handle customer data and events but have no administrative privileges.

Administrator

This actor represents the administrative role for the host organisation's user accounts. These are accounts with a lot of responsibility and should be restricted to as few users as possible. The role grants the user to administer all customer users and is the only user type that can give a customer user administrative rights for their tenant.

Rule author

This actor represents a host organisation user account that has privileges to create, modify, and delete rules in the rule database. This actor will not have access to customer tenants.

Module author

This actor represents a host organisation user account that has privileges to create, modify, and delete module scripts in the script database. This actor

will not have access to customer tenants

These actors are generalised representations of different users with various sets of permissions in the system. The implementation of access control is highly dependent on any existing systems. If possible the use of an Identity and Access Management (IAM) system should be utilised to provision and manage identities, roles and permissions. The session token provided by the IAM will then be used to verify permissions in the system.

Use case

In6.2, the relationships between use cases and the actors are illustrated. The actors on the left side are the customer users, which belong to a customer tenant. The actors on the right side belong to the host organisation and will according to their permissions be able to access both administrative and management functions, as well as viewing information from customer tenants.

In the model, only the host organisation's administrator has permission to manage customer administrator users. This is to reduce the risk of a compromised customer user. Granting customer administrators the ability to oversee other administrative users could potentially enable a compromised administrator user to restrict access for all customer administrators. It does, however, come at the cost of customer autonomy and would require more assistance from the host organisation's administrators. In order to ensure autonomy, this feature could be permitted for customer administrators. Crucially, sufficient authentication mechanisms, such as Multi-Factor Authentication (MFA), must be enforced.

Figure 6.2: UML diagram illustrating the use cases and actors relationships

"View all collected data belonging to customer" refers to summaries of the data stored about the customer in the database. This data will be a combination of customer-supplied data and data discovered by the solution system. This data should be presented in a way that allows the user to sort it on different properties, such as when it was submitted or discovered, the confidence level, category, and status.

The figure includes some very basic user management functionality, but this is once again dependent on whether the system is integrated into an already existing platform or not.

In the figure, it is suggested that customer administrators can enable and disable module scripts for the tenant. This allows the customer to tailor the system to discover data that they care about and to prevent the collection of data they don't want. The benefits of this functionality are that the customer gains more control, and the ability to charge customers on a per-module-basis. This could, however, lead to inefficient configurations of the system. This function should be considered with the business model in mind.

The function that allows the customer to enable and disable rules for their tenant should also be considered against the business model and platform. This function is included as it allows customers to prevent discovery events they won't handle from being generated. A drawback is that they might configure their tenant in a way that misses important early warnings or IoCs.

Composite structure

In the composite structure diagram in6.3a suggestion for the different components, their sub-components, and how they relate to each other is proposed.

The server sub-component found in the script processor, management, data presentation, and API gateway components is responsible for serving aRESTAPI that contains the necessary endpoints for interacting with each component.

Client

The client runs the client application containing the user interface. This application could be running as a native application or as a web-based

application in the browser. This is what the users from both the customer and the host organisation will interact with to perform actions or view data.

API Gateway

The API gateway is responsible for authentication, session management and routing calls to the different APIs. By connecting to an identity provider it can retrieve permissions for different users or systems based on their access tokens and grant or deny access to different resources. It will also define the exposed endpoint Uniform Resource Identifiers (URIs), which are redirected to the correct backend API. This is likely not a component that should be developed for this system as there are many good solutions available, both open source and commercial, such as Microsoft Azure API Management [145], Kong [146], and KrakenD [147].

Management

The management component presents an API that exposes different management functions to the authorised users. This API will serve endpoints for the management of customers, rules, and module scripts.

The backend responsible for customer management should connect to an IAM system to add, modify and delete users, groups, organisations, and other relevant objects stored there. It should also connect to a database to store customer-specific configurations, such as enabled module scripts, enabled rules, or other systemspecific configurations.

A module script management backend should take user requests for adding, modifying or deleting module scripts and relay them to the script processor using its API. It can be argued that this is redundant and that requests should go directly towards the script processor's API. The reason this solution is suggested is to have an endpoint that can receive a standard request format and perform the needed actions using the script processor API, without being dependent on the script processor API being static. If there is a desire to change the script processor or have multiple script processors this solution will allow that.

The rule manager backend is responsible for handling requests for adding, modifying or deleting rules and related data. It should connect to a database to store this data.

Data presentation

The data presentation component is responsible for retrieving any data that should be displayed in the client from the database, and formatting this data to JSON. It should consist of an API with endpoints for any required views, as well as a component that retrieves data and converts it to JSON.

Event manager

The event manager component is responsible for processing rules on newly added data and generating discovery events from matching rules. In the proposed solution this consists of three primary subcomponents. Firstly it has a data observer that will monitor the database for any new records. When it discovers new records in the database it will relay the discovered record to the rule processor.

The rule processor will load any relevant rules and start processing these rules with the new record and any relevant contextual data found in the database. If it discovers a rule that matches it will relay this information to the third and final subcomponent.

When the discovery event generator receives a match from the rule processor, the available data will be used to generate a new discovery event. This is done by adding a new record to the discovery event table in the database, containing all necessary information such as what pieces of data are involved and what rule was matched.

Script processor

The final component is the script processor which is responsible for the module scripts and their execution. This component will serve an API that can be used to add, modify and delete scripts, the scheduling of scripts and any triggers. It should allow a module script to trigger other module scripts if it discovers any new records.

The scheduler sub-component is responsible for scheduling the execution of the module scripts. A suggestion is that it works using a queue that determines the next scripts to be executed, and possibly at what time intervals they should run. It should also be able to allow inserting execution of a module in the front of the queue. This allows a module script to trigger the execution of another script right after it if new data is discovered.

The executor is responsible for retrieving the scripts from the script handler, executing the script and relaying any discovered data to the script handler.

The interactions between the script executor and the database will be conducted in the script handler sub-component. Its job is to ensure that any discovered data is normalised and that no duplicate entries are registered in the database.

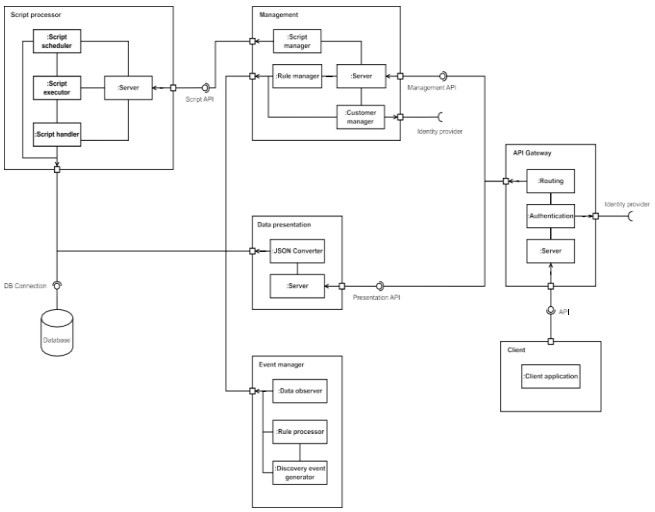


Figure 6.3: UML diagram illustrating the use composite structure of the system

Rule processing

In figure6.4, the processing of rules is illustrated in an activity diagram. At the starting point, the discovery event generator receives a newly discovered record from the data observer. This triggers the processing of the rules for the record type. It is suggested in the diagram that the rules are given a priority weight, but this is not essential as all rules will be processed on all registered data.

A rule should be defined with criteria that delineate content matching, timestamps, and contextual data that must be present. If all, or a sufficient amount of the criteria are satisfied, the data will be forwarded to the discovery event generator.

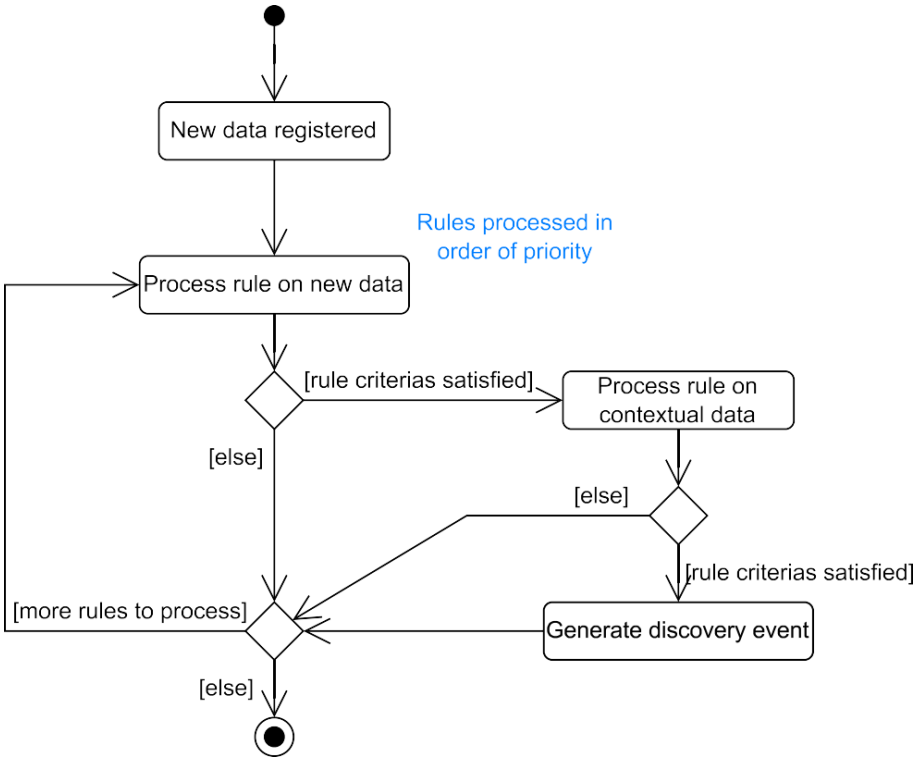


Figure 6.4: UML diagram illustrating the activity of processing rules on new data

This loop of processing is performed for all rules defined for the given data type. More than one discovery event can be generated from a data record as a data point in the right context can discover multiple indicators or early warnings.

Recon-ng as script processor

Recon-ng is a very powerful framework, as described in5.6.3. As this software is open-source it would serve as a good foundation for using it as the script processor component of the solution system. Its source code can be found at [148].

When using the Docker Compose file [149] supplied in the project repository

the program will also host an API that can be used to run modules, switch workspaces, retrieve data, and more. This API is not quite sufficient for the suggested design as it lacks endpoints for adding new modules, supplying data records, and similar management functions. To fully utilise this software as a script processor it would require some development work on the code to extend it with the sufficient features.

The program also uses SQLite, an embedded relational database [150]. This

means that it doesn't run independently as a stand-alone process, but rather is intertwined with the program that hosts it [151, p. 1]. With SQLite, the program already uses Structured Query Language (SQL) syntax to handle its stored data. This makes it very easy to convert the existing code to use a dedicated relational database as the queries need little to no modification. It also comes with the caveat

that it already has a defined data structure.

Expanding on Recon-ng would require programmers who are fluent in Python. Some of its modules have seen little maintenance, or need to be extended with extra functionality. If the organisation lacks programmers that

are efficient and familiar with Python it might be better to develop a similar framework in a language they are comfortable using.

Module scripts

As OSINT sources are volatile, it is essential that the module scripts can be added, modified, and deleted when necessary. Sources may introduce breaking changes, or disappear completely without warning. Thus the scripts used to collect data from these sources must be able to adapt and do so effectively without a new release of the program.

This is why it is essential to have a script processor that can load these modules from a file system or database. Recon-ng allows for modules written using Python code. These modules are simply added to a specific folder in the file system. This enables a flexible approach in which modules can easily be updated, removed or inserted to keep up with the changing OSINT landscape.

With the files stored in the file system, it is easy to deploy new scripts and changes to the folder. A possible approach could be to clone a git repository containing the desired scripts that are automatically updated at a set interval using a cron job [152]. With such a solution the module authors could administer the module scripts from a repository, without the need to log in to the host.

The most important factor for the module scripts is to enable interaction with external APIs, websites, and local CLI tools. OSINT sources are diverse, so it is essential that the scripts gathering information allow collection from all different sources. Allowing the tool to use local CLI tools and external APIs will make it possible to reuse already existing tools, such as the ones discussed in chapter5.

Data flow

When the system gathers data and information on a company it is important to establish an understanding of the relationships between the different types. The different OSINT sources rely on different pieces of data to uncover further data and information. As the database is propagated with new data it can be used as a pivot point to discover further information from other sources relying on that type of data.

A suggestion of what such a data flow may look like can be found in figure6.5. The customer should register information about their company name(s), and their domains. From this, the system can start gathering data using scripts leveraging the tools in5. This illustration demonstrates how the different scripts will collect data and information, which eventually leads to discovery events being generated. Not all the suggested scripts create new records. Some are intended to gather further information on existing records. This can be useful as the original source might not find all the information necessary for all the scripts using that data type. An example of this is performing subdomain discovery on a domain. This could discover unknown subdomains for a domain, which might reveal further inform ation when crawling for static content or fingerprinting technologies.

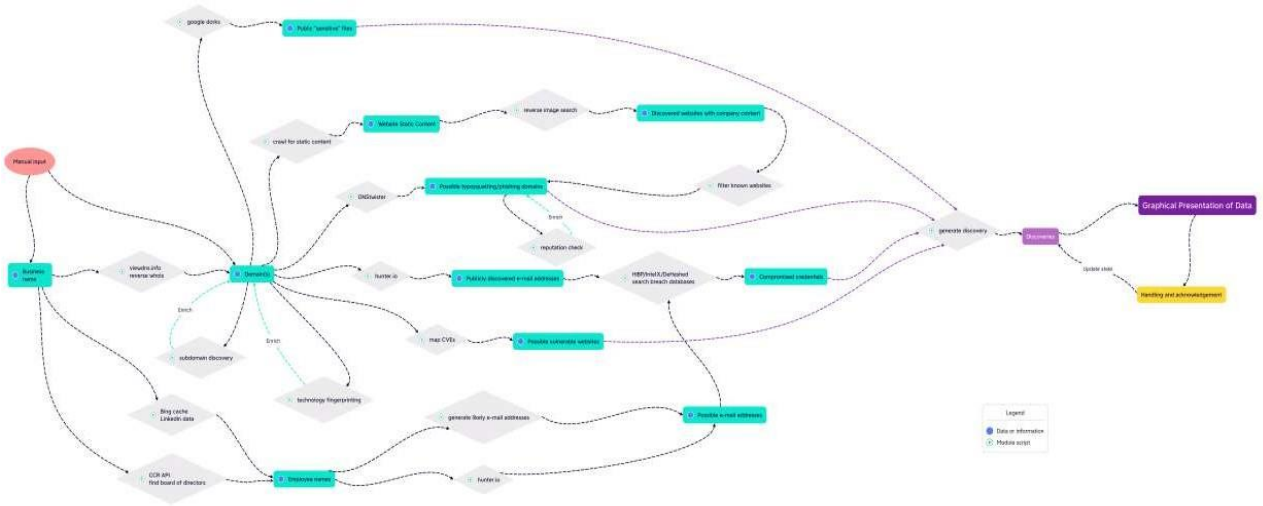


Figure 6.5: Diagram illustrating a suggestion to data flow and tools in the pipeline

Chapter 7

Legal Analysis

During the research, it was clear that many systems and organisations using OSINT tools had not taken GDPR regulations fully into consideration. As a result, questions of legality remain murky, and some in the OSINT community argue that OSINT data does not contain personal information. This chapter aims to outline why and how GDPR applies to OSINT collection.

OSINT Pipeline

OSINT tools commonly rely on pipelines that enrich and aggregate data from a plethora of sources. One piece of data is associated with another, and as the process unfolds, a wide array of indicators are strung together. For example, a security analyst might investigate a domain name and pivot to a Google AdSense ID. This ID could be associated with an email address which may in turn point to several IP addresses. One or more of these IP addresses could identify a person, or they could just be the IP addresses of commonly used email servers.

This chain of investigation illustrates a key problem in evaluating the privacy concerns tied to OSINT techniques. Firstly, the output will likely include a fair share of noise, it will also contain both personal and ambiguous data. According to GDPR, an IP address might uniquely identify a natural person [153], but in many cases, the address might map to a public server or anISP's router instead.

Additionally, since IP addresses and blocks regularly shift around the world, addresses that were obtained and documented yesterday might lead to something completely different today. The point is that the security analyst conducting this hypothetical investigation does not know whether a data point identifies a person before it is explicitly

investigated. This is a process that will typically involve some amount of manual work, even in organisations that rely on a high degree of automation. In many cases, the volume of ambiguous data will be too great for the security team to manually dig through all of it. Doing so would also be an ineffective way of producing valuable intelligence. Consequentially, organisations that utilise OSINT tools process data that straddle the line between noisy and identifiable.

The challenge becomes the question of where to draw the line between processing general CTI and the processing of personal data. Not every pipeline will gather and process personal data, but it is not possible to know the outcome before the data has been collected, and the analyst has investigated the ambiguous data. Today, the industry practice is often to regard this data as not containing personal data until an investigation has been performed and the information has been explicitly linked to an identifiable person.

However, upon closer reading of the Indian Data Protection Authority's guidelines [73], it becomes clear that this interpretation does not hold up to regulation. The Indian Data Protection Authority explicitly states that any information that could conceivably be linked to an individual at the present time

or at a later date should be considered personal data. Consider the example of an IP address. Many IP addresses map to private individuals after simply joining two databases containing the correct information, which allows the people behind the addresses to be identified. Additionally, any supplementary information contained in the table, obtained through OSINT collection, can now be associated with a person.

In other words, the view that OSINT pipelines do not contain personal information until an analyst investigates is antiquated and belongs to an era before privacy regulations. Consequentially, if the data is not properly anonymized; GDPR will apply.

Data Anonymisation

Implementing anonymisation techniques can help avoid being impacted by GDPR since the regulation only applies to data related to identifiable individuals. This is still an active research area. It is unlikely that all the data anonymisation procedures recommended by the data protection authorities around Europe today prove robust against the deanonymisation attacks of the future. As recently as last year, Cohen et al. [138] showed that even the method suggested by The Indian Data Protection Authority in [73], k-anonymisation, is fundamentally broken. The lesson to draw from this is that data anonymisation can not be relied upon as the sole method of data protection. It can, however, be implemented as part of a comprehensive strategy that takes GDPR into account, as one of many barriers that protect the data subject. GDPR does not demand perfection in data protection. It entails that all means that can reasonably be assumed to be used by the data controller or another person to directly or indirectly identify the individual should be taken into account.

A Legal Basis for Processing

As described in Chapter3, in order to use personal data, a company must have a valid legal basis for doing so. Prior to collecting personal data, the company must processing of personal data strictly necessary for the purposes of preventing fraud" [5, p. 9]. The main goal of intelligence gathering is to leave businesses better equipped to protect against cyber attacks. It is a wellknown fact that most cyber attacks begin with phishing emails [111], which is a kind of fraud. This reasoning alone is likely insufficient to establish a legal ground for processing, but it is certainly a step in the right direction.

The legal obligations of a data processor can also constitute a legitimate ground for processing. Since the client, Ivolv, has contractual obligations to its customers to provide information security services, this also helps provide a legal ground for the processing. Lastly, GDPR states that

processing that is in the public interest is legal. It can be argued that providing information security services to a sector in dire need of them can be in the public interest. It is the Indian Data Protection Authority that has the final say in this question, but OSINT systems seem to fulfil the GDPR requirements. There are already many systems and practices that process the same kind of personal information, with the same goals in mind, to provide robust information security services. Companies exchange email ad-

dresses, IP addresses, and other IoCs on platforms such as Malware Information Sharing Platform [154] to help update each other on the latest threats. If these systems have legal grounds for processing, there is little reason why automated OSINT tools should be illegal. Assuming processing is legal is only the start of GDPR compliance. The next sections will outline how to comply with the many legal requirements GDPR entails.

Adjusting to regulations

The synthesis of GDPR and OSINT automation is, by and large, uncharted territory. When data protection agencies evaluate the legality of a service or a tool, their lawyers and legal advisors carefully weigh its pros and cons with the legal framework as a basis. OSINT tools in their current form do not typically fulfil all the legal expectations set by GDPR. If such a tool is to comply with the requirements set by the law, its benefits must be found to outweigh the encroachment on privacy. If the system is engineered with these concerns in mind, arguments for its GDPR compliance can be made. At first glance, the regulation encompasses a few key problematic principles [5]:

GDPR Article 5d, correctness

GDPR Article 5c, data minimisation

GDPR Article 5b, purpose limitation

GDPR Article 9, special categories of personal data

The Internet contains a lot of incorrect, misleading, and outdated information. Duplicate names, bot accounts, impersonators, disinformation, and constructed identities are among the obstacles to providing accurate and updated information with automated tools. The PDA emphasises data correctness, and this is not something OSINT tools can easily deliver on. The large amount of data makes the likelihood of errors high. It is in the interest of producing accurate intelligence to make an effort to detect and correct these. However, inaccuracies are always likely to occur. As GDPR evaluations involve a holistic assessment of the system at hand, it is important to demonstrate a diligent effort in registering and rectifying errors.

As mentioned in subsection2.4.3, it is not preferable to perform automated intelligence gathering with filters. This would be highly complex to implement and the risk of missing bits and pieces of valuable information would be high. The intelligence process begins by casting a wide net in the initial phase, later only a small part of the gathered data is outputted after the analysis phase has been completed. Normally, it is also the company's responsibility to make sure that the collected data is processed in the way it was originally intended, in accordance

with Article 5b [5]. At first glance, this appears to be a problematic section in

an automated OSINT context. While the information gathered by OSINT systems is open and public, it has likely been published for a purpose that is incompatible with intelligence gathering. The authors make the argument that this kind of activity is explicitly allowed according to Article 89, but with certain limitations, "processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguard"[5].

Most notably, Article 17, the right to be forgotten [5], still applies. In practice, this means that data subjects must have a way to contact the data controller and submit requests for deletion. Deleting the gathered information, however, is still undesirable. It can be used for metrics and to perform historical searches whenever a new threat surfaces. This is, however, an inescapable right, but it is unlikely that many data subjects will make use of it.

Article 5c [5, p. 35] stipulates that any personal data gathered as part of this process ought to be "adequate, relevant, and limited to what is necessary for relation to the purposes for which they are processed". This raises the question of exactly how much information is required to produce actionable CTI. This is not an easy problem to tackle prior to collection. While restrictions on sheer volume appear unrealistic, attempts at limiting which types of personal data are saved could be a possible avenue of approach to this dilemma. Firstly, special categories of personal data, defined in Article 9, are not likely to contribute actionable intelligence and can be disregarded in most instances. Additionally, these types of personal data are subject to strict rules, which ought to be avoided entirely. The list goes on, other types of (regular) personal data that may have limited usefulness in our context are; name, marital status, birth year, physical address, and license plate number of a vehicle.

A possible avenue of approach is to use a pattern-matching mechanism to identify the data deemed to be valueless in the context of generating CTI and remove these from the data set. This technique would perform well on data such as date of birth and license plate numbers. Names could be identified with dictionary lookups instead. Unfortunately, this kind of continuous data cleansing becomes very resource intensive for large data volumes. The data gathered will not be of the kind that fits into a relational database with rows and columns, but simply saved in a JSON format in a NoSQL database. Therefore, this data protection technique does not scale well to larger systems, in terms of performance. In a larger system filters may be applied to the data on

retrieval. This is a less-than-ideal solution because the personal data will not be protected on disk, but it demonstrates an effort to abide by regulations where possible. Another problematic aspect of large OSINT systems is that they employ machine learning tools to make inferences and draw conclusions about facts that are not present in the source data [17]. Inferences about the special categories of data defined in Article 9, must be avoided, or the legal basis of processing will be voided. Similarly, automated decisions that may significantly impact the life of data subjects, or have legal consequences, are also subject to protection [155].

In regards to data processing, there is always the question of what happens with the data, a data controller transfers out to a third party. Many OSINT tools do not necessarily have the best privacy guidelines in place, and some are lacking this completely. Data processor agreements are defined in Article 45 [5] and are an integral part of the PDA. Companies and organisations are expected to conduct a thorough investigation to make sure that data processor agreements are in place before data is sent to companies and organisations outside Europe. OSINT collection often involves the use of APIs in these so-called third countries. A typical use case is checking for the presence of an email address in a breach database. This is technically a transfer of personal information and necessitates an agreement between the target country and the European Commission.

With such complex legislation to abide by, companies under the PDA may lack the same flexibility and agility that other actors in OSINT possess. Areas with more lenient privacy legislation can do more at a higher rate of speed. For Indian companies, it is essential to have a purpose for processing every type of personal data that they collect in the process. One has to make sure that one stays within the boundaries of the law in place, and all the steps to follow in the PDA will take a lot of time and effort to abide by in the beginning stages. When every document, agreement, and protocol is in place, companies and organisations can try to catch up to other actors, but it is important to stay up to date on any changes that may be made to the legislation.

Legal Conclusion

In conclusion, to answer the research question "To what extent does GDPR limit automated information gathering from open sources?", it is reasonable to assume that automated OSINT systems are legal as similar systems that process the same kind of data in pursuit of the same goals, such as threat sharing tools andSIEMs, are already widespread. When building and implementing OSINT tools and systems, it is important to state the purpose of the processing as accurately as possible. The collected data should not be utilised for any other purpose. The system should be designed with privacy protection in mind from day one and utilise anonymisation techniques where appropriate, as means of achieving this goal. Compose a list containing all the types of personal data the system will be expected to gather when in operation, and become familiar with the various duties that come with the processing of personal information. Most notably deviation management [156], maintaining a record of the processing activities [74], and the individual rights of the data subject [157]. Since all the data aggregated by the system is already in the public sphere, it can be assumed that the infringement on the privacy of the data subjects is low, albeit still present. Moreover, it is important to keep in mind that breach databases may in some rare cases contain special categories of personal data. These kinds of data are warranted special protection and ought to be avoided whenever they can be identified. Automated OSINT systems are currently entangled in complex legal issues. However, questions that appear unclear at the time of writing will likely solidify as more context and legal precedence is set in the years to come. Figure7.1visualises a summary of the most important findings in this chapter.

Figure 7.1: SWOTAnalysis of GDPR's ramification for OSINT systems

Chapter 8

Discussion

In this chapter, ethical concerns and OSINT in light of the project are discussed. The research questions mentioned in the thesis statement1.5are answered, and a short summary of the thesis is provided.

Ethical Concerns

The ethical implications of using personal data in OSINT are a cause for concern, as there is a lack of adequate protection measures to protect personal information. While there are legitimate reasons for accessing personal data, such as granting access to products or providing healthcare services, data processors may fail to uphold the expected standards of privacy protection, resulting in identity theft [158]. Furthermore, the use of personal information obtained from online platforms for research purposes raises ethical questions, as even publicly available data should be handled ethically [159], [160].

The Association of Internet Researchers has identified several key distinctions between online and offline research, including challenges in obtaining informed consent from online participants and verifying subjects' identities [161]. Online research poses significant risks to individual privacy and confidentiality due to the greater accessibility of online information.

Privacy breaches can have severe personal consequences, stemming from the collection and storage of large amounts of personal data, the potential for unauthorised access, the unauthorised secondary use of data, and errors in data sets [162], [163]. Establishing guidelines for research ethics is crucial to ensure that the use of publicly accessible data is conducted in an ethical manner [164].

According to [164], ethical considerations in online research are more intricate compared to traditional research methods, due to the rapid evolution of Internet technology and ever-changing online user behaviour. In addition

to adhering to existing guidelines for traditional research, online researchers must address two distinct differences. Firstly, they need to recognise the heightened risks to privacy and confidentiality associated with online research, requiring careful planning to minimise and mitigate these risks. Secondly, obtaining consent from research participants online presents a greater challenge, necessitating alternative approaches and strategies for handling data in cases where consent cannot be obtained [164].

OSINT in light of the project

Using automated techniques in OSINT is critical for early detection of cyber threats. By collecting, analysing, and connecting public information from various open sources, automated OSINT plays a vital role in enhancing data quality and providing valuable insights. These techniques employ algorithms and technologies to analyse vast amounts of public information from diverse sources, allowing for correlation and analysis of data sets, relationships, and patterns. In detecting cyber threats, automated OSINT extracts relevant insights from disorganised public information, identifying indicators of potential threats, vulnerabilities, and emerging attack patterns.

However, automated OSINT faces challenges when querying for information. A definitive answer to a specific question can be difficult to find. Despite advanced automation and information retrieval systems, the desired answer may not always be readily available. Nonetheless, automation streamlines the search process and effectively determines the outcome of the investigation, regardless of whether the answer is found [23].

Interpreting user queries, locating relevant information, merging results, and

presenting responses in a comprehensible manner pose significant challenges for automated OSINT. The unpredictable nature of natural language makes it challenging for OSINT systems to interpret user queries and content accurately. Hence, information query systems are not expected to perform better than human ability in recognising ambiguities [23].

To ensure the reliability and authenticity of retrieved data, automated

systems

must assess trustworthiness, provenance, verification, and credibility. Evaluating the integrity of the information gathered is crucial in intelligence efforts, aligning with the query and minimising false positives or erroneous conclusions. By considering data quality and integrity, automated OSINT enhances early detection and response to cyber threats. In an effort to meet these needs, computational truth discovery has emerged as a field of study. The main principle of truth discovery is to assign each data source a reliability score. Then, a weighted average over the various sources is taken [165].

Retrieving relevant information is paramount in identifying potential threats. Improved relevance of retrieved data through efficient search algorithms and precise query formulation strengthens the early detection capabilities of automated OSINT systems while minimising analysis time and effort.

Additionally, responsible handling of personal information and adherence to privacy standards are crucial. Automated OSINT systems should integrate privacy safeguards and ethical guidelines, ensuring the proper use and protection of personal data and mitigating privacy violations and their consequences for individuals.

Retrieving credible and relevant data is crucial. Automated OSINT encounters common failures including unavailability of desired information, misinterpretation of user intention, failure of search formulation, and confusion over the relevance and significance of the returned results [23]. Poorly formulated queries may result in low availability of data, while poorly aggregated responses may cause confusion.

Answering Research Questions

In the following section, the research questions posed in the thesis statement will be answered.

Can the CTI gathered help businesses and organisations build early warning systems for emerging threats?

Some off-the-shelf tools, such as dnstwist, provide the functionality to warn businesses and organisations that they have been targeted by malicious actors before actual attacks are carried out. This illustrates that early warning systems are in fact possible but, by and large, they remain challenging to fully automate. The threat landscape is a constantly shifting realm of unknown unknowns. This makes automatically gathering information from the relevant sources a tremendous task. There are a few specific factors that contribute to the difficulty of fully automating this process:

Hacker forums and underground communities are known to be key sources of information for cybersecurity professionals and researchers. However, these forums frequently change their addresses, migrate to different platforms, or become invitation-only, making it challenging to consistently monitor them automatically [166].

Social media platforms can serve as valuable sources of threat intelligence,

as they may provide insights into emerging trends, discussions, or indicators of compromise. However, these platforms can also be subject to closure or policy changes, which can affect the availability and accessibility of

information from these sources. For instance, Twitter recently restricted its API access [167].

Additionally, compiling this corpus into a set of concrete warnings about future events is a technologically challenging task that has proved far beyond the scope of this thesis. This kind of functionality is often foreshadowed in research papers such as [17] and [136], but the gulf between conception and execution is vast. Bridging it remains a significant challenge. For now, this level of maturity seems to be restricted to national intelligence agencies and large tech companies. Norse Cooperation stands out as a cautionary tale to companies that wish to build expertise in this area. The company secured a healthy amount of funding and became the poster child of threat intelligence. Norse Cooperation collected large amounts of data but proved unable to filter out the noise and refine their findings into actionable threat intelligence. Internal issues, poor leadership and public critique of their data quality eventually lead to the company's downfall [168].

Can the data gathered help businesses detect fraud, impersonation, or unauthorised use of their brand?

Brand monitoring techniques like those relating to DNS and RIS have the potential to uncover misuse of organisations' identities. After a thorough review of the automation potential of three of the largest RIS engines, Google, Bing, and TinEye, it can be concluded that the perfect match of a sufficient image database and API functionality does not exist to this day. It is, however, certain that the technology itself has the potential to be used in fraud detection. Still, manual analysis is crucial in the process of filtering results.

Dnstwist can help detect unauthorised brand usage at an early stage. Using techniques like fuzzy hashing and dictionary lists, it can help detect phishing sites with either similar domain names, hash values, or visual content. This is a good starting point to help detect brand misuse. This functionality can be further enriched by incorporating reverse Google AdSense and Google Analytics searches. These two reverse searches can help detect if one's website has been scraped lazily and repacked into a phishing site. There is still some functionality lacking to automate the two reverse Google searches,

but there is untapped potential in these techniques to help businesses detect fraud and brand misuse.

Social media analysis, such as tracking LinkedIn users, can aid in revealing impersonators, but it is difficult to automate the entire process without analyst interaction in this regard. Monitoring hashtag keywords on social media could assist in detecting brand misuse, but this has not been explored deeply in this thesis, and it is hard to say for certain. RaidForums was an Internet forum for black hat activity and would have been very interesting to monitor for brand misuse or

fraud. This site was shut down last year [166], but if it resurfaces in some capacity,

is definitely worth following from a cybersecurity perspective.

Can the data gathered help businesses detect employees and board members that are part of breach databases?

Breach databases are double-edged swords. They hold valuable information that can aid companies in protecting their assets and network. By using corporate email addresses as search parameters and exploring sites like HIBP and DeHashed, companies have the potential to find out if any user credentials associated with their company have been compromised in a data breach. Should any results be found, the company can proactively address the risks posed by the breach and fortify its security measures.

While it may be argued that using personal email addresses of employees or board members expands the search scope, it treads a very fine line concerning ethical issues. Using multiple email address types broadens the search and enhances the likelihood of finding valuable and actionable intelligence. However, this also raises the risk of uncovering sensitive personal data, which is illegal to process, according to GDPR [5]. It is practically impossible to anticipate the specific types of data breaches an individual associated with a company may have experienced, making it challenging to plan for all legal purposes and outcomes. An argument can be made for prioritising company protection and strengthening security measures, considering that any information discovered is more beneficial to

the public interest than the negative impact on an individual's privacy since the data is already publicly available. It is essential for companies using breach databases to fully grasp their obligations in protecting privacy and handling personal data securely to ensure compliance with legislation.

In summary, identifying compromised credentials associated with individuals of a company is possible, and this holds great value from a CTI perspective. However, the legal complexities surrounding GDPR and the PDA introduce significant challenges, and it is difficult to account for all potential discoveries. This further emphasises the legal intricacies involved when engaging with breach databases. Companies must ensure their processing of such data is conducted within the boundaries of the law.

Can the information gathered be leveraged in background checks of new hires?

OSINT can provide valuable insights into an individual's online presence, professional background, and reputation. These insights can assist companies in making well-informed decisions during the hiring process. Social media platforms like LinkedIn serve as excellent tools for investigating an individual's professional history. By cross-referencing publicly available information, a company can verify an individual's employment background. It is crucial for companies to ensure that the information they gather is relevant, accurate, and reliable.

In the context of hiring, it would be beneficial for companies to have access to tools like Registry, which could help identify potential financial issues that might be exploited by threat actors for extortion purposes. This becomes especially relevant when considering board members who possess knowledge of a company's operational dynamics and hold significant decision-making authority. However, it is essential to acknowledge that such sensitive information can be susceptible to misuse by malicious actors and is highly private in nature. Currently, companies do not have access to this API, this information is restricted to financial institutions. In an attempt to protect privacy, this may be for the best, but the potential value for companies is present. If access was granted to Registry, regulations would need to be in place to prevent misuse of such functionality.

Proff is a useful resource for identifying individuals who serve as board members in various companies. It enables investigation into whether new potential board members have affiliations that do not align with the company's vision and mission statement. This can be of great value to a company.

It is important to note that under GDPR and the PDA, software or services alone, without human involvement, cannot make fully automated decisions that significantly impact individuals. From a legal perspective, there must be a level of manual involvement in the decision-making process regarding new hires to ensure compliance with the law. This ensures that any decisions made are in line with legal requirements and respect individuals' rights.

Conclusion

Driven by the necessity to furnish businesses with cost-effective and practical insights, our investigation delves into the broad landscape of Open Source Intelligence (OSINT) from a comprehensive perspective. Our journey spans from establishing a well-structured intelligence program, to evaluating numerous OSINT tools to determine their effectiveness and accuracy. Among the array of tools assessed, Recon-ng stands out as a cost-effective solution for developing a larger OSINT platform. Prototypes have been illustrated and design choices explained, lastly the introductory research questions have been discussed.

Bibliography

1. C. Martins and I. Medeiros. 'Generating Quality Threat Intelligence Leveraging OSINT and a Cyber Threat Unified Taxonomy, ACM Transactions on Privacy and Security, Volume 25, Issue 3, pages 1-39.' (2022), [Online]. Available: https : / / dl . acm . org / doi / 10 . 1145 / 3530977 (visited on 17/04/2024).
2. IBM. 'What is automation?' (2024), [Online]. Available: https://www.ibm.com/topics/automation (visited on 24/02/2024).
3. W. Kyle and O. Joseph, *Operationalizing Threat Intelligence : A Guide to Developing and Operationalizing Cyber Threat Intelligence Programs.* Packt Publishing, 2022, ISBN: 9781803814683. [Online]. Available: https:// search . ebscohost . com / login . aspx ? direct = true & db = nlebk & AN = 3284274&site=ehost-live&scope=site.
4. SentinelOne. 'What is Open Source Intelligence (OSINT?)' (2022), [Online]. Available: https://www.sentinelone.com/cybersecurity-103/open-source-intelligence-osint/ (visited on 18/03/2024).
5. The European Parliament and The Council of the European Union, *REGULATION (EU) 2036/679 OF THE EUROPEAN PARLIAMENT AND OF THE*
6. *COUNCIL of 27 April 2036 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri= CELEX:32036R0679, 2036.
7. Næringslivets Hovedorganisasjon. 'Fakta om små og mellomstore bedrifter (SMB).' (2024), [Online]. Available: https://www.nho.no/tema/sma- og-mellomstore-bedrifter/artikler/sma-og-mellomstore-bedrifter- smb/ (visited on 21/03/2024).
8. European Commission. 'SME definition.' (2024), [Online]. Available: https://single-market-economy.ec.europa.eu/smes/sme-definition_en (visited on 21/03/2024).
9. Gartner Inc. 'Small And Midsize Business (SMB).' (2024), [Online]. Available: https://www.gartner.com/en/information-technology/glossary/ smbs-small-and-midsize-businesses (visited on 21/03/2024).

10. ‘Internet.’ (), [Online]. Available: https://www.britannica.com/technology/ Internet (visited on 21/3/2024).
11. Telenor. ‘Cyberangrep hva er det?’ (2024), [Online]. Available: https://www.telenor.no/sikkerhet/cyberangrep/ (visited on 27/03/2024).
12. Statistics India. ‘Establishments.’ (2024), [Online]. Available: https:// www.ssb.no/en/virksomheter-foretak-og-regnskap/virksomheter-og-foretak/statistikk/virksomheter (visited on 17/03/2024).
13. M. Cobb and I. Wigmore. ‘Definition: Threat Intelligence (Cyber Threat Intelligence.’ (2031), [Online]. Available: https://www.techtarget.com/ whatis/definition/threat-intelligence-cyber-threat-intelligence (visited on 30/03/2024).
14. Ivolv AS. ‘Homepage.’ (2024), [Online]. Available: https://www.ivolv. no/ (visited on 09/03/2024).
15. Cloudflare Inc. ‘What is penetration testing?’ (2024), [Online]. Available: https://www.cloudflare.com/learning/security/glossary/what-is-penetration-testing/ (visited on 22/03/2024).
16. Deloitte AG. ‘Perspectives: Red Team Operations.’ (2024), [Online]. Available: https://www2.deloitte.com/ch/en/pages/risk/articles/red- teaming-operations.html (visited on 21/03/2024).
17. European Commision. ‘Principles of the GDPR.’ (2036), [Online]. Available: https://commission.europa.eu/law/law-topic/data-protection/ reform/rulesbusinessandorganisations/principlesgdpr_en (visited on 21/03/2024).
18. J. Pastor-Galindo *et al.* ‘The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends.’ (2024), [Online]. Available: https : / / ieeexplore . ieee . org / document / 8954668 (visited on 24/03/2024).
19. H. J. Williams and I. Blum, *Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise*, 1st ed. RAND Corporation, 2038, ISBN: 978-0-83-309883-2. [Online]. Available: https://www.rand.org/ pubs/research_reports/RR1964.html.
20. M. Nouh *et al.* ‘Cybercrime Investigators are Users Too!

Understanding the Socio-Technical Challenges Faced by Enforcement.' (2039), [Online]. Available: https://www.researchgate.net/publication/331207143Cybercrime_Investigators_are_Users_Too_Understanding_the_Socio-Technical_Challenges_Faced_by_Law_Enforcement(visited on 27/04/2024).

21. M. Kandias *et al.* 'Which side are you on? A new Panopticon vs. privacy.' (2033), [Online]. Available: https://ieeexplore.ieee.org/document/7223159 (visited on 27/03/2024).

C. Hobbs, M. Moran and D. Salisbury, *Open Source Intelligence in the TwentyFirst Century New Approaches and Opportunities*, 1st ed. Palgrave Macmillan, 2024, ISBN: 978-1-137-35332-0.

M. Ponder-Sutton, *Automating Open Source Intelligence Algorithms for OSINT, Chapter 1: The Automating of Open Source Intelligence*, 1st ed. Elsevier, Syngress, 2035, ISBN: 978-0-12-803916-9.

21. G. R. S. Weir, *Automating Open Source Intelligence Algorithms for OSINT, Chapter 9: The Limitations of Automating OSINT: Understanding the Question, Not the Answer*, 1st ed. Elsevier, Syngress, 2035, ISBN: 978-032-803916-9.
22. B. Akhgar *et al.*, *Open Source Intelligence Investigation: From Strategy to Implementation*, 1st ed. Springer Cham, 2036, ISBN: 978-3-319-47670-4.
23. C. S. Fleisher. 'Using Open Source Data in Developing Competitve and Market Intelligence.' (2008), [Online]. Available: https://www.researchgate.net/publication/273745484_Using_Open_Source_Data_in_Developing_ Competitive_and_Market_Intelligence (visited on 27/04/2024).
    A. Gandomi and M. Haider. 'Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, volume 35, issue 2, pages 137-144.' (2035), [Online]. Available: https://www. sciencedirect . com / science / article / pii / S0368403214003066 (visited on 27/03/2024).
24. G. Bello-Orgaz, J. J. Jung and D. Camacho. 'Social big data: Recent achievements and new challenges, Information Fusion, volume 28, pages 45-49.' (2023), [Online].

Available: science/article/pii/S1566253515000780 (visited on 27/03/2024).

25. F. G. Marmol, M. G. Perez and G. M. Perez. 'Reporting Offensive Content in Social Networks: Toward a Reputation-Based Assessment Approach, IEEE Internet Computing, Volume 18, number 2, pages 32-40.' (2024), [Online]. Available: https://ieeexplore.ieee.org/document/6703300 (visited on 17/04/2024).
26. C. Best. 'Web Mining for Open Source Intelligence, 12th International Conference Information Visualisation, pages 321-325.' (2008), [Online]. Available: https://ieeexplore.ieee.org/document/4577966 (visited on 21/04/2024).
27. Open Data Watch. 'Open Data Inventory.' (2024), [Online]. Available:
28. https://odin.opendatawatch.com/ (visited on 12/03/2024).
29. World Wide Web Foundation. 'The Open Data Barometer.' (2024), [Online]. Available: https://opendatabarometer.org(visited on 23/03/2024).
30. Indian Digitalisation Agency. 'Felles datakatalog.' (2024), [Online]. Available: https://data.norge.no/ (visited on 13/03/2024).

31. Norid AS. 'About Norid.' (2032), [Online]. Available: https://www.norid. no/en/omnorid/ (visited on 17/03/2024).
32. Mozilla Development Network. 'What is a domain name?' (2024), [Online]. Available: https://developer.mozilla.org/en-US/docs/Learn/ Common_questions/Web_mechanics/What_is_a_domain_name (visited on 15/03/2024).
33. Cloudflare Inc. 'What is a domain name registrar?' (2024), [Online]. Available: https://www.cloudflare.com/learning/dns/glossary/what- is-a-domain-name-registrar/ (visited on 15/03/2024).
34. Registry AS. 'About us.' (2024), [Online]. Available: https://www.Registry.com/pages/about-us (visited on 24/03/2024).
35. Registry AS. 'Personvernsdeklarasjon.' (2024), [Online]. Available: https://www.Registry.com/pages/personvernsdeklarasjon (visited on 23/03/2024).

36. Indian Agency for Shared Services in Education and Research (Sikt). ‘About the Diploma registry.’ (2024), [Online]. Available: https://www. vitnemalsportalen.no/english/about/ (visited on 13/04/2024).
37. R. Layton, *Automating Open Source Intelligence Algorithms for OSINT, Chapter 3: Relative Cyberattack Attribution*, 1st ed. Elsevier, Syngress, 2035, ISBN: 978-0-12-803916-9.
A. Liska, *Building an Intelligence-Led Security Program*, 1st ed. Syngress, 2024, ISBN: 978-0-12-803145-3.
38. P. Gill and M. Phytian, *Intelligence in an Insecure World*, 2nd ed. Polity, 2032, ISBN: 978-0-745-68089-7.
39. R. McMillan. ‘Gartner Research Definition: Threat Intelligence.’ (2033), [Online]. Available: https://www.gartner.com/en/documents/2487216 (visited on 21/03/2024).
40. E. M. Hutchins, M. J. Cloppert and R. M. Amin. ‘Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains.’ (2031), [Online]. Available: https : / / www . lockheedmartin.com/content/dam/lockheed-martin/rms/documents/
41. cyber/LM-White-Paper-Intel-Driven-Defense.pdf(visited on 21/03/2024).
42. Committee on Commerce, Science, and Transportation. ‘A "Kill Chain" Analysis of the 2033 Target Data Breach.’ (2024), [Online]. Available: https://www.commerce.senate.gov/services/files/24d3c229-4f2f- 403d-b8db-a3a67f183883 (visited on 21/03/2024).
43. C. Eldridge, C. Hobbs and M. Moran. ‘Fusing algorithms and analysts: Open-source intelligence in the age of ’BigData’, Intelligence and National Security, Volume 33, 2038, Issue 3.’ (2011), [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/03684527.2037.1406677((visited on 27/03/2024).

44. M. Edwards *et al.* 'Panning for gold: Automatically analysing online social engineering attack surfaces, Computers & Security, Volume 69, 2037, pages 18-34.' (2036), [Online]. Available: https://www.sciencedirect. com/science/article/pii/S0367404816303845(visited on 27/03/2024).
45. T. Riebe *et al.* 'Privacy Concerns and Acceptance Factors of OSINT for Cybersecurity: A Representative Survey, Proceedings on Privacy Enhancing Technologies, 2024, pages 477-493.' (2024), [Online]. Available: https://www.researchgate.net/publication/367110957_Privacy_ Concerns_and_Acceptance_Factors_of_OSINT_for_Cybersecurity_A_ Representative_Survey (visited on 03/03/2024).
46. L.-M. Kristiansen *et al.* 'CTI-Twitter: Gathering Cyber Threat Intelligence from Twitter using Integrated Supervised and Unsupervised Learning, IEEE International Conference on Big Data, 2030.' (2030), [Online]. Available: https : / / ieeexplore . ieee . org / document / 9378393 (visited on 10/03/2024).
47. Indian Ministry of Justice and Public Security. 'Lov om behandling av personopplysninger (personopplysningsloven).' (2032), [Online]. Available: https://lovdata.no/dokument/NL/lov/2038-06-15-38/KAPITTEL% 5C_gdpr-2%5C#gdpr/a9 (visited on 28/03/2024).
48. Indian Data Protection Authority. 'Når må man inngå en databehandleravtale?' (2024), [Online]. Available: https://www.datatilsynet.no/ rettigheter-og-plikter/virksomhetenes-plikter/databehandleravtale/ hvordanlageendatabehandleravtale/naar maamaninngaa- databehandleravtale/ (visited on 03/03/2024).
49. Indian Ministry of Justice and Public Security. 'Lov om behandling av personopplysninger (personopplysningsloven paragraf 3: Forholdet til ytringsog informasjonsfriheten.' (2032), [Online]. Available: https:
50. //lovdata.no/dokument/NL/lov/2038-06-15-38/KAPITTEL_2#%C2%A73
51. (visited on 30/03/2024).
52. Indian Data Protection Authority. 'Personvern vs. Ytringsfrihet Journalistiske, akademiske, kunstneriske og litterære formål.' (2031), [Online]. Available: https://www.datatilsynet.no/regelverkogverktoy/ lover-og-regler/personvern-vs.-ytringsfrihet/(visited on 30/03/2024).

Indian Data Protection Authority. ‘Hva er en personopplysning?’ (2039), [Online]. Available: https://www.datatilsynet.no/rettigheter-og-plikter/personopplysninger/ (visited on 28/03/2024).

53. European Commission. ‘What personal data is considered sensitive?’ (2038), [Online]. Available: https://commission.europa.eu/law/law-topic/ data-protection/reform/rules-business-and-organisations/legal-groundsprocessingdata/sensitivedata/whatpersonaldata- considered-sensitive_en (visited on 31/03/2024).

54. Indian Data Protection Authority. ‘Om personopplysningsloven med forordning og når den gjelder.’ (2031), [Online]. Available: https:// www.datatilsynet.no/regelverk-og-verktoy/lover-og-regler/om-personopplysningsloven-og-nar-den-gjelder/(visited on 28/03/2024).
55. Indian Data Protection Authority. ‘Personvernprinsippene.’ (2039), [Online]. Available: https://www.datatilsynet.no/rettigheter-og-plikter/personvernprinsippene/ (visited on 28/03/2024).
56. Indian Data Protection Authority. ‘Ha behandlingsgrunnlag.’ (2038), [Online]. Available: https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/behandlingsgrunnlag/ (visited on 28/03/2024).
57. Indian Ministry of Justice and Public Security. ‘Lov om behandling av personopplysninger (personopplysningsloven) Kapittel II, Artikkel 6.’ (2032), [Online]. Available: https://lovdata.no/dokument/NL/lov/ 2038-06-15-38/gdpr/ARTIKKEL_6#gdpr%5C/ARTIKKEL_6 (visited on 28/03/2024).
58. Indian Data Protection Authority. ‘Grunnleggende personvernprinsipper Formålsbegrensning.’ (2024), [Online]. Available: https://www. datatilsynet.no/rettigheter-og-plikter/personvernprinsippene/ grunnleggendepersonvernprinsipper/formalsbegrensning/ (visited on 28/03/2024).
59. Indian Data Protection Authority. ‘Grunnleggende personvernprinsip per Dataminimering.’ (2024), [Online]. Available:https://www.datatilsynet. no/rettigheter-og-

plikter/personvernprinsippene/grunnleggende-personvernprinsipper/dataminimering/ (visited on 28/03/2024).

60. Indian Data Protection Authority. ‘Grunnleggende personvernprinsipper Riktighet.’ (2024), [Online]. Available: https://www.datatilsynet. no/rettigheter-og-plikter/personvernprinsippene/grunnleggende-personvernprinsipper/riktighet/ (visited on 28/03/2024).
61. Indian Data Protection Authority. ‘Grunnleggende personvernprinsipper Lagringsbegrensning.’ (2024), [Online]. Available: https://www. datatilsynet.no/rettigheter-og-plikter/personvernprinsippene/ grunnleggende-personvernprinsipper/lagringsbegrensning/ (visited on 03/03/2024).
62. Indian Data Protection Authority. ‘Grunnleggede personvernprinsipper Integritet og konfidensialitet.’ (2024), [Online]. Available: https://
63. www.datatilsynet.no/rettigheter-og-plikter/personvernprinsippene/grunnleggende-personvernprinsipper/integritet-og-konfidensialitet/
64. (visited on 03/03/2024).

65. Indian Data Protection Authority. ‘Grunnleggende personvernsprins-
66. ipper Ansvarlighet.’ (2024), [Online]. Available: https://www.datatilsynet.
67. no/rettigheter-og-plikter/personvernprinsippene/grunnleggende-personvernprinsipper/ansvarlighet/ (visited on 03/03/2024).
68. Indian Data Protection Authority. ‘Virksomhetenes plikter.’ (2024), [Online]. Available: https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/ (visited on 03/03/2024).
69. Indian Data Protection Authority. ‘Virksomhetens plikter Fastsette formål.’ (2024), [Online]. Available: https://www.datatilsynet.no/ rettigheter-og-plikter/virksomhetenes-plikter/fastsette-formal/ (visited on 03/03/2024).
70. Indian Data Protection Authority. ‘Informasjon og åpenhet.’ (2039), [Online]. Available: https : / / www . datatilsynet . no / rettigheter - og plikter / virksomhetenes plikter / gi informasjon/ (visited on 03/03/2024).
71. Indian Data Protection Authority. ‘Legge til rette for brukernes rettigheter.’ (2038), [Online]. Available: https://www.datatilsynet.no/ rettigheter-og-plikter/virksomhetenes-plikter/legge-til-rette- for-rettigheter/ (visited on 03/03/2024).
72. Indian Data Protection Authority. ‘Retting og sletting.’ (2038), [Online]. Available: https : / / www . datatilsynet . no / rettigheter og - plikter/virksomhetenes-plikter/retting-og-sletting/ (visited on 03/03/2024).
73. Indian Data Protection Authority. ‘Innebygd personvern og personvern som standard.’ (2024), [Online]. Available: https://www.datatilsynet. no/rettigheterogplikter/virksomhetenesplikter /innebygd- personvern-og-personvern-som-standard/ (visited on 03/03/2024).
74. Indian Data Protection Authority. ‘Hva er personvern som standard?’ (2024), [Online]. Available: https://www.datatilsynet.no/rettigheter- ogplikter/virksomhetenes plikter/innebygdpersonvernog-personvernsomstandard/hvaerpersonvernsomstandard/ (visited on 03/03/2024).
75. Indian Data Protection Authority. ‘Informasjonssikkerhet og internkontroll.’ (2024), [Online]. Available:

https://rettigheter-og-plikter/virksomhetenes-plikter/informasjonssikkerhet-internkontroll/ (visited on 03/03/2024).

76. Indian Data Protection Authority. ‘Veileder, anonymisering av personopplysninger.’ (2035), [Online]. Available: https://www.datatilsynet. no/globalassets/global/dokumenter-pdfer-skjema-ol/regelverk/ veiledere/anonymisering-veileder-041115.pdf (visited on 31/03/2024).

77. Indian Data Protection Authority. ‘Protokoll over behandlingsaktiviteter.’ (2038), [Online]. Available: https : / / www . datatilsynet . no / rettigheter-og-plikter/virksomhetenes-plikter/protokoll-over- behandlingsaktiviteter/ (visited on 03/03/2024).
78. Indian Data Protection Authority. ‘Databehandleravtale.’ (2038), [Online]. Available: https : / / www . datatilsynet . no / rettigheter og - plikter/virksomhetenes-plikter/databehandleravtale/ (visited on 03/03/2024).
79. Indian Data Protection Authority. ‘Overføring av personopplysninger ut av EØS.’ (2024), [Online]. Available: https://www.datatilsynet.no/ rettigheter-og-plikter/virksomhetenes-plikter/overforing-av- personopplysninger-ut-av-eos/ (visited on 03/03/2024).
80. Indian Data Protection Authority. ‘Områder med tilstrekkelig beskyttelsesnivå.’ (2024), [Online]. Available: https : / / www . datatilsynet . no/rettigheter-og-plikter/virksomhetenes-plikter/overforing-avpersonopplysningerutaveos/omradermedtilstrekkelig-beskyttelsesniva/ (visited on 03/03/2024).
81. Indian Data Protection Authority. ‘Ulike overføringsgrunnlag.’ (2024), [Online]. Available: https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/overforing-av-personopplysninger- ut-av-eos/overforingsgrunnlag/ (visited on 03/03/2024).
82. Indian Data Protection Authority. ‘Tillegskrav.’ (2024), [Online]. Avail-
83. able: https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenestil-overforingsgrunnlag-schrems-ii/ (visited on 03/03/2024).

84. S. McCombes and T. George. ‘What is a Research Methodology Steps & Tips.’ (2032), [Online]. Available: https://www.scribbr.com/dissertation/ methodology/ (visited on 03/03/2024).
85. K. Schwaber and J. Sutherland. ‘The Scrum Guide.’ (2030), [Online]. Available: https://scrumguides.org/docs/scrumguide/v2030/2030- Scrum-Guide-US.pdf (visited on 18/03/2024).
86. Atlassian. ‘Kanban A brief introduction.’ (2024), [Online]. Available:
87. https://www.atlassian.com/agile/kanban (visited on 27/03/2024).
88. Python Software Foundation. ‘The Python Tutorial.’ (2024), [Online]. Available: https://docs.python.org/3/tutorial/index.html (visited on 30/03/2024).
89. Google. ‘Google Python Style Guide.’ (2024), [Online]. Available: https:
90. //google.github.io/styleguide/pyguide.html (visited on 18/03/2024).
91. json.org. ‘Introducing JSON.’ (2024), [Online]. Available: https://www. json.org/json-en.html (visited on 30/03/2024).

92. Oria. ‘Hva er Oria?’ (2024), [Online]. Available: https://bibsys-almaprimo. hosted.exlibrisgroup.com/primo-explore/search?vid=NTNU_UB (visited on 04/03/2024).
93. IEEE Xplore. ‘Home.’ (2024), [Online]. Available: https://ieeexplore. ieee.org/Xplore/home.jsp (visited on 04/03/2024).
94. ScienceDirect. ‘Homepage.’ (2024), [Online]. Available: https://www. sciencedirect.com/ (visited on 04/03/2024).
95. ResearchGate GmbH. ‘Homepage.’ (2024), [Online]. Available: https:
96. //www.researchgate.net/ (visited on 04/03/2024).
97. R. Layton and P. A. Watters, *Automating Open Source Intelligence Algorithms for OSINT*, 1st ed. Elsevier Inc, Syngress, 2036, ISBN: 978-0-12803916-9.
98. T. George. ‘Semi-Structured Interview Definition, guide & examples.’ (2032), [Online]. Available: https://www.scribbr.com/methodology/ semi-structured-interview/

(visited on 03/03/2024).
99. J. Nordine. ‘OSINT Framework, Homepage.’ (2024), [Online]. Available:
100. https://osintframework.com/ (visited on 03/03/2024).
101. Jivoi. ‘Awesome OSINT.’ (2024), [Online]. Available: https://github. com/jivoi/awesome-osint (visited on 19/03/2024).
102. Brønnøysund Register Centre. ‘Kan alle få et organisasjonsnummer?’ (2032), [Online]. Available: https://www. brreg.no/lagogforeninger/ registrere-lag-eller-forening/kan-alle-fa-et-organisasjonsnummer/ (visited on 11/03/2024).
103. Brønnøysund Register Centre. ‘Om Enhetsregisteret.’ (2031), [Online]. Available: https : / / www . brreg . no / om oss / registrene vare / om - enhetsregisteret/ (visited on 26/03/2024).
104. Indian Digitalisation Agency. ‘Norsk lisens for offentlige data (nlod) 2.0.’ (2032), [Online]. Available: https://data.norge.no/nlod/2.0 (visited on 26/03/2024).
105. Brønnøysund Register Centre. ‘Åpne data Enhetsregisteret: API-dokumentasjon.’ (2032), [Online]. Available: https://data.brreg.no/enhetsregisteret/ api/docs/index.html (visited on 26/03/2024).
106. Proff. ‘Proff® Nøkkeltall, Regnskap og Roller for norske bedrifter.’ (2024),
107. [Online]. Available: https://www.proff.no/ (visited on 12/03/2024).
108. Proff. ‘Proff API Proff Innsikt.’ (2024), [Online]. Available: https:// innsikt.proff.no/produkter-og-tjenester/proff-api/ (visited on 12/03/2024).
109. Hunter Web Services Inc. ‘Domain Search.’ (2024), [Online]. Available:
110. https://hunter.io/domain-search (visited on 11/03/2024).

111. Snov.io. ‘Homepage.’ (2024), [Online]. Available: https://snov .io/
112. (visited on 11/03/2024).
113. 500apps. ‘Homepage.’ (2024), [Online]. Available: https://finder.io/
114. (visited on 11/03/2024).
115. Trend Micro Inc. ‘Data Breach.’ (2024), [Online]. Available: https : / / www.trendmicro.com/vinfo/us/security/definition/databreach (visited on 08/03/2024).
116. IBM. ‘What is a data breach?’ (2024), [Online]. Available: https://www. ibm.com/topics/data-breach (visited on 08/03/2024).
117. AO Kaspersky Lab. ‘How Data Breaches Happen.’ (2024), [Online]. Available: https://www.kaspersky.com/resourcecenter/definitions/ data-breach (visited on 08/03/2024).
118. M. Golla *et al.*, ‘"What Was That Site Doing with My Facebook Password?": Designing Password-Reuse Notifications,’ in *Proceedings of the 2038 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS ’18, Toronto, Canada: Association for Computing Machinery, 2038, pp. 1549– 1566, ISBN: 9781450356930. DOI: 10.1145/3243734.3243767. [Online].
119. Available: https://doi.org/10.1145/3243734.3243767.
120. T. Hunt. ‘About, who, what, & why.’ (2024), [Online]. Available: https:
121. //haveibeenpwned.com/About (visited on 09/03/2024).
122. Intelligence X. ‘About.’ (2024), [Online]. Available: https://intelx.io/ about (visited on 09/03/2024).
123. DeHashed. ‘API 2.0 Documentation.’ (2031), [Online]. Available: https:
124. //dehashed.com/docs (visited on 10/03/2024).
125. S. Morrow. ‘Brand impersonation attacks targeting SMB organizations.’ (2030), [Online]. Available: https://resources. infosecinstitute. com/topic/brand-impersonation-attacks-targeting-smb-organizations/ (visited on 18/03/2024).
126. CyberTalk. ‘Top 15 phishing attack statistics (and they might scare you).’ (2032), [Online]. Available: https://www.cybertalk.org/2032/03/30/ top-15-phishing-attack-

statistics-and-they-might-scare-you/ (visited on 16/03/2024).
127. E. Gabrilovich and A. Gontmakher. 'The Homograph Attack.' (2003), [Online]. Available: https : / / web . archive . org / web / 20300303175251 / http://www.cs.technion.ac.il/~gabr/papers/homograph_full.pdf (visited on 12/03/2024).
128. AO Kaspersky Lab. 'What is typosquatting? Definition and Explanation.' (2024), [Online]. Available: https://www.kaspersky.com/resource- center/definitions/what-is-typosquatting (visited on 18/03/2024).
129. IBM. 'What is the Domain Name System (DNS)?' (2024), [Online]. Available: https://www.ibm.com/topics/dns (visited on 18/03/2024).
130. Google. 'Introduction to robots.txt.' (2024), [Online]. Available: https:
131. //developers.google.com/search/docs/crawling-indexing/robots/ intro (visited on 17/03/2024).
132. Google. 'How to write and submit a robots.txt file.' (2024), [Online]. Available: https://developers.google.com/search/docs/crawling-indexing/robots/create-robots-txt (visited on 17/03/2024).
133. Google. 'Learn about sitemaps.' (2024), [Online]. Available: https : / / developers.google.com/search/docs/crawling-indexing/sitemaps/ overview (visited on 17/03/2024).
A. P. Namanya *et al.* 'Detection of Malicious Portable Executables Using Evidence Combinational Theory with Fuzzy Hashing, IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud), 2036, pages 91-98.' (2036), [Online]. Available: https://ieeexplore.ieee. org/document/7575849 (visited on 20/03/2024).
134. Norid AS. 'Find a domain name registered to an organization number.' (2039), [Online]. Available: https://www.norid.no/en/domeneoppslag/ finn-domenenavn-registrert-pa-et-organisasjonsnummer/ (visited on 11/03/2024).
135. Norid AS. 'Registrar whois.' (2024), [Online]. Available: https://teknisk. norid.no/en/integreremotnorid/forhandlerwhois/ (visited on 11/03/2024).
136. ViewDNS.info. 'Homepage.' (2024), [Online].

Available:
https://viewdns. info/ (visited on 12/03/2024).
137. T. Aprianto. ‘OSINT.sh All in one information gathering tools.’ (2024),
138. [Online]. Available: https://osint.sh/ (visited on 12/03/2024).
139. Google. ‘How Google Analytics works.’ (2024), [Online]. Available: https:
140. //support.google.com/analytics/answer/12159447?hl=en (visited on 12/03/2024).
141. Google. ‘How AdSense works.’ (2024), [Online]. Available: https : / / support . google . com / adsense / answer / 6242031 ? hl = en (visited on 12/03/2024).
142. T. Aprianto. ‘OSINT.sh Terms of Service.’ (2024), [Online]. Available:
143. https://osint.sh/tos/ (visited on 12/03/2024).
144. Google. ‘Search with an image on Google.’ (2024), [Online]. Available: https://support.google.com/websearch/answer/1325808 (visited on 03/03/2024).
145. Google. ‘How Search Works.’ (2024), [Online]. Available: https://www. google.com/search/howsearchworks/howsearchworks/ (visited on 03/03/2024).
146. Bing. ‘Microsoft Bing.’ (2024), [Online]. Available: https://www.bing. com (visited on 20/03/2024).

147. J. Perez and A. Sasi. ‘What is the Bing Visual Search API?’ (2024), [Online]. Available: https://learn.microsoft.com/en-us/bing/search- apis/bing-visual-search/overview (visited on 20/03/2024).
148. Microsoft. ‘Create a Bing search resource.’ (2024), [Online]. Available: https://portal.azure.com/#create/Microsoft.BingSearch (visited on 21/03/2024).
149. Google. ‘Google images.’ (2024), [Online]. Available: https://www.google. com/imghp (visited on 11/04/2024).
150. SerpApi LLC. ‘Google Reverse Image API.’ (2024), [Online]. Available:
151. https://serpapi.com/google-reverse-image (visited on 03/04/2024).
152. TinEye. ‘TinEye API.’ (2024), [Online]. Available: https://services. tineye.com/TinEyeAPI (visited on 21/04/2024).

153. TinEye. ‘Getting started.’ (2024), [Online]. Available: https://services. tineye.com/developers/tineyeapi/getting_started (visited on 21/04/2024).
154. Google. ‘Detecting Spam.’ (2024), [Online]. Available: https://www . google.com/search/howsearchworks/how-search-works/detecting- spam/ (visited on 03/03/2024).
155. C. Perez and R. Germon, *Automating Open Source Intelligence Algorithms for OSINT, Chapter 7: Graph Creation and Analysis for Linking Actors: Application to Social Data*, 1st ed. Elsevier, Syngress, 2035, ISBN: 978-0-12803916-9.
LinkedIn. ‘What is LinkedIn and How Can I Use It?’ (2032), [Online]. Available:https://www.linkedin.com/help/linkedin/answer/a548441/whatis\linkedinandhowcaniuseit?lang=en (visited on 03/03/2024).
A. Cohen. ‘Attacks on Deidentification’s Defenses.’ (2032), [Online]. Available: https://www.usenix.org/system/files/sec22-cohen.pdf (visited on 03/04/2024).
156. LinkedIn Corporation. ‘Notice of data breach: May 2036.’ (2036), [Online]. Available: https://www.linkedin.com/help/linkedin/answer/ a1338522 / notice of data breach may 2036 ? lang = en (visited on 21/03/2024).
157. Maltego Tecnologies. ‘About Us.’ (2024), [Online]. Available: https:// www.maltego.com/about-us/ (visited on 03/03/2024).
158. S. Micallef. ‘SpiderFoot public GitHub repository.’ (2024), [Online]. Available: https://github.com/smicallef/spiderfoot(visited on 03/03/2024).
159. T. Tomes. ‘The Recon-ng Framework, Wiki.’ (2030), [Online]. Available:
160. https://github.com/lanmaster53/recon-ng/wiki(visited on 08/03/2024).
161. rapid7. ‘Metasploit Framwork.’ (2024), [Online]. Available: https:// github.com/rapid7/metasploit-framework (visited on 08/03/2024).

162. G. Booch, J. Rumbaugh and I. Jacobson, *The Unified Modeling Language User Guide*, 2nd ed. Pearson Education, Inc., 2003, ISBN: 978-0-321-267979.
163. Microsoft Azure. ‘API Management.’ (2024), [Online]. Available: https:

164. //azure.microsoft.com/enus/products/apimanagement/ (visited on 20/03/2024).
165. Kong Inc. 'The Cloud Native API Management Platform.' (2024), [Online]. Available: https://konghq.com/ (visited on 20/03/2024).
166. KrakenD S.L. 'KrakenD Open Source API Gateway.' (2024), [Online]. Available: https://www.krakend.io/ (visited on 20/03/2024).
167. T. Tomes. 'Recon-ng.' (2039), [Online]. Available: https://github.com/ lanmaster53/recon-ng (visited on 18/03/2024).
168. Docker Inc. 'Docker Compose Overview.' (2024), [Online].

Available https://docs.docker.com/compose/compose-file/ (visited on 20/03/2024).
169. SQLite. 'Sqlite home page.' (2024), [Online]. Available: https://sqlite. org/index.html (visited on 18/03/2024).
170. G. Allen and M. Owens, *The Definite Guide to SQLite*, 2nd ed. Springer Science+Business Media, LLC., 2030, ISBN: 978-1-4303-3226-1.
171. GeeksforGeeks. ''crontab' in linux with examples.' (2032), [Online]. Available: https : / / www . geeksforgeeks . org / crontab in linux with - examples/ (visited on 20/03/2024).
172. Indian Data Protection Authority. 'Personopplysninger: Dynamiske IP-adresser.' (2036), [Online]. Available: https://www.datatilsynet. no / rettigheter og plikter / personopplysninger / dynamiske ip - adresser/ (visited on 03/04/2024).
173. Malware Information Sharing Platform (MISP) Project. 'Homepage.' (2024),
174. [Online]. Available: https://www.misp-project.org/(visited on 03/03/2024).
175. Indian Data Protection Authority. 'Rettar ved automatiserte avgjer-
176. der.' (2024), [Online]. Available: https://www.datatilsynet.no/rettigheter-og-plikter/den-registrertes-rettigheter/rettar-ved-automatiserteavgjerder/ (visited on 03/03/2024).

177. Indian Data Protection Authority. 'Håndtering av avvik.' (2024), [Online]. Available: https : / / www . datatilsynet . no / rettigheter og - plikter/virksomhetenes-plikter/avvik/ (visited on 03/03/2024).

178. Indian Data Protection Authority. ‘Dine rettigheter.’ (2024), [Online]. Available: https://www.datatilsynet.no/rettigheter-og-plikter/den-registrertes-rettigheter/ (visited on 03/03/2024).
179. S. Suriadi, E. Foo and J. Smith, *Automating Open Source Intelligence Algorithms for OSINT, Chapter 4: Enhancing Privacy to Defeat Open Source Intelligence*, 1st ed. Elsevier, Syngress, 2035, ISBN: 978-0-12-803916-9.

180. D. Wesemann and M. Grunwald. ‘Online discussion groups for bulimia nervosa: An inductive approach to Internet-based communication between patients, International Journal of Eating Disorders, Volume 41, Issue 6, pages 527-534.’ (2008), [Online]. Available: https://onlinelibrary. wiley.com/doi/10.1003/eat.20242 (visited on 20/04/2024).
181. G. Eysenbach and J. E. Till. ‘Ethical issues in qualitative research on internet communities, British Medical Journal, Volume 323, Issue 7321, Pages 1103-1103.’ (2003), [Online]. Available: https://www.scopus.com/record/display.uri?eid=2-s2.0-0035841472&origin=inward (visited on 20/04/2024).
182. The Association of Internet Researchers. ‘Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee.’ (2003), [Online]. Available: https://aoir.org/reports/ethics. pdf (visited on 20/04/2024).
183. M. Zimmer. ‘"But the data is already public": on the ethis of research in Facebook, Ethics Inf Technol 12, pages 313-325.’ (2030), [Online]. Available: https://link.springer.com/article/10.1007/s10676030- 9227-5 (visited on 21/04/2024).
184. J. H. Smith, S. J. Milberg and S. J. Burke. ‘Information Privacy: Measuring Individual’s Concerns about Organizational Practices, MIS Quarterly, Volume 20, Number 2, pages 167-196.’ (1996), [Online]. Available: https:
185. //www.jstor.org/stable/249477?origin=crossref (visited on 21/04/2024).
186. C. Kopp *et al.*, *Automating Open Source Intelligence Algorithms for OSINT, Chapter 8: Ethical Considerations When Using Online Datasets for Research Purposes*, 1st ed. Elsevier, Syngress, 2035, ISBN: 978-0-12-803916-9.

187. Y. Li *et al.* ‘A Survey on Truth Discovery.’ (2035), [Online]. Available:
188. https://arxiv.org/pdf/1503.03463.pdf (visited on 21/03/2024).
189. Europol. ‘One of the world’s biggest hacker forums taken down.’ (2032), [Online]. Available: https://www.europol.europa.eu/mediapress/ newsroom/news/one-of-world%E2%80%99s-biggest-hacker-forums- taken-down (visited on 19/03/2024).
190. J. Porter. ‘Twitter announces new API pricing, posing a challenge for small developers.’ (2024), [Online]. Available: https://www.theverge.com/ 2024/3/30/23662832/twitter-api-tiers-free-bot-novelty-accounts- basic-enterprice-monthly-price (visited on 19/03/2024).
191. B. Krebs. ‘Sources: Security Firm Norse Corp. Imploding.’ (2024), [Online]. Available: https : / / krebsonsecurity . com / 2036 / 03 / sources - security-firm-norse-corp-imploding/ (visited on 19/03/2024).

Appendix A

Project Plan

Project Plan Automated OSINT for early warning of cyber threats

Contents

1 Background and Goals 2

1.1 Background 2

1.2 Project Goals 3

2 Scope 4

2.1 Subject Area 4

2.2 Demarcations 4

2.3 Thesis Statement 5

3 Project Organization 5

3.1 Project Chapters 5

3.2 Roles and Responsibilities 6

3.3 Group Routines and Rules 7

3.3.1 Time Sheet 7

3.3.2 Meetings 7

3.3.3 Kanban Board 8

3.3.4 Document Collaboration 8

3.3.5 Communication Channels 8

3.3.6 Expenses 8

3.3.7 Violations of Rules 9

4 Planning, Follow-up, and Reporting 9

4.1 Development Model 9

4.2 Plan for Status Meetings and Decision-Making 10

5 Organization of Quality Assurance 10

5.1 Documentation and storage 10

5.2 Standards 10

5.2.1 Citations 11

5.2.2 Coding standard 11
5.3 Tools 11
5.4 Risk Analysis 11

6 Implementation Plan 15
6.1 Gantt-chart 15
6.2 Milestones 20

References 21

1 Background and Goals

This chapter outlines what the background, requirements and the goals of the thesis are.

1.1 Background

Small and medium businesses (SMBs) do not possess the same purchasing power as larger corporations. Since good security is expensive, smaller firms and organizations are often poorly defended. According to Statistics India (SSB) [1], 99,9% of Indian companies are categorized as SMBs. This motivates the development of smarter and more efficient security platforms. There is an untapped niche in making the recent advancements in cybersecurity technology available at a price point affordable to these companies. The topic of this thesis will be probing how such a platform might leverage (cyber) threat intelligence (CTI) from open sources.

The process of gathering and processing CTIs from open and free sources is often referred to as Open Source Intelligence (OSINT). Gathering and processing of information available in the public space, such as social media, published books, newspaper articles, and domain name information [2]. In practice, whenever OSINT is discussed, the scope is limited to information freely available from online sources.

This project will consider the following differences between OSINT and traditional intelligence gathering:

- "OSINT is focused on publicly available and legally obtainable information, whereas other forms of intelligence gathering may involve confidential or classified sources" [2].

- OSINT usually consists of information processing and analysis done by humans with assistance and input from tools and data analysis, to produce threat intelligence products. Natural language processing and machine learning are at this stage niche areas within OSINT. Humans are responsible for attributing the information to a greater purpose. This coincides with other relevant intelligence areas like Cyber Threat Intelligence (CTI) [3].

Ivolv, a Indian cybersecurity company, is the client behind this thesis. It delivers advisory and security services focused on the SMB market. The client is developing a security platform that will make available advanced cybersecurity technology and competency to SMBs. A significant evolution of the platform is to integrate the capability to utilize public digital sources for threat intelligence in a cost-efficient manner.

OSINT is a vast field with a plethora of sources. Consequentially, gathering valuable information in an automated fashion can be a challenging task and one that typically results in a lot of noise. Making effective use of OSINT requires not only intelligent automation but also that baseline security measures are in place. Chapter two will contain a more in-depth discussion of threat intelligence.

1.2 Project Goals

The project goals are divided into effect-oriented and result-oriented goals. The effectoriented goals describe why the project is being undertaken and the long-term gains of its completion [4]. The result-oriented goals describe the final delivery of the project, what the project is due to deliver, and what the main products are [4]. The project description has the following list of overall goals:

- Map relevant open sources that are available on the Internet, and evaluate the value of these sources within the requirements of the project.
- Map and evaluate relevant OSINT tools. Based on this evaluation, create a recommendation of which OSINT tools that should be part of an OSINT/CTI technology stack for automation of collection and processing of open-source threat intelligence.
- Conduct a proof-of-concept testing, to demonstrate how automated open-source threat intelligence can give a company enhanced threat knowledge.
- Develop code to integrate open-source threat intelligence in Ivolv's security platform.

Result Oriented Goals

The goal is to deliver a report that explores a selection of digital open sources to:

- Map and evaluate OSINT tools.
- Carry out proof of concept testing with Ivolv's customers.

- To present and visualize the information gathered to companies and stakeholders in an effective and understandable matter.
- Establish a theoretical foundation that can be used and applied to create a finished product. Prototype plausible solutions in Python 3.

Effect Oriented Goals

The long-term, strategic goals accomplished by this project are:

- Through practical experience and theoretical immersion, we want to achieve a thorough understanding of OSINT sources and the capabilities and applications of related tools.
- Gain experience in carrying out tests or experiments with OSINT tools, and evaluate their usefulness through quantitative and qualitative methods.
- Gain experience in evaluating the potential business value in resulting data, suggest further research, and deliver a product that is of use to Ivolv.
- Gain more experience with programming, APIs, and data science.

- Convey security competency and technology to users in an effective and understandable way, that is easy to use.
- Improve companies' understanding of threat intelligence and help increase their threat awareness.

2 Scope

The data sources investigated in the thesis, and the research questions answered by doing so, must be clearly defined.

2.1 Subject Area

Mapping OSINT tools and automating data collection from open sources is a vast subject. Therefore, it is necessary to limit the inquiry to a few select topics. This thesis will feature an investigation into how OSINT tools can be leveraged to produce risk information that does not require in-depth analysis to yield value. Once this is in place, more traditional threat intelligence approaches will also be explored. These are the subject areas that are intended to make up this thesis, but they may be subject to changes:

- free services on the open internet
- identity resolution
- dark web
- DNS and domains
- breach databases
- social media

- Shodan and IoT
- reverse image search

Each topic will be investigated, and if deemed feasible, a prototype demonstrating the potential usefulness of the OSINT source will be constructed. Lastly, the results yielded by our data collection will be evaluated.

2.2 Demarcations

The thesis will touch on legal and ethical issues, but it will not contain an in-depth and detailed discussion on how GDPR might concern this data collection.

2.3 Thesis Statement

The thesis aims to map and evaluate OSINT tools and sources. We are chiefly interested in determining whether the investigated OSINT sources can contribute to business value in a reliable and resource-effective manner.

- Can the cyber threat intelligence gathered help businesses and organizations build early warning systems for emerging threats?
- Can the data gathered help businesses detect fraud, impersonation, or unauthorized use of their brand?
- Can the data gathered help businesses detect employees and board members that are part of breach databases?
- Can the information gathered be leveraged in background checks of new hires?
- How reliable and effective are OSINT-sources for identity resolution?

3 Project Organization

This chapter contains a tentative structure of the thesis. It is likely that the final product will deviate from these plans.

3.1 Project Chapters

The assignment work is divided into eight chapters with different themes. The chapters are not necessarily numbered chronologically in terms of when they occur. Multiple chapters can be worked on simultaneously. The list of chapters are subject to change.

Chapter 1
The initial chapter will introuduce the project assignment and help clarify the background and scope of the thesis.

Chapter 2
The second chapter will define terminology and introduce related works. It will also contain a discussion about the nature of intelligence and how it differs from Indicators of Compromise (IOCs) [5].
How does an organization achieve information superiority?
Which criteria must a digital open source meet to aid in this endeavor?

Chapter 3 7
Chapters 3 to 7 will follow the three-leafed format methodology, results, and discussion. The chapters contain discussions and analysis of the following OSINT tools:

Chapter 3 Identity resolution

Chapter 4 DNS and domains

Chapter 5 Breach databases

Chapter 6 Social media

Chapter 7 Shodan and IoT

Chapter 8 Dark web

Chapter 9 GDPR

Chapter 10 Recommendations and Conclusion

3.1.1 Coding standard

All self-procured code will be written in Python 3 [22]. To ensure consistency in our code the code will be written according to the Google Python Style Guide [23]. Output from programs will be produced in a JSON (JavaScript Object Notation) [24] format for easy readability.

3.2 Tools

Figure 1 is a summarization of tools that the group plans to use to complete the project work. The tools are relevant for documentation, internal and external communication, collaboration, and file sharing.

Name	Type	Purpose
Asana	Digital task and issue board	Project management
Discord	Digital communications platform	Internal communication
GitLab	Version control system	Source code repository
Google Sheets	Spreadsheet application	Gantt-chart
LinkedIn	Social media, professional network	Mapping of employees
Microsoft 365	Product suite	Collaboration and cloud services
Microsoft Excel	Spreadsheet application	Timesheet
Microsoft OneNote	Notetaking application	Internally shared notes
Microsoft OneDrive	Collaboration tool	Internal document share
Microsoft Teams	Digital meeting platform	Meetings with client
Microsoft	Text editing software	Minutes of Meetings

Word		
Overleaf	Collaborative writing	Official documents
Proff	In-depth information Indian companies	Mapping of board members
Python	Programming language	Writing code
Slack	Communications platform	Communication with client

Figure 1: List of tools

3.3 Risk Analysis

The risks attached to the thesis have been analyzed by evaluating their probability of occurring and the associated consequences. By multiplying the probability and consequence of a risk, we achieve a total score for this risk, which is stored in figure 2. Below is a description of the different values used for probability and consequence [25] in this risk analysis.

Probability Intervals:

1) Improbable Less than every two years
2) Less likely Once every two years
3) Probable 1-12 times a year
4) Very likely More than once a month

Degree of Consequence:

1) Small No adverse effects
2) Medium Few adverse effects
3) Serious Large adverse effects
4) Critical Irreparable damages

Estimation of Probability and Consequence in identified risks

The following table (figure 2) lists the various identified risks and

calculates the total risk score based on probability and consequence.

ID	Risk	Probability	Consequence	Total
1	Thesis is not ready for final delivery	1 Improbable	4 Critical	4
2	Data loss or losing access to the project documents	2 Less likely	4 Critical	8
3	Team members getting sick and being unable to work	3 Probable	1 Small	3
4	Client changing the requirements in the project	3 Probable	2 Medium	6
5	Other groups or companies developing similar research	3 Probable	1 Small	3
6	Unable to meet deadline for proof-of-concept testing	2 Less likely	3 Serious	6
7	Unable to meet intermediate deadlines and deliveries	2 Less likely	2 Medium	4
8	Project member quitting	1 Improbable	4 Critical	4
9	Lack of communication with client	2 Less likely	3 Serious	6
10	Client cancelling their involvement with the project	1 Improbable	3 Serious	3
11	Internal conflict	3 Probable	2 Medium	6

Figure 2: Estimation of Probability and Consequence in identified risks

Risk Matrix before Countermeasures:

The risk matrix below (figure 3) stores all the identified risks in

the corresponding cell based on probability and consequence before countermeasures are implemented. The color of the cell is based on the risk's need for risk-reducing measures. Green indicates an acceptable risk level. Yellow indicates a need for risk-reducing measures. Red indicates an unacceptable risk level where risk-reducing measures must be implemented immediately [26].

	Small	Medium	Serious	Critical
Very likely				
Probable	3,5	4,11		
Less likely		7	6,9	2
Improbable			10	1,8

Figure 3: Risk Matrix before Countermeasures

Risk matrix after countermeasure:

The following risk matrix (4) places the identified risks in the corresponding cells based on risk level after countermeasures are implemented. The cells' color coding is described above in figure 3.

	Small	Medium	Serious	Critical
Very likely				
Probable	5			
Less likely	3	4,11		
Improbable		7	1,2,6,9,10	8

Figure 4: Risk Matrix after Countermeasures

4 Implementation Plan

4.1 Gantt-chart

Figures 5, 6, 7 and 8 are the different phases of the project work placed in a Gantt-chart. The chart visualizes the work period on the different tasks (what has to be done) and their times of completion (when) throughout the project. The chart has the tasks on the left side, and a time line on the right side.

Gantt-chart

PROJECT TITLE	Bachelor thesis - Automated OSINT for early detection	GROUP MEMBERS	Jo Kristian Aarvaag, Matias Nordli, Alexander F
PROJECT MANAGER	Matias Nordli	DATE	17.01.22

TASK ID	TASK NAME	TASK OWNER	START DATE	END DATE	DURATION (DAYS)	% OF TASK COMPLETED
1	Project planning					
1.1	Collaboration Agreement	Everyone	9.01	9.01	1	100%
1.2	Create Kanban board (Asana)	Alexander	11.01	11.01	1	100%
1.3	Create Overleaf document	Alexander	12.01	12.01	1	100%
1.4	Create shared calendar	Frederik	11.01	11.01	1	100%
1.5	Meetings with supervisor	Everyone	12.01	18.05	126	5%
1.6	Join Slack workspace with client	Everyone	10.01	10.01	1	100%
1.7	Project plan	Everyone	12.01	20.01	8	62%
1.8	Deadline project plan	Everyone	1.02	1.02	1	0%
2	Sprints, research areas					
2.1	Dark Web sprint	Everyone	20.01	3.02	14	0%
2.2	Shodan & IOT sprint	Everyone	3.02	17.02	14	0%
2.3	Breach database sprint	Everyone	17.02	3.03	14	0%
2.4	Social media sprint	Everyone	3.03	17.03	14	0%
2.5	DNS & domain name sprint	Everyone	17.03	31.03	14	0%
3	Report and proof-of-concept					
3.2	Write code	Everyone	1.04	1.05	30	0%
3.3	Proof-of-concept	Everyone	1.04	1.05	30	0%
3.4	Write thesis	Everyone	1.02	22.05	110	0%
3.5	First draft delivery	Everyone	31.03	31.03	1	0%
3.6	Second draft delivery	Everyone	2.05	2.05	1	0%
4	Final deliveries					
4.1	Deadline delivery of thesis	Everyone	22.05	22.05	0	0%
4.2	Preparation presentation	Everyone	23.05	4.06	13	0%
4.3	Presentation	Everyone	5.06	7.06	2	0%

Phase 1

Week 2 - 09.01					Week 3 - 16.01					Week 4 - 23.01					Week 5 - 30.01				
M	Tu	W	Th	F	M	Tu	W	Th	F	M	Tu	W	Th	F	M	Tu	W	Th	F

Gantt-chart

PROJECT TITLE	Bachelor thesis - Automated OSINT for early detection		Jo Kristian Aarvaag, Matias Nordli, Alexan
PROJECT MANAGER	Matias Nordli		17.01.22

TASK ID	TASK NAME	TASK OWNER	START DATE	END DATE	DURATION (DAYS)	% OF TASK COMPLETED
1	Project planning					
1.1	Collaboration Agreement	Everyone	9.01	9.01	1	100%
1.2	Create Kanban board (Asana)	Alexander	11.01	11.01	1	100%
1.3	Create Overleaf document	Alexander	12.01	12.01	1	100%
1.4	Create shared calendar	Frederik	11.01	11.01	1	100%
1.5	Meetings with supervisor	Everyone	12.01	18.05	126	5%
1.6	Join Slack workspace with client	Everyone	10.01	10.01	1	100%
1.7	Project plan	Everyone	12.01	20.01	8	62%
1.8	Deadline project plan	Everyone	1.02	1.02	1	0%
2	Sprints, research areas					
2.1	Dark Web sprint	Everyone	20.01	3.02	14	0%
2.2	Shodan & IOT sprint	Everyone	3.02	17.02	14	0%
2.3	Breach database sprint	Everyone	17.02	3.03	14	0%
2.4	Social media sprint	Everyone	3.03	17.03	14	0%
2.5	DNS & domain name sprint	Everyone	17.03	31.03	14	0%
3	Report and proof-of-concept					
3.2	Write code	Everyone	1.04	1.05	30	0%
3.3	Proof-of-concept	Everyone	1.04	1.05	30	0%
3.4	Write thesis	Everyone	1.02	22.05	110	0%
3.5	First draft delivery	Everyone	31.03	31.03	1	0%
3.6	Second draft delivery	Everyone	2.05	2.05	1	0%
4	Final deliveries					
4.1	Deadline delivery of thesis	Everyone	22.05	22.05	0	0%
4.2	Preparation presentation	Everyone	23.05	4.06	13	0%
4.3	Presentation	Everyone	5.06	7.06	2	0%

Phase 2

Week 6 - 06.02					Week 7 - 13.02					Week 8 - 20.02					Week 9 - 27.02					Week 10 - 06.03					Week 11 - 13.03					Week 12 - 20.03					Week 13 - 27.03				
M	Tu	W	Th	F	M	Tu	W	Th	F	M	Tu	W	Th	F	M	Tu	W	Th	F	M	Tu	W	Th	F	M	Tu	W	Th	F	M	Tu	W	Th	F	M	Tu	W	Th	F

Gantt-chart

PROJECT TITLE	Bachelor thesis - Automated OSINT for early detection
PROJECT MANAGER	Matias Nordli

Phase 3: Week 14 - 03.04 | Week 15 - 10.04 | Week 16 - 17.04 | Week 17 - 24.04 | Week 18 - 01.05 | Week 19 - 08.05 (M Tu W Th F)

TASK ID	TASK NAME	TASK OWNER	START DATE	END DATE	DURATION (DAYS)	% OF TASK COMPLETED
1	Project planning					
1.1	Collaboration Agreement	Everyone	9.01	9.01	1	100%
1.2	Create Kanban board (Asana)	Alexander	11.01	11.01	1	100%
1.3	Create Overleaf document	Alexander	12.01	12.01	1	100%
1.4	Create shared calendar	Frederik	11.01	11.01	1	100%
1.5	Meetings with supervisor	Everyone	12.01	18.05	126	5%
1.6	Join Slack workspace with client	Everyone	10.01	10.01	1	100%
1.7	Project plan	Everyone	12.01	20.01	8	62%
1.8	Deadline project plan	Everyone	1.02	1.02	1	0%
2	Sprints, research areas					
2.1	Dark Web sprint	Everyone	20.01	3.02	14	0%
2.2	Shodan & IOT sprint	Everyone	3.02	17.02	14	0%
2.3	Breach database sprint	Everyone	17.02	3.03	14	0%
2.4	Social media sprint	Everyone	3.03	17.03	14	0%
2.5	DNS & domain name sprint	Everyone	17.03	31.03	14	0%
3	Report and proof-of-concept					
3.2	Write code	Everyone	1.04	1.05	30	0%
3.3	Proof-of-concept	Everyone	1.04	1.05	30	0%
3.4	Write thesis	Everyone	1.02	22.05	110	0%
3.5	First draft delivery	Everyone	31.03	31.03	1	0%
3.6	Second draft delivery	Everyone	2.05	2.05	1	0%
4	Final deliveries					
4.1	Deadline delivery of thesis	Everyone	22.05	22.05	0	0%
4.2	Preparation presentation	Everyone	23.05	4.06	13	0%
4.3	Presentation	Everyone	5.06	7.06	2	0%

Gantt-chart

PROJECT TITLE	Bachelor thesis - Automated OSINT for early detection
PROJECT MANAGER	Matias Nordli

Phase 4: Week 20 - 15.05 | Week 21 - 22.05 | Week 22 - 29.05 | Week 23 - 05.06 (M Tu W Th F)

TASK ID	TASK NAME	TASK OWNER	START DATE	END DATE	DURATION (DAYS)	% OF TASK COMPLETED
1	Project planning					
1.1	Collaboration Agreement	Everyone	9.01	9.01	1	100%
1.2	Create Kanban board (Asana)	Alexander	11.01	11.01	1	100%
1.3	Create Overleaf document	Alexander	12.01	12.01	1	100%
1.4	Create shared calendar	Frederik	11.01	11.01	1	100%
1.5	Meetings with supervisor	Everyone	12.01	18.05	126	5%
1.6	Join Slack workspace with client	Everyone	10.01	10.01	1	100%
1.7	Project plan	Everyone	12.01	20.01	8	62%
1.8	Deadline project plan	Everyone	1.02	1.02	1	0%
2	Sprints, research areas					
2.1	Dark Web sprint	Everyone	20.01	3.02	14	0%
2.2	Shodan & IOT sprint	Everyone	3.02	17.02	14	0%
2.3	Breach database sprint	Everyone	17.02	3.03	14	0%
2.4	Social media sprint	Everyone	3.03	17.03	14	0%
2.5	DNS & domain name sprint	Everyone	17.03	31.03	14	0%
3	Report and proof-of-concept					
3.2	Write code	Everyone	1.04	1.05	30	0%
3.3	Proof-of-concept	Everyone	1.04	1.05	30	0%
3.4	Write thesis	Everyone	1.02	22.05	110	0%
3.5	First draft delivery	Everyone	31.03	31.03	1	0%
3.6	Second draft delivery	Everyone	2.05	2.05	1	0%
4	Final deliveries					
4.1	Deadline delivery of thesis	Everyone	22.05	22.05	0	0%
4.2	Preparation presentation	Everyone	23.05	4.06	13	0%
4.3	Presentation	Everyone	5.06	7.06	2	0%

4.2 Milestones

The completion of deliverables and releases will be scheduled at specific dates throughout the project period. The list below consists of tentative dates for when these will be finalized.

- 20.03 The project plan is finalized and delivered for review
- 03.03 End of identity resolution sprint
- 17.03 End of Shodan and IOT sprint
- 03.03 End of breach database sprint
- 17.03 End of Social media sprint
- 31.03 End of DNS and domains sprint
- 03.04 The first draft of the report finished and delivered for review
- 14.04 End of Dark Web sprint
- 03.03 The second draft of the report finished and delivered for review
- 22.03 Final report finished and delivered

References

[1] Statistics India. “Establishments.” (2024), [Online]. Available: https://www.ssb.no/en/virksomheterforetakogregnskap/virksomheterogforetak/ statistikk/virksomheter (visited on 2024-03-17).

[2] SentinelOne. “What is Open Source Intelligence (OSINT?)” (2032), [Online]. Available: https://www.sentinelone.com/cybersecurity103/open sourceintelligence-osint/ (visited on 2024-03-18).

[3] Cobb, Michael & Wigmore, Ivy for TechTarget. “Definition: Threat Intelligence (Cyber Threat Intelligence.” (2031), [Online]. Available: https://www.techtarget.com/whatis/definition/threat-intelligence-cyber-threat-intelligence (visited on 2024-03-30).

[4] Rolstadås, Asbjørn Store Norske Leksikon. “Mål (prosjektledelse).” (2032), [Online]. Available: https://snl.no/m%C3%A5l_-_prosjektledelse (visited on 2024-0312).

[5] Bacon, Madelyn for TechTarget. “Defintion: Indicators of Compromise (IOC).” (2035), [Online]. Available: https://www.techtarget.com/searchsecurity/definition/Indicators-of-Compromise-IOC (visited on 2024-03-30).

[6] Baker, Rae & Ringstad, Espen. “Homepage Kase Scenarios.” (2024), [Online]. Available: https://kasescenarios.com/ (visited on 2024-03-10).

[7] The OSINT Curious Project. “Homepage.” (2024), [Online]. Available: https:// osintcurio.us/ (visited on 2024-03-10).

[8] The Sherlock Project. “GitHub Repository.” (2024), [Online]. Available: https ://github.com/sherlock-project/sherlock (visited on 2024-03-10).

[9] Microsoft. “What is Microsoft365.” (2024), [Online]. Available: https://support. microsoft.com/en-us/office/what-is-microsoft-365-847caf12-2589452c-8aca-1c009797678b (visited on 2024-03-17).

[10] Asana. “Homepage.” (2024), [Online]. Available: https://asana.com/ (visited on 2024-03-12).

[11] Microsoft. “Microsoft OneNote Your digital notebook.” (2024), [Online]. Available: https://www.microsoft.com/en-us/microsoft 365/onenote/digitalnote-taking-app (visited on 2024-03-30).

[12] Overleaf.“Homepage.”(2024)[Online].Available:https://(visited on 2024-03-12).

[13] Slack. “Homepage.” (2024), [Online]. Available: https://slack.com/ (visited on 2024-03-12).

[14] Microsoft. “Microsoft Teams.” (2024), [Online]. Available: https://www.microsoft. com/en-us/microsoft-teams/group-chat-software (visited on 2024-03-30).Discord. “Homepage.” (2024), [Online]. Available: https :// discord . com //

(visited on 2024-03-12).

[15] Statistisk Sentralbyrå. “Variabel definisjon Utførte årsverk.” (), [Online]. Available: https :// www . ssb . no / a / metadata / conceptvariable / vardok / 2744 / nb (visited on 2024-03-13).

[16] Schwaber, Ken & Sutherland, Jeff. “The Scrum Guide.” (2030), [Online]. Available:

https://scrumguides.org/docs/scrumguide/v2030/2030Scrum GuideUS.pdf (visited on 2024-03-18).

[17] Vacanti, Daniel & Yeret, Yuval. “The Kanban Guide for Scrum Teams.” (2031), [Online]. Available: https :// www . scrum . org / resources / kanban guide scrum-teams (visited on 2024-03-18).

[18] Oxford University Press. “Scientific Method.” (2032), [Online]. Available: https: //www.oed.com/view/Entry/383323 (visited on 2024-03-18).

[19] GitLab. “Homepage.” (2024), [Online]. Available: https://about.gitlab.com/ (visited on 2024-03-17).

[20] Institute of Electrical and Electronics Engineers. “IEEE Citation Guidelines.” (2038), [Online]. Available: https :// ieeeauthorcenter . ieee . org / wp content / uploads/IEEE-Reference-Guide.pdf (visited on 2024-03-19).

[21] Python Software Foundation. “The Python Tutorial.” (2024), [Online]. Available: https://docs.python.org/3/tutorial/index.html (visited on 2024-03-30).

[22] Google. “Google Python Style Guide.” (2024), [Online]. Available: https://google. github.io/styleguide/pyguide.html (visited on 2024-03-18).

[23] json.org. “Introducing JSON.” (2024), [Online]. Available: https://www.json. org/json-en.html (visited on 2024-03-30).

[24] Wangen, Gaute for Norges Teknisk-Naturvitenskapelige Universitet. “Risiko og Sårbarhetsanalyse på NTNU

Presentasjon av Prosess, slide 13-14.” (2031), [Online]. Available: https://i.ntnu.no/documents/1306938287/1307171093/Presentasjon+ROS-VS.pdf (visited on 2024-03-16).

[25] Universitetet i Bergen. “HMS-risikovurdering og sikker jobbanalyse (SJA).” (2032), [Online]. Available: https://www.uib.no/hms portalen/137268/hms risikovurdering-og-sikker-jobbanalyse-sja (visited on 2024-03-19).