SAFETY4RAILS

CYBER-PHYSICAL THREAT DETECTION WITH CAPABILITIES MATRIX INTELLIGENCE

Deliverable 4.3

Lead Author : INNO

Contributors: TREE, IC, CS, ERARGE, ICOM, UMH

Dissemination level: Public

Security Assessment Control: passed

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 883532.

D4.3 CYBER-PHYSICAL THREAT DETECTION WITH CAPABILITIES MATRIX INTELLIGENCE		
Deliverable number:	D4.3	
Version:	1.0	
Delivery date:	27/05/2021	
Deliverable due date:	31/05/2021	
Dissemination level:	Public	
Nature:	Report	
Main authors:	Marco Tiemann Lesley Badii	INNO INNO
Contributor(s):	Alejandro Prada Nespral Zsuzsanna Keri Eros Cazzato	TREE CS IC
Internal reviewer(s):	Eros Cazzato Stephen Crabbe Nacho Diaz Andreas Georgakopoulos Atta Badii Uli Siebold Antonio De Santiago Laporte	IC Fraunhofer UMH WINGS READ IC MdM
External reviewer:	Reto Biedermann Ryan Faulkner	IC INNO

Document control			
Version	Date	Author(s)	Change(s)
0.1	15/03/2021	Marco Tiemann	ToC release
0.2	19/04/2021	Marco Tiemann Lesley Badii	Initial contributions
0.3	26/04/2021	Marco Tiemann Alejandro Prada Nespral	Overall updates TISAIL description
0.4	10/05/2021	Marco Tiemann Alejandro Prada Nespral Zsuzsanna Keri	Revisions after internal reviews
0.5	11/05/2021	Marco Tiemann	Revisions after internal reviews
1.0	27/05/2021	Marco Tiemann	Further revisions after external reviews

DISCLAIMER AND COPYRIGHT

ABOUT SAFETY4RAILS

SAFETY4RAILS is the acronym for the innovation project: **Data-based analysis for SAFETY and security protection FOR detection, prevention, mitigation and response in trans-modal metro and RAILway networkS**. Railways and Metros are safe, efficient, reliable and environmentally friendly mass carriers, and they are becoming even more important means of transportation given the need to address climate change. However, being such critical infrastructures turns metro and railway operators as well as related intermodal transport operators into attractive targets for cyber and/or physical attacks.**The SAFETY4RAILS project delivers methods and systems to increase the safety and recovery of track-based inter-city railway and intra-city metro transportation.** It addresses both cyber-only attacks (such as impact from WannaCry infections), physical-only attacks (such as the Madrid commuter trains bombing in 2014) and combined cyber-physical attacks, which are important emerging scenarios given increasing IoT infrastructure integration.

SAFETY4RAILS concentrates on rush hour rail transport scenarios where many passengers are using metros and railways to commute to work or attend mass events (e.g. large multi-venue sporting events such as the Olympics). When an incident occurs during heavy usage, metro and railway operators have to consider many aspects to ensure passenger safety and security, e.g. carry out a threat analysis, maintain situation awareness, establish crisis communication and response, and they must ensure that mitigation steps are taken and communicated to travellers and other users. **SAFETY4RAILS will improve the handling of such events through a holistic approach.** It will analyse the cyber-physical resilience of metro and railway systems and deliver mitigation strategies for an efficient response, and, in order to remain secure given everchanging novel emerging risks, it will facilitate continuous adaptation of the SAFETY4RAILS solution; this will be validated by two rail transport operators and the results will support the re-design of the final prototype.

TABLE OF CONTENTS

List of tables

List of figures

Executive summary

This deliverable report, D4.3 “Cyber-Physical Threat Detection with Capabilities Matrix Intelligence”, reports on the initial work that has been carried out in Task 4.2 ”OSINT Technologies for Cyber-Physical Intelligence (SCADA)”. The report encompasses both work towards inventorisation of threats with related modeling and work towards the development and integration of tools for gathering and storing threat data from OSINT sources as laid out in the Description of Action (DoA).

The deliverable introduces the concept of Open Source Intelligence (OSINT) in Section 2. Section 3 describes the design of the OSINT system designed in the project. Section 4 discusses data modelling, data sources and data acqusition that are relevant to the activities that are reported on. Section 5 concludes the deliverable with a summary and a matrix that relates the specified data-related and development activities with the functionalities specified in Section 2 of the deliverable.

1. Introduction

This introductory section provides an overview of the activities in task T4.2 on which the deliverable reports and outlines the structure and contents of the present deliverable D4.3.

1.1 Activity Overview

This document reports on the initial activities undertaken in task T4.2 “OSINT Technologies for Cyber-Physical Intelligence (SCADA)”. We therefore first introduce the activities, goals and means that are important for this task. This will provide the necessary understanding for the remainder of this deliverable.

The work reported on in this deliverable is concerned with open source intelligence, which can be generally described as analytics using openly available data (see Section 2.1 for a more differentiated definition). The main aims of this task are to gather, process and make accessible open source intelligence that relates to cyber-physical vulnerabilities and threats of concern within the scope of the SAFETY4RAILS project and of course to those who may benefit from these activities beyond the project consortium. Achieving these aims involves the following main activities:

1. Specification of a data model that suitably represents the application domain in terms of components, their potential vulnerabilities and past and present threats that are relevant related to the components under consideration. The data model should be suitable for representing cyber, physical and cyber-physical data in line with the overall project ambition.
2. Creation of a database implementation using the developed data model; the database implementation should support typical database functionalities and should also readily integrate into the overall SAFETY4RAILS system infrastructure, for instance by supporting the messaging system used in the project.
3. Development of data gathering components that collect relevant open source intelligence data from web sources. This should include commonly used sources such as malware repositories and threat intelligence feeds that are relevant for the application domain as well as less formal sources including social media feeds[1]. The gathered data should be processed so that the extracted information can be stored in the database using the data model developed for the project. The data gathering process should be automated as far as possible while maintaining a suitable level of specificity and relevance for the application domain that minimises risks of information overload by users of the database.
4. Development of alerting and simple statistics processing functionalities in order to create necessary alerting functionalities for vulnerable components and in order to generate overview summaries of key risks in the application domain for more strategic review and reaction purposes.
5. Populating the database with component, vulnerability and initial threat data in order to bootstrap the system as well as facilitate testing, demonstration and evaluation activities.

Two deliverable documents are part of the work in the task, this deliverable and the deliverable D4.2 “Framework and Methodology of Critical Components Based on OSINT”, due in project month 14. Furthermore, the task contributes to the data modelling and software development activities in the project.

1.2 Organisation of this Deliverable

The main aims of this deliverable are to document the methods for cyber-physical threat detection using open source intelligence that the task is concerned with throughout the overall task duration. The deliverable

[1] In the project, we will investigate the usage of social media sources operated by organisations concerned with publishing information on vulnerabilities and additionally sources created by us that simulate private user social media feeds. With these approaches we intend to avoid gathering any real personal data from social media sources. We will liaise with our project ethics partners and review this approach prior to, during and after implementation.

furthermore summarises the contributions and focus areas of the individual components and other contributions in the task through a capability matrix for open source intelligence.

The deliverable is organised into the following sections:

- Section 2 “Open Source Intelligence” introduces the concept of open source intelligence generally as well as specifically in the context of SAFETY4RAILS. The section presents the main functionalities to be provided via open source intelligence in SAFETY4RAILS.
- Section 3 “Open Source Intelligence System Design” describes the specification of the design chosen for implementing the OSINT system in SAFETY4RAILS. It also introduces the relevant background information on tools that are included as part of the design.
- Section 4 “Data Model, Data Sources & Data Acquisition” is concerned with a general description of the data model needed for task T4.2 (including the relation to data models developed in other tasks of the project), with the identification of data sources for the task and with the process of data acquisition required for the execution of the task activities during the project.
- Section 5 “Conclusion” briefly summarises the contents of this deliverable and looks forward towards the second deliverable related to task T4.2. The section concludes with a matrix that relates the task activities and outcomes to the functionalities discussed in Section 2 of the deliverable.

A bibliography and an annex with glossary and acronym explanations are appended at the end of this deliverable.

2. Open Source Intelligence

This section introduces the field of Open Source Intelligence in general as well as in terms of how it is applied within SAFETY4RAILS. Based hereon, the main functionalities of open source intelligence for SAFETY4RAILS are specified as a guiding framework for the research and innovation efforts described here.

2.1 Introduction

The term “open source intelligence” (OSINT) is a term that originated and is frequently used in the intelligence community (including military intelligence). In this context, (1) defines it as “unclassified information that has been deliberately discovered, discriminated, distilled and disseminated to a select audience in order to address a specific question.” Further definitions that also focus on a definition from a military perspective as can be found for instance in (2) generally define OSINT similarly with some variations in terms of the scope of what is included in under “open source” (publicly available, legally available, unclassified information, etc.). (3) more generally defines OSINT as “all information that can be derived from overt collection”.

As the described activities are not undertaken in a military context, the following more generic definition in (2) is useful:

> “Open-source intelligence (OSINT) is a multi-factor […] methodology for collecting, analyzing and making decisions about data accessible in publicly available sources to be used in an intelligence context. In the intelligence community, the term ‘open’ refers to overt, publicly available sources (as opposed to covert or clandestine sources).”

The term “open source” in the context of OSINT is in the remainder used to denote publicly available sources of potential intelligence value, and “intelligence” is defined as the process of collecting, analysing and using the collected information in order to generate useful information for use by others.

2.2 Open Source Intelligence in SAFETY4RAILS

What is the rationale for and what is the purpose of having open source intelligence as a part of the activities in SAFETY4RAILS?

Based on the general definition above, OSINT in SAFETY4RAILS should be used to gather and evaluate “open source” data in order to support the goals and functionalities envisioned for the project, including identifying weaknesses in live systems, the automation of systems dedicated to monitoring cyber vulnerabilities and supporting the prevention of risks and threats as identified in deliverable D2.1 “Grid Analysis of End-User Needs and Workshop Minutes”. One key aspect of the project is the ambition to cover a wide range of cyber, physical and combined cyber-physical systems in order to increase their security as well as overall rail infrastructure safety. Open source intelligence can provide valuable information in these areas by identifying vulnerabilities, threats and, where feasible from open source data, specific information on vulnerabilities or suspicious behaviour identified in publicly accessible systems and services.

It is useful to take note of current typical real-world usage of open source intelligence, particularly mediated by technological solutions such as the ones proposed in SAFETY4RAILS, separately for cyber, physical and combined cyber-physical security in general as well as in the railway domain. The following table summarises the view of the task participants based on their expertise in the field.

TABLE 1: USE OF OSINT IN CYBER, PHYSICAL AND CYBER-PHYSICAL SCENARIOS

Domain	*General Use*	*Rail Operator Use*
Cyber security	OSINT is used fairly widely, sources such as malware repositories or threat data feeds are available for general IT infrastructure and also for Supervisory Control and Data Acquisition	Railway use of OSINT in the cyber security domain is generally comparable with OSINT use in industrial sectors; standard not domain-specific threat repository

	(SCADA) devices of major manufacturers; standard data sources, organisational structures and software tools for managing threat repositories are available	systems are used by some operators; security matters may be outsourced to third parties
Physical security	OSINT is not used widely in order to identify physical threats such as for instance natural hazards from open source data outside of the security domain	Little to no use of OSINT in order to identify physical threats in the railway domain
Cyber-physical security	OSINT usually only used by more sophisticated operators in the security domain	No use of OSINT in the cyber-physical domain; cyber and physical risks are not usually managed by the same departments or responder groups

Generally, OSINT tools used in the railway domain focus on cyber security and are not specialised for use in the railway domain. The goal of developing open source intelligence tools in SAFETY4RAILS will therefore include a) the development of specialised solutions for the railway domain in terms of data gathering and processing, b) extending the coverage of OSINT technologies to also cover physical and where practical cyber-physical security and c) automating data processing for OSINT acquisition and processing as far as possible given the quality of real-world data sources available. These specific goals can be defined in more detail as functionalities to be provided by OSINT for use in SAFETY4RAILS.

2.3 Open Source Intelligence Functionalities in SAFETY4RAILS

In order to be useful for the type of end user organisations addressed by SAFETY4RAILS, an OSINT system must target the specific potential sources of risks, such as specific deployed component models that are used by an end user organisation, in order to ensure that intelligence is gathered for the relevant infrastructure and environment. The availability of an inventory of systems and components to be included into OSINT processes is therefor a prerequisite for the targeted use of OSINT in SAFETY4RAILS. These data are used to customise search processes and filter out irrelevant data from OSINT analyses; the OSINT system should provide an easy to use API to add and update this information as changes are made to the railway operator infrastructure[2].

The data processing functionalities that should be provided via OSINT are derived from the requirements specified for the OSINT functionalities in SAFETY4RAILS, set out in deliverable D1.4. Here, we organise the resulting functionalities by first specifying a processing pipeline for the OSINT system in SAFETY4RAILS and then describing the functionalities to be provided at each stage of the processing pipeline.

FIGURE 1: SAFETY4RAILS OSINT PROCESSING PIPELINE

[2] Section 4.1 describes the integration of such data from a data model perspective and Section 4.2.3 explains initial steps for creating a data model to use during development. It is expected that for component data as well as taxonomy data for later stages of development, demonstration and evaluation will be integrated from other tasks in the project (see also Section 4.1.3). The system API depicted in Figure 3 exposes methods to create, read, update and delete relevant entries in the MISP database.

This processing pipeline is applied for each data processing activity. It is envisioned that the execution of this processing pipeline is a frequently recurring activity that is either carried out in regular intervals, triggered by notifications of available data updates or activated by changes in the infrastructure data entered by railway operators.

We describe each of the processing steps together with the associated functionalities to be provided in the table below.

TABLE 2: FUNCTIONALITIES ORGANISED BY PROCESSING PIPELINE STEPS

Processing Step	*Description (in italics) and functionalities*
Data Acquisition	*This processing step encompasses all activities that involve retrieval of OSINT data (push and pull)*
	FUNC-DA-01: Retrieval from typical cyber security data sources such as specialised search engines and data feed; data sources and retrieval from searches are preselected from data sources considered relevant for the domain; potentially retrieval from social media sources such as cyber security Twitter feeds
	FUNC-DA-02: Search for and identification of vulnerable and potentially compromised devices that are accessible on the Internet and relevant for the application domain
	FUNC-DA-03: Retrieval from relevant physical security data sources including specialised providers, general news feeds and potentially social media sources such as Twitter posts related to relevant tags
Pre-Processing and Analytics	*This processing step encompasses necessary activities for parsing the acquired data and for attempting to extract relevant information from any data that may not be sufficiently well structured for relatively simple parsing*
	FUNC-PP-01: Parsing of standard format data feeds and standard format search query results from specialised data sources
	FUNC-PP-02: Analytics and entity extraction from semi-structured and unstructured data sources such as free text and social media communications in order to identify potentially relevant data
Storage and Representation	*This processing step involves storing the processed data in a database while adhering to the database structure*
	FUNC-SR-01: Operation of a suitable database management system or similar infrastructure that provides typical utility functionalities
Data Set Analytics	*This processing step involves all analytics activities that are carried out on the database instead of a single data point prior to addition to the database*
	FUNC-DS-01: Analysis of newly added data points in order to identify new vulnerabilities or threats that are identified based on the newly added data point

	FUNC-DS-02: Generation of statistics in order to identify trends and rankings of threats and vulnerabilities
Data Access and Messaging	*This processing step consists of responding to data retrieval requests via an API and of messaging updates to components that expect notifications of particular state changes in the database*
	FUNC-DM-01: Exposing data access functionalities to authorised system components within SAFETY4RAIL
	FUNC-DM-02: Communication of data updates to components that require data to be pushed to them

These functionalities will be revisited and related to the specific implementation tasks that are defined in the next section as well as to the data sources that are described later in this deliverable in order to specify how the defined functionalities are realised in the project.

3. Open Source Intelligence System Design

This section describes the design of the software system that forms part of the overall solution created in SAFETY4RAILS. First, relevant development background that influences the system design is introduced. Then, the components that form part of the system design are introduced and the system architecture is described including the integration into the overall SAFETY4RAILS system.

3.1 Background

One important way in which the SAFETY4RAILS project accomplishes its objectives is by leveraging existing specialised software solutions and integrating them into an overarching system that is tailored to the target application domain. The majority of SAFETY4RAILS technical tasks hence involve the integration of and extension and/or customisation of an existing software solution. Task T4.2 includes the integration of TISAIL, a threat intelligence solution for the railway sector developed by TREE Technology. We introduce TISAIL here as background to help understand the system design. We also briefly introduce the Malware Information Sharing Platform (MISP[3]) that is used both in TISAIL and in the OSINT system design.

3.1.1 TISAIL

TISAIL is a threat intelligence platform for the railway sector that is being developed by SAFETY4RAILS project partner TREE Technology. Threat intelligence is here defined as follows (4):

> "Threat intelligence (TI) is evidence-based knowledge (e.g. context, mechanisms, indicators, implications and action-oriented advice) about existing or emerging menaces or hazards to assets. It can be used to inform decisions regarding the subject's response to that menace or hazard." (Gartner)

TISAIL incorporates three different stages as part of the threat intelligence process:

1. Using automated processes for discovering potential threats using threat intelligence feeds, malware repositories, vulnerability reports and detection rules.
2. Carrying out malware analysis processes.
3. Extracting Indicators of Compromise (IoC) and enriching the gathered information in order to generate threat data and notifications for use by other systems.

TISAIL uses the MISP (5) data model for data representation and uses MISP internally for data storage. A domain-specific taxonomy for threats is used to help decision makers identify and classify threats and take suitable actions more quickly. Figure 2 illustrates the general operation of TISAIL.

[3] More information on MISP is available on the project website at https://www.misp-project.org (accessed 19.04.2021).

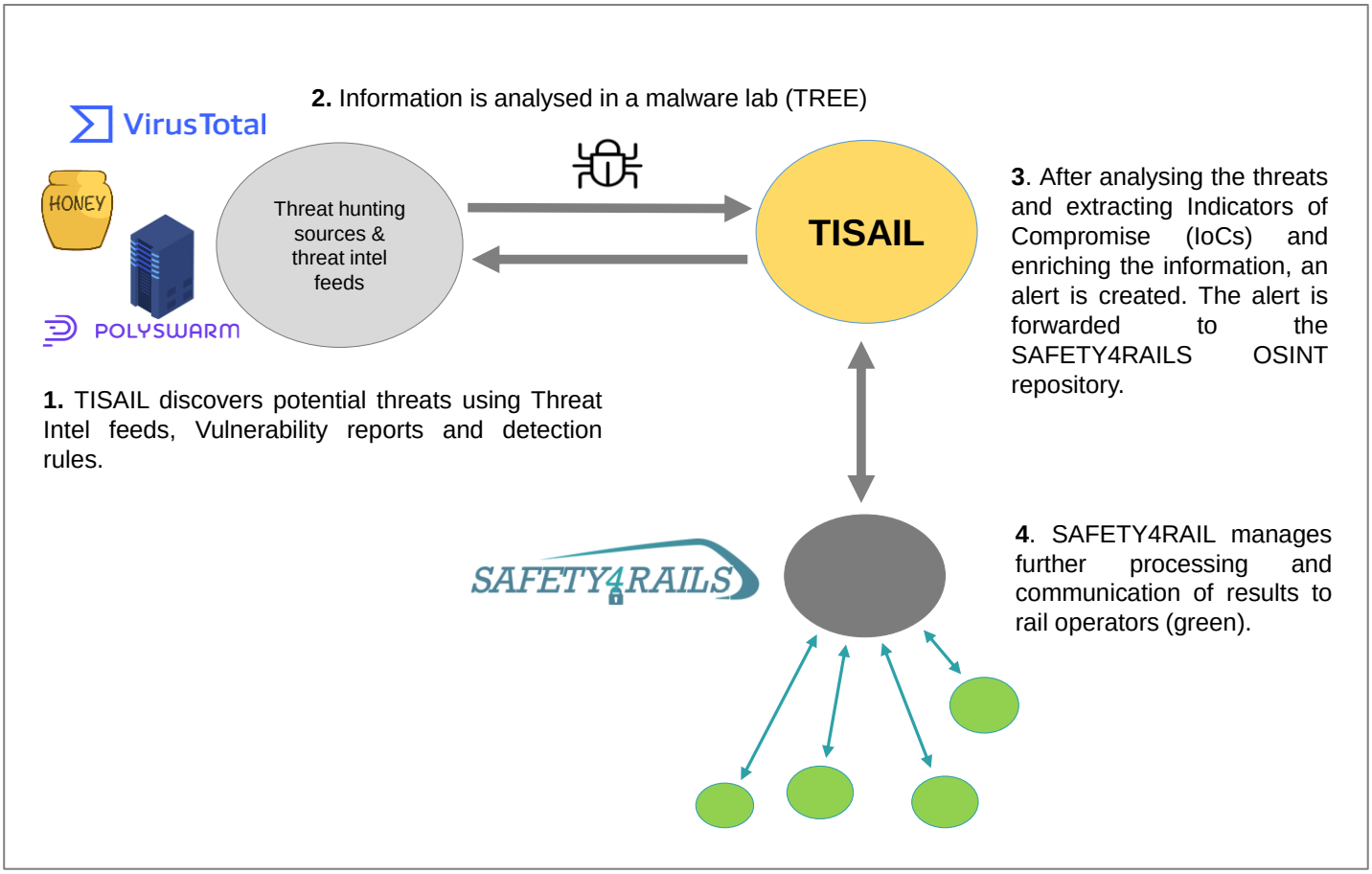

FIGURE 2: TISAIL PROCESSING OVERVIEW

The format of the TISAIL alerts is based on the MISP standard format that is used to exchange threat information among MISP instances. The MISP format is JSON-based. It includes different standards for facilitating the representation of technical and non-technical information about malware, cyber-attacks and threat actors. MISP uses several data models for different purposes; TISAIL uses the MISP core format, the MISP object template and the MISP taxonomy format for communicating information. Details of each standard are explained in Section 4.1.

The communicated threat information may include IoCs, malicious file indicators or event information about a threat actor, such as its Tactics, Techniques and Procedures (TTPs).

TISAIL is a proprietary solution developed by TREE Technology, and the TISAIL infrastructure operates in a secured Amazon Web Services (AWS) environment in order to protect the gathered data. This is relevant for the design of the overall solution in SAFETY4RAILS because other partners in the project will not be able to directly access TISAIL, and they will not be able to directly integrate software components with TISAIL itself. Since TISAIL uses the MISP data format and MISP repository as part of the overall solution, a suitable solution to address this limitation is to deploy a SAFETY4RAILS MISP instance with which TISAIL can communicate natively and with little additional effort.

3.1.2 Malware Information Sharing Platform

The Malware Information Sharing Platform (MISP) is an open source, community-driven platform for the exchange and sharing of threat intelligence and IoCs about targeted malware and attacks. The development of the platform was supported by NATO and is currently used widely in order to store and exchange threat data

within trust groups, especially in the cyber security domain. Beyond its applicability in cyber security, MISP provides a number of useful features for the scope of SAFETY4RAILS out of the box, including the following[4]:

- A widely used threat data model that is expressed in JSON format and can be customised for domains beyond the typical cyber security application domain of MISP;
- Integration of a database management system, a standard REST API, data synchronisation mechanisms as well as the ability to export threat data for direct import into commonly used intrusion detection systems (IDS) and custom tools (in the following formats: STIX, OpenIOC, plain text, CSV, XML, JSON);
- A correlation finding component that can automatically identify relationships between attributes and indicators from malware or data submitted to the MISP instance. This component was also extended with an advanced visualisation tool in the most recent rounds of development, to support analysis work within MISP. The cyber threat analysis process is further supported by the availability of a MITRE ATT&CK matrix integration;
- A flexible tool for importing and integrating MISP and OSINT data feeds using standard data formats with typically used feeds being available as part of a default installation. Further feeds can be integrated for free or on a subscription[5] basis, mostly offered by nation-level, CERT/CSIRT organisations or private cyber threat intelligence providers;
- Well-defined expansion points and available expansion modules that enable customisation of MISP functionalities and for the easy integration and use of expansion modules that are made publicly available by the large user community.

MISP development has been co-funded by the European Union through the Connecting Europe facility. The core MISP platform and already developed modules are released under a GPLv3 licence (6), which means that modifications to the core software or modules must also be published as open source using the same licence. This does not apply to customised developments that use MISP APIs and similar software development interfaces[6].

3.2 Components

This subsection specifies the components that have been identified as necessary for the OSINT system design. They are briefly described in the tables included in each of the following subsubsections and presented in a structured form in Section 3.3. The subsubsections are organised analogously to the data processing pipeline structure shown in Section 2.

3.2.1 Data Acquisition

Data acquisition components acquire data from open sources that may be relevant in order to provide intelligence concerning vulnerabilities, threats and general risks that are relevant to the application domain and can be gathered from open sources. Table 3 lists the currently specified data acquisition components. Please note that the term components is used in a wide sense and also includes specific system configurations (for instance for the retrieval of threat intelligence feed data). The table reflects the state of planning at the time of finalisation of this deliverable.

[4] Please see https://www.misp-project.org/features.html (accessed 19.04.2021) for a more complete listing of MISP platform features.
[5] Subscription services may or may not be considered to be OSINT depending on the definition applied. Our aim in SAFETY4RAILS is to avoid subscription services within the context of the project, but to support their use in general for post-project deployment scenarios.
[6] See https://www.misp-project.org/license/ (accessed 19.04.2021) for the MISP description of the applicability of the GPL license for MISP.

TABLE 3: DATA ACQUISITION COMPONENTS

ID	*Title*	*Description*
DA-TIS-01 (TISAIL)	**Internet-exposed asset crawlers**	Development of a set of crawlers that search for IT/OT assets/components that are exposed on the Internet and gather information about them for correlation with a list of railway keywords as well as a list of products used by railway companies. In case of a match, the exposed asset will be stored and an alert will be created.
DA-TIS-02 (TISAIL)	**Vulnerability and exploit scanning crawlers**	Development of a set of crawlers that search for known vulnerabilities and exploit ICS assets/components that have been defined as relevant in the railway domain
DA-TIS-03 (TISAIL)	**Malware repository crawlers with Yara[7] rules**	Development of Yara rules for tracking malware families targeting ICS/SCADA Systems. The rules will be deployed in malware databases such as MalwareBazaar.
DA-TIS-04 (TISAIL)	**Malware social media crawlers with Yara rules**	Development of a set of crawlers that will retrieve IoC, TTPs and context about malware and Threat actors from binaries from Threat Intel feeds and social media sources (e.g., Twitter)
DA-TIS-05 (TISAIL)	**Phishing campaign monitoring crawlers**	Development of a crawler that monitors domain names for alterations of railway domain names, detecting potential phishing campaigns
DA-TIS-06 (TISAIL)	**Threat intel feed selection and configuration**	Selection and configuration of relevant threat intelligence feeds and social media that provide relevant information for processing in TISAIL
DA-MSP-01	**Threat intel feed selection and configuration for cyber threats**	Selection and configuration of domain-relevant threat intelligence feeds that relate to cyber security threats relevant to the railway domain and the particular system configurations of concern
DA-MSP-02	**Structured data source integration for physical threat data**	Development of crawlers that retrieve data concerning physical threat detection from structured data sources in well-defined scenarios (natural hazards, public safety emergencies) where those data source are available as open sources
DA-MSP-03	**Social media data source integration for physical threat data**	Development of crawlers that retrieve data concerning physical threat detection from social media sources (either real social media sources or simulated data sources)

[7] See https://blog.malwarebytes.com/security-world/technology/2017/09/explained-yara-rules/ (accessed, 05/05/2021) for a description of Yara rules.

Identifiers for data acquisition components contain “TIS” when they are to be implemented as part of TISAIL in order to differentiate them from components that are not implemented directly as part of that system. Identifiers for components contain “MSP” when they are to be implemented as part of the project MISP instance.

3.2.2 Pre-Processing and Analytics

Data Pre-processing and analytics components handle incoming data and process in order to select, convert, verify or otherwise evaluate or convert the original data into data that are suitable for input into the system database and relevant for use in the application domain. Table 4 lists the currently specified pre-processing and analytics components.

TABLE 4: PRE-PROCESSING AND ANALYTICS COMPONENTS

ID	*Title*	*Description*
PA-MSP-01	**Threat intel feed automated relevance filtering**	Development of a filter rule system to filter out threat intelligence that does not relate to assets/components that has been declared as being in use by railway domain partners
PA-MSP-02	**Structured physical threat automated data feed filtering**	Development of a filter rule system to filter out structured threat data that does not relate to the relevant assets declared by railway domain partners
PA-MSP-03	**Semantic social media data parsing for physical threat detection**	Extraction of entities from social media messages in order to identify messages that may relate to the relevant assets declared by railway domain partners

Identifiers for data acquisition components contain “TIS” when they are to be implemented as part of TISAIL in order to differentiate them from components that are not implemented directly as part of that system. Identifiers for components contain “MSP” when they are to be implemented as part of the project MISP instance.

3.2.3 Storage and Representation

Storage and data representation components manage and model the data gathered and processed in data gathering and pre-processing steps. Table 5 lists the currently specified storage and data representation components.

TABLE 5: STORAGE AND REPRESENTATION COMPONENTS

ID	*Title*	*Description*
SR-TIS-01 (TISAIL)	**Data repository operation**	Secure operation of the TISAIL data repository for data persistence and management
SR-TIS-02 (TISAIL)	**Data model customisation**	Customisation of the TISAIL data model in order to suitably represent relevant threats (based on existing data model used with TISAIL)
SR-MSP-01	**Data repository operation**	Secure operation of the SAFETY4RAILS MISP repository for data persistence and management

SR-MSP-02	**Data model customisation**	Development of an integrated data model for cyber, physical and cyber-physical domain elements included assets/components modelled as MISP objects with taxonomy linkage

Identifiers for data acquisition components contain “TIS” when they are to be implemented as part of TISAIL in order to differentiate them from components that are not implemented directly as part of that system. Identifiers for components contain “MSP” when they are to be implemented as part of the project MISP instance.

3.2.4 Data Set Analytics

Data set analytics components compute analytics over the overall dataset available in the open source intelligence database. This includes processing when specific new or new types of data points are added and processing in specified intervals in order to e.g. generate overall statistics or ranking data. Table 6 lists the currently specified data set analytics components.

TABLE 6: DATA SET ANALYTICS COMPONENTS

ID	*Title*	*Description*
DS-MSP-01	**Rule evaluation over database triggered by specific conditions**	Integration of a rule engine or similar mechanism that enables the evaluation of rules when specific changes to the database have been detected
DS-MSP-02	**Generation of summary statistics**	Development of a component that generates summary statistics for threats and vulnerabilities including the generation of ranked lists of threats and vulnerabilities for a given time period

Identifiers for data acquisition components contain “TIS” when they are to be implemented as part of TISAIL in order to differentiate them from components that are not implemented directly as part of that system. Identifiers for components contain “MSP” when they are to be implemented as part of the project-specific MISP instance.

3.2.5 Data Access and Messaging

Data access and messaging components enable access to the database containing gathered and processed OSINT data either upon request or proactively through the message broker when so specified. Table 7 lists the currently specified data access and messaging components.

TABLE 7: DATA ACCESS AND MESSAGING COMPONENTS

ID	*Title*	*Description*
DM-TIS-01	**Synchronisation mechanism with SAFET4RAILS MISP instance**	Integration of a synchronisation mechanism that synchronises data generated by TISAIL into the SAFETY4RAILS MISP instance
DM-MSP-01	**Custom API endpoint development for typical complex queries**	Development of a component that operates an extended MISP API endpoint with convenience methods that reduce the complexity of requests for frequent complex API client requests

DM-MSP-02	**Message broker integration for SAFETY4RAILS system communication**	Integration and configuration of a client that integrates MISP with the project-wide message broker system

Identifiers for data acquisition components contain “TIS” when they are to be implemented as part of TISAIL in order to differentiate them from components that are not implemented directly as part of that system. Identifiers for components contain “MSP” when they are to be implemented as part of the project MISP instance.

3.3 Architecture

The system architecture of the OSINT system can overall be derived straightforwardly from the processing pipeline structure and the division of system components into TISAIL and generic MISP components. Figure 3 presents a conceptual component diagram of the overall OSINT system.

The diagram shows a conceptual view where components are labelled with the component ID given in the component description tables in Section 3.2. Multi-component diagram elements are labelled with non-specific identifiers such as “DA-TIS-XX” in order to denote that all of the types that belong to the family of components denoted by the remainder of the ID (e.g. DA-TIS-01, DA-TIS-02, …) are shown there as a group. Components that are not software components are added to the diagram as notes (these are data models and the message integration of the TISAIL repository and the MISP repository). The “open source world of available data” is indicated using the cloud symbol outside of the overall OSINT system border.

Please note that the TISAIL equivalents of the “PA-MSP-XX” and “DS-MSP-XX” components for the TISAIL system are also part of the overall OSINT system, but are already integrated as part of the TISAIL repository and are hence not listed here as separately developed or integrated components.

The architecture integrates the proprietary TREE TISAIL solution into the overall system architecture in a simple and practical way that enables TREE to securely host, use and further develop their system as part of the project work and that allows the other technical partners in the project to develop and contribute their own IP and developments to the project through the project MISP repository. Additional overhead caused by this remains minimal by using MISP instance synchronisation between the TISAIL repository and the project MISP instance.

The API interface and Broker Client extend outside of the boundary of the OSINT system in order to indicate that the system exposes a single API facade to the outside world and integrates with the designated message broker system that is envisioned to be used in the SAFETY4RAILS system. It is expected that API access to the OSINT API from within the SAFETY4RAILS system will exclusively involve queries for data retrieval and not queries for creating, writing or updating data stored in the data repository. This may be revised in future, particularly in order to allow for updating information on railway operator infrastructure assets, which are required in order to be able to gather data on vulnerabilities and threats for the correct types and individual assets deployed by specific railway operators.

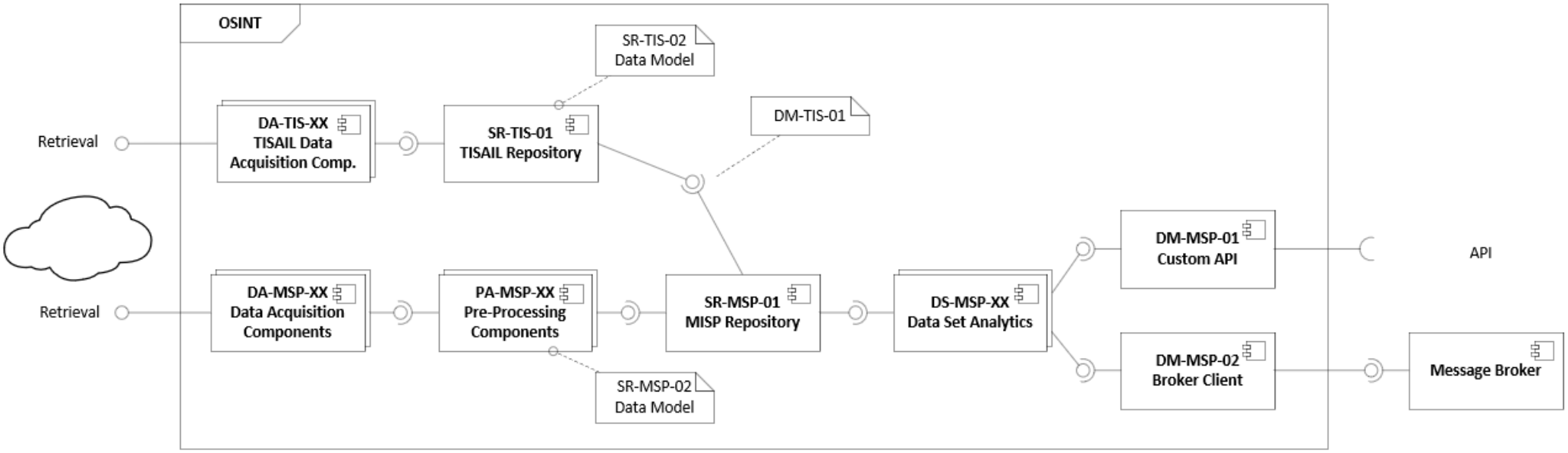

FIGURE 3: OSINT SYSTEM ARCHITECTURE COMPONENT DIAGRAM

The expected consumers of the OSINT data processing activities carried out in the system within SAFETY4RAILS are primarily Task T4.4 concerned with predictive modelling of cascading effects on interconnected infrastructures, Task T4.5 concerned with the implementation of real-time monitoring, and the overall S4RIS information system that is managed in WP6. More detailed specifications regarding this can be found in deliverable D1.4 “Specification of the Overall Technical Architecture” (please note that this is a document with limited public access). Finally, it should be noted that while specific consuming systems are envisioned in the project and their needs are considered during the development of the OSINT system, the system remains open and accessible by all systems that are part of the overall SAFETY4RAILS infrastructure.

We close with some remarks concerning the practical implementation of the system in terms of use of programming languages and similar matters. The relevant boundary conditions for the technical implementation of the OSINT system are primarily determined by the interfaces and access functionalities provided by the MISP system. MISP provides both a REST API and an API access wrapper library written in Python. In addition, MISP uses standard backend systems including a MySQL database backend to which developers can gain direct access in order to carry out modifications that are not supported by the MISP system itself.

For the integration of TISAIL and the SAFETY4RAILS MISP instance, the standard instance data synchronisation features of MISP can be used in order to propagate relevant data from TISAIL to the MISP instance as it becomes available. On the reverse, the MISP instance will synchronise data about assets that are necessary for TISAIL to identify the relevant vulnerabilities and threats for specific railway operators and their infrastructure.

4. Data Model, Data Sources & Data Acquisition

This section introduces the abstract data model that defines the key entity types that are relevant in the context of SAFETY4RAILS OSINT. Furthermore, this section lists basic data sources and data acquisition requirements that need to be met within the project in order to enable the task to successfully carry out the defined activities.

4.1 Data Model

The “world of concern” for SAFETY4RAIL generally involves both cyber and physical assets and their security. The purpose of open source intelligence in the project is to improve the security of these assets by gathering information about potential vulnerabilities, past, present and future threats and to identify assets that have been compromised by a threat where that is possible using open source data.

4.1.1 A General Data Model

Two specific challenging properties of this problem are the inclusion of cyber and physical security-relevant open source intelligence within a single system and the aim of covering as much as possible of the highly diverse and multi-faceted cyber and physical infrastructure that exists within railway operators. Because of the vast and diverse array of possible assets and related security issues, we view the data to be gathered via OSINT from a high-level perspective at theoretical level. Figure 4 illustrates this simple overall model.

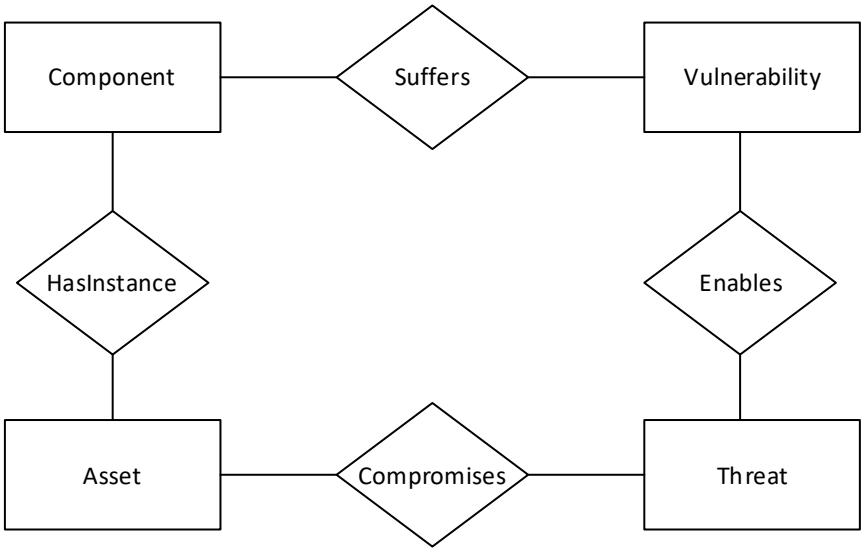

FIGURE 4: HIGH-LEVEL ER-DIAGRAM OF OSINT ENTITIES OF CONCERN

In this diagram, we reduce the view of our “world of concern” to the following four entities:

1. Components are specific and uniformly describable classes of “things” that are relevant within the context of railway security. Examples for a component are for instance a specific model of a SCADA component used within the railway infrastructure.
2. Assets are specific instances of components that are uniquely identifiable and are deployed in or potentially connected to the railway infrastructure of concern. Examples for an asset would be an actual SCADA system deployed within a railway infrastructure.
3. Vulnerabilities describe known risks of compromise or weaknesses that can be exploited by threats. Vulnerabilities can be defined at different levels: a direct connection to the Internet in itself could for instance be considered to be a vulnerability of a system.
4. Threats describe potential attacks or other types of threats to assets that may cause adverse effects to assets with relevant vulnerabilities.

This basic model is an example of a threat-driven data model (7) (8). It can be extended with representations of adversaries and effects of a threat (which can be referred to as an attack when it is executed) and related to observable events, attacks and/or incidents.

For the purpose of SAFETY4RAILS, our initial focus will be on profiling assets (and their generic super class of components) in terms of vulnerabilities and on identifying threats that may impact assets via component-level vulnerabilities. Our practical starting point for implementing a useful data model and OSINT data gathering system is MISP with its threat reporting model and a highly customisable data modelling environment.

4.1.2 MISP Data Models

MISP is primarily a system that receives and processes messages that provide information on (mostly cyber security) threats, incidents and elements connected to those in the form of events. (9) provide and describe a basic example of a MISP format event as follows:

```
{
    "Event": {
        "date": ":2002-03-12",
        "threat_level_id": "1",
        "info": "testevent",
        "published": false,
        "analysis": "0",
        "distribution": "0",
        "Attribute": [{
            "type": "domain",
            "category": "Network activity",
            "to_ids": false,
            "distribution": "0",
            "comment": "",
            "value": "test.com"
        }]
    }
}
```

FIGURE 5: EXAMPLE OF A MISP THREAT EVENT WITH A SINGLE ATTRIBUTE (9)

Beyond events, MISP data models can contain objects and can be structured and further defined using taxonomies, which are also used in order to tag events with relevant information. MISP galaxies[8] are more complex data model environments that can bundle and group larger data sets with complex relations that can be used dynamically in MISP (10). Galaxies and taxonomies can be used in order to model complex domains such as the one envisioned for the SAFETY4RAILS project.

While the MISP format is flexible and easy to extend, it also provides both a broad coverage of relevant event types, attributes and related data types and specifies and implementations of formalised concepts such as the estimative language model of likelihoods and confidence in sources, data and methodologies as defined in ICD 203 and JP 2-0 [9]. Use of such formalised concepts is also of key importance for the specification of concepts such as the impact of a threat or an executed attack on an asset.

4.1.3 Integration into SAFETY4RAILS Data Model Environment

SAFETY4RAILS undertakes substantial work in the areas of defining and elaborating concepts for railway security both in the cyber and physical domains, including identifying critical IT & OT components in the railway

[8] See https://www.misp-project.org/galaxy.html (accessed 19.04.2021) for an overview over standard galaxies that are available with MISP.
[9] See https://www.misp-project.org/taxonomies.html#_estimative_language (accessed 19.04.2021) for the natural language description of this concept.

domain (T3.1) as well as in particular threats (T3.2, T3.3). Risks and vulnerability modelling and assessment are investigated in T5.1. All of these activities both gather instances and characterise them with attribute descriptions and develop taxonomies in order to organise and structure them into data models.

Since the timings of these activities and the development of the OSINT system overlap, data model work in task T4.2 initially focuses on developing methods for the acquisition of open source data and the development and use of a lightweight taxonomy and model that can easily be reorganised to correspond to the final models produced by tasks T3.1, T3.2, T3.3 and T5.1. To this end, work in the task will initially be based on developments for MISP taxonomies in the railway domain that were carried out as part of the Shift2Rail EU initiative and there in particular the X2RAIL-1 project[10] and the 4SECURail project[11]. We will also consider integration with one the MITRE ATT&CK® matrices[12].

4.2 Data Sources & Data Acquisition

The data used in open source intelligence are the key part of the definition of the term OSINT, but the definition of what the data are is particularly concerned with the accessibility of the data under consideration (see Section 2.1). In this section we briefly present the initially specified data sources for cyber security and physical security respectively and outline the framework for the development of the test and demonstration data for the task.

4.2.1 Open Source Cyber Security Data

A number of key services provide cyber security information that is relevant for the railway domain. In the first instance, the majority of these services will provide access to general cyber security data, which need to be filtered in order to identify the subset that is relevant for the railway domain. Note that this subset is still likely to include a range of general cyber security threats, for instance ones relating to information displays and terminals operating legacy Microsoft Windows operating systems.

TABLE 8: INITIAL SELECTION OF OPEN SOURCE CYBER SECURITY DATA SOURCE CANDIDATES

Name	*Description*	*References*[13]
VirusTotal	Malware repository	https://www.virustotal.com
PolySwarm	Malware and online threat detection repository	https://polyswarm.io
MalwareBazaar	Malware repository	https://bazaar.abuse.ch
Hybrid Analysis	Malware repository	https://www.hybrid-analysis.com
Any.run	Malware analysis sandbox and repository	https://any.run
Shodan	Network security search engine	https://www.shodan.io
Censys	Internet-wide scanning service	https://censys.io

[10] See https://projects.shift2rail.org/s2r_ip2_n.aspx?p=X2RAIL-1 (accessed 19.04.2021) for further information on the X2RAIL-1 project.
[11] See https://www.4securail.eu/ (accessed 19.04.2021) for further information on the 4SECURail project.
[12] See https://attack.mitre.org/matrices/enterprise/ (accessed 25.05.2021) for further information on MITRE ATT&CK®.
[13] All web resources accessed 10.05.2021.

ZoomEye	Address and port scanning service	https://www.zoomeye.org
OTX Threat Feed	Threat Intel feed by AT&T (Former AlienVault)	https://otx.alienvault.com
DigitalSide Threat Intel	Regular data feeds on threat intelligence for cyber security	https://osint.digitalside.it
CERT data feeds	Regular data feeds on breaches, vulnerabilities and system misuse	List of EU CERTS at https://www.enisa.europa.eu/
Twitter	Public feeds of threat alert accounts that are operated by organisations and that are not individual user accounts	https://twitter.com

In addition to data repositories and search engines, relevant domain data for the identification of frequently used railway domain security-relevant websites and relevant security-related social media accounts will be elicited for the demonstration of phishing specialising in the railway security domain.

4.2.2 Open Source Physical Security Data

Open source physical security data is generally not as readily available via structured online data services and data sources as open source cyber security data.

TABLE 9: INITIAL SELECTION OF OPEN SOURCE PHYSICAL SECURITY DATA SOURCE CANDIDATES

Name	*Description*	*References*[14]
GTD Global Terrorism Database	Regularly updated database of terrorist incidents	https://www.start.umd.edu/gtd/
PredictHQ Natural Hazards	Natural hazard indicator feed and database	https://www.predicthq.com
Twitter	Public feeds of threat alert accounts that are operated by organisations and that are not individual user accounts	https://twitter.com
News agency news feeds	News agency news feeds for local events including weather and traffic events	For example https://www.reuters.com/

In the initial data selection, terrorist incidents and natural hazard events have been selected as initial candidates for open source physical security data. These sources may be revised in order to align the task activities with selected demonstration and evaluation scenarios in the project.

[14] All web resources accessed 10.05.2021.

4.2.3 Development and Demonstration Data

For development and demonstration purposes, it is desirable to have an easily controllable development and demonstration environment available. Two such development and demonstration activities are relevant in task T4.2.

First, it is useful to work from a small-scale world model in order to gradually expand testing until the system crosses over to fully specified and modelled demonstration and real environments. To this end, “Lummerland Rail” will be used as a toy example railway for data modelling and technical testing purposes. While the name Lummerland Rail references a children’s book series[15], the toy example railway used in T4.2 uses modern technology and devices.

Second, it is necessary for development and testing to be able to create controlled data sources that respond with search results, data feed, news feed and social media feed data that can be triggered either manually or automatically for automated system testing purposes. To this end, task T4.2 will develop locally deployable test data sources that are representative of the open source data sources that are used in production scenarios while being under full control of the task developers.

[15] See https://en.wikipedia.org/wiki/Jim_Button_and_Luke_the_Engine_Driver (accessed 23.04.2021) for further information on the source for the naming of the railway toy example.

5. Conclusion

This section presents a brief overall summary of the content of this deliverable, presents the outlook for the next deliverable for task T4.2 and sets out a summary capabilities matrix that indicates which of the functionalities identified and described throughout this deliverable address which of the required functionalities that have been outlined in Section 2.

5.1 Summary

This deliverable has provided an overview of the approach to cyber-physical threat detection in terms of functionalities and components that are necessary in order to integrate the envisioned OSINT approach to data acquisition to identify cyber-physical threats in the railway domain. The general system architecture for the identified components was described. The general approach to data representation and main issues concerning data sources and data acquisition were discussed.

The information provided in this document represents the state as envisioned at a relatively early stage in the development of SAFETY4RAILS. The follow-up deliverable to this deliverable is D4.2 “Framework and Methodology of Critical Components Based on OSINT”, which will be based on the final implementation concept and data model for the work carried out in task T4.2. This deliverable will be completed in the second year of development in the project.

5.2 Capability Matrix

In order to review in which way the components proposed for development contribute to the identified relevant system functionalities that the OSINT system in SAFETY4RAILS should provide, the table below relates the required functionalities to components in a capability matrix.

In the table functionalities and components respectively are identified by their unique IDs and briefly characterised by (shortened) summary texts intended to help with recalling the respective functionalities and component characterisations. In each component row, an “x” signifies that a component contributes towards enabling a functionality either partially or completely.

TABLE 10: CAPABILITIES MATRIX RELATING SUGGESTED FUNCTIONALTIIES AND SUGGESTED COMPONENTS

	FUNC-DA-01 Retrieval from cyber security data sources	FUNC-DA-02 Identification of vulnerable & compromised assets	FUNC-DA-03 Retrieval from physical security data sources	FUNC-PP-01 Parsing of standard format data feeds	FUNC-PP-02 Analysis of unstructured data sources	FUNC-SR-01 Operation of a suitable database system	FUNC-DS-01 Analysis of newly added data for threat intelligence	FUNC-DS-02 Generation of statistics to identify trends and ranks	FUNC-DN-01 Exposing data access functionalities	FUNC-DM-02 Messaging data updates to recipient components
DA-TIS-01 Internet exposed asset crawlers	X									
DA-TIS-02 Vulnerability and exploit scanning crawlers		X								
DA-TIS-03 Malware repository crawlers with Yara rules	X	X		X						
DA-TIS-04 Malware social media crawlers with Yara rules	X				X					
DA-TIS-05 Phishing campaign monitoring crawlers		X								
DA-TIS-06 Threat intel feed selection and configuration	X			X		X				
DA-MSP-01 Threat intel feed selection and configuration for cyber threats	X			X		X				
DA-MSP-02 Structured data source integration for physical threat data			X	X						
DA-MSP-03 Social media data source integration for physical threat data			X							
PA-MSP-01 Threat intel feed automated relevance filtering				X	X					
PA-MSP-02 Structured physical threat automated data feed filtering				X	X					

PA-MSP-03 Semantic social media data parsing for physical threat detection					X					
SR-TIS-01 Data repository operation						X				
SR-TIS-02 Data model customisation						X				
SR-MSP-01 Data repository operation						X				
SR-MSP-02 Data model customisation						X				
DS-MSP-01 Rule evaluation over database							X			
DS-MSP-02 Generation of summary statistics								X		
DM-TIS-01 Synchronisation with SAFETY4RAILS MISP instance						X				
DM-MSP-01 Custom API endpoint development									X	
DM-MSP-02 Message broker integration										X

BIBLIOGRAPHY

1. **Steele, R.D.** Open Source Intelligence. [book auth.] L.K. Johnson. *Handbook of Intelligence Studies.* New York : Routledge, 2007, pp. 129-147.

2. **Wikimedia Foundation.** Open-Source Intelligence. *Wikipedia.* [Online] [Cited: 19 04 2021.] https://en.wikipedia.org/wiki/Open-source_intelligence.

3. **Lowenthal, M.M.** Open-Source Intelligence: New Myths, New Realities. [book auth.] R.Z. George and R.D. Kline. *Intelligence and the National Security Strategist: Enduring Issues and Challenges.* Lanham : Rowman & Littlefield, 2004, pp. 273-278.

4. **Gartner.** Security Threat Intelligence Products and Services Reviews and Ratings. *Gartner.* [Online] [Cited: 19 04 2021.] https://www.gartner.com/reviews/market/security-threat-intelligence-services.

5. **Wagner, C., et al.** MISP: The Design and Implementation of a Collaborative Threat Intelligence Sharing Platform. *Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security.* Vienna : ACM, 2016, pp. 49-56.

6. **Free Software Foundation.** GNU Affero General Public License v3.0 or later. [Online] [Cited: 19 04 2021.] https://spdx.org/licenses/AGPL-3.0-or-later.html.

7. **Kellet, M. and Bernier, M.** *Cyber Threat Data Model. High-Level Model and Use Cases.* s.l. : Defense Research and Development Canada, 2016. DRDC-RDDC-2016-D080.

8. **Magar, A.** *State-of-the-Art in Cyber Threat Models and Methodologies.* 2016. DRDC-RDDC-2016-C132.

9. *Distributed Security Framework for Reliable Threat Intelligence Sharing.* **Preuveneers, D., et al.** 2020, Security and Communication Networks, pp. 1-15.

10. **MISP Project.** Best Practices in Threat Intelligence. *MISP Project.* [Online] [Cited: 19 04 2021.] https://www.misp-project.org/best-practices-in-threat-intelligence.html.

ANNEXES

The annexes to this deliverable contain auxiliary information. Annex I. “Glossary and Acronyms” contains an overview of abbreviations and acronyms used in this deliverable.

ANNEX I. GLOSSARY AND ACRONYMS

The following table lists and defines or describes the abbreviations and acronyms used in this deliverable.

TABLE 11: GLOSSARY AND ACRONYMS

Term	*Definition/Description*
AWS	Amazon Web Services
DoA	Description of Action
IDS	Intrusion Detection System
IoC	Indicator of Compromise
IT	Information Technology
JSON	JavaScript Object Notation
MISP	Malware Information Sharing Platform
OSINT	Open Source Intelligence
OT	Operational Technology
SCADA	Supervisory Control and Data Acquisition
SOC	Security Operations Centre
TI	Threat Intelligence
TTP	Tactics, Techniques and Procedures

SAFETY4RAILS

Partners:

Fraunhofer EMI
UIC
Metro de Madrid
EGO Genel Müdürlüğü
RFI Rete Ferroviaria Italiana – Gruppo Ferrovie dello Stato Italiane
Leonardo
ceis avisa partners
STAM Mastering Excellence
IC information company
RMIT University
University of Reading
etra I+D
Demokritos National Centre for Scientific Research
RINA
WINGS ICT Solutions
Newcastle University
EOS European Organisation for Security
Innova Integra
Laurea Ammattikorkeakoulu University of Applied Sciences
CybeServices Made in Europe
erGE
MTRS
Intracom Telecom
Universitas Miguel Hernández
TREE Technology
Elbit Systems C4I and Cyber
ProRail
Comune di Milano

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 883532.