A I 4 C Y B E R

Artificial Intelligence for next generation CYBERsecurity

Deliverable title

Data Management Plan - Initial version

Deliverable ID:

D1.2

Preparation date:

01/03/2023

Editor/Lead beneficiary (name/partner):

Erkuden Rios / TECNALIA

Internally reviewed by (name/partner):

Jason Mansell / TECNALIA,

Panagiotis Radoglou / UOWM

Abstract:

This document details the Data Management Plan (DMP) of AI4CYBER project which defines the procedures and criteria established in the project regarding the data handling strategy of the project, while guaranteeing ethics and privacy in project activities and outcomes.

Dissemination level		
PU	Public, fully open	X
SEN	Sensitive, limited under the conditions of the Grant Agreement	
EU-R	Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No 2015/444	

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450

Disclaimer: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.

AI4CYBER consortium

1	tecnalia MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE	Fundación Tecnalia Research & Innovation (TECNALIA, Spain)
2		University of Western Macedonia (UOWM, Greece)
3	montimage	Montimage EURL (MI, France)
4	THALES	Thales Six GTS France SAS (TSG, France)
5	SEARCH-LAB SECURITY EVALUATION ANALYSIS AND RESEARCH LABORATORY	Search Lab (SLAB, Hungary)
6	FRONTENDART	Frontendart Szoftver KFT (FEA, Hungary)
7	EOS EUROPEAN ORGANISATION FOR SECURITY	European Organisation for Security (EOS, Belgium)
8	PDM	PDM E FC Projecto Desenvolvimento Manutencao Formacao e Consultadorialda (PDMFC, Portugal)
9	iTTi	ITTI Sp. z o.o. (ITTI, Poland)
10	ΔΕΗ	Public Power Corporation S. A. (PPC, Greece)
11	Hospital do Évora Espírito Santo E.P.E.	Hospital Do Espirito Santo de Evora EPE (HES, Portugal)
12	CaixaBank	Caixabank S.A. (CXB, Spain)
13	MINDS	Metamind Innovations P.C (MINDS, Greece)

Table of contents

AI4CYBER consortium 2
Table of contents 3
List of figures 5
List of tables 6
Executive summary 7
1 Introduction 8
1.1 Objective of the document 8
1.2 Structure of the document 8
1.3 Relationships with other deliverables 8
1.4 Contributors 8
1.5 Acronyms and abbreviations 8
1.6 Revision history 9
1.7 Change log 9
2 AI4CYBER Data Management Plan 10
2.1 Data summary 10
2.1.1 Project deliverables 10
2.1.2 Scientific publications 11
2.1.3 Other publications 12
2.1.4 Research datasets 12
2.2 FAIR Data 13
2.2.1 Making data findable, including provisions for metadata 13
2.2.2 Making data openly accessible 14
2.2.3 Making data interoperable 14
2.2.4 Increase data re-use (through clarifying licences) 15
2.3 Data security and privacy 15
2.3.1 Personal data protection management 15
2.3.2 Security management 18
2.4 Ethics aspects 19
2.5 Data Management Roles and Responsibilities 20
2.6 Monitoring and update of the DMP 21
3 AI4CYBER research datasets 23
3.1 Dataset for AIFix4SecCode 23
3.2 JiraMiner dataset 25
3.3 IEC 60870-5-104 Intrusion Detection Dataset 26
3.4 OCPP_DDOS_FakeEVs_Dataset 28
3.5 AI4AppSec dataset 30
3.6 AI4AuthLog dataset 31
3.7 AI4NetHealth dataset 32
3.8 AI4WinEvent dataset 34
3.9 CXB-InsiderThreat-AzureAD dataset 35
3.10 CXB-InsiderThreat-CyberArk dataset 36
3.11 CXB-InsiderThreat-Prisma dataset 38
3.12 SIMARGL2021 dataset 39
3.13 French COVID19 study 41
4 Conclusion 43
References 44
Appendix A. Dataset description template 45

Appendix B. Informed consent form..47

List of figures

Figure 1: Data Management structure in AI4CYBER 20
Figure 2: Data Management Plan updates.. 22

List of tables

Table 1: Public deliverables in AI4CYBER.. 10
Table 2: Data Management Roles in AI4CYBER... 20

Executive summary

As part of the AI4CYBER project management, the *Ethics, Privacy, Data management and open data pools* management task is in charge of defining and revising the Data Management Plan (DMP) of the project.

The present document is the initial version of the DMP which sets out the methodology to be followed by AI4CYBER consortium members throughout the project duration to handle the data used and created in the project.

The report presents the categories of the data that will be created and processed (including storage) within AI4CYBER.

The deliverable describes how the FAIR principles (as laid down by the EC guidelines) will guide the preparation and processing of all the open data generated in AI4CYBER.

The plan describes the data security and data privacy measures adopted in the project to ensure no privacy-sensitive and no security-sensitive information is disclosed in the project data.

Ethics aspects in relation to the data are also described as part of the plan.

The roles and responsibilities of the partners and key participants in the project in the data management plan are clarified, providing the necessary contact data.

This initial version of the DMP shall undergo a regular revision since the plan may be adjusted to project needs in the future. Therefore, the timeline for monitoring and updating the DMP is presented, and the procedure explained.

The document also provides the collection of initially identified datasets that will be used in the project's research activities. The datasets are described according to the best knowledge of the partners and follow a common description template that will be used in the project to manage the datasets. The template itself is shown in Appendix A.

1 Introduction

1.1 Objective of the document

This document is deliverable D1.2 *Data Management Plan – Initial version* of AI4CYBER project [1].

The document details the Data Management Plan (DMP) of AI4CYBER project which defines the procedures and criteria established in the project regarding the data handling strategy of the project, while guaranteeing ethics and privacy in project activities and outcomes.

Pursuant to European Commission's goal to advance Open Science policies, the DMP of AI4CYBER defines the ways in which data is collected, generated and/or processed throughout the project's lifetime.

The present report is the initial version of the plan which describes the data currently used by the project and the initial procedures designed for data life-cycle management. The report will be updated at the end of the project in deliverable D1.4 *Data Management Plan – Final version* (M36) to reflect any potential updates in the procedures adopted, as well as the final data types generated in the project.

The deliverable is the result of Task 1.4 *Ethics, Privacy, Data management and open data pools* devoted to the analysis of the privacy, ethical, and societal impact of the proposed technologies and project activities and to the design of needed procedures for minimising related risks.

1.2 Structure of the document

This document is structured as follows:

Section 2 describes the DMP, which is structured in the following main components, as suggested in the Guidelines on FAIR Data Management in Horizon 2020 [2]: i) Data summary, ii) FAIR Data principles, iii) Data security and privacy aspects, iv) Ethics aspects, and v) DMP monitoring and update procedures. These guidelines still apply to Horizon Europe.

Section 3 lists the datasets that were initially identified by the AI4CYBER partners to be used in the project research activities.

Finally, Section 4 concludes the document providing a summary of the report and a description of future work.

Appendix A shows the dataset description template to be adopted in the project.

Appendix B provides the Informed consent form to be used to gain signed informed consent from data subjects involved in project activities that process personal data.

1.3 Relationships with other deliverables

The DMP presented in the document relates to the following deliverable:

- D1.4 Data Management Plan – Final version (M36) will present the final data management methodology followed in the project and it will also report the final data generated in the project.

1.4 Contributors

The following partners have contributed to this deliverable:

- TECNALIA as coordinator of the report edition.
- ALL partners as contributors to the design and description of the procedures and initial datasets.

1.5 Acronyms and abbreviations

DMP	Data Management Plan	PSO	Project Security Officer

DPO	Data Protection Officer	SAB	Security Advisory Board
OSS	Open Source Software	URL	Uniform Resource Locator
PC	Project Coordinator	WP	Work Package
PoC	Proof of Concept		

1.6 Revision history

Version	Date issued	Author	Organisation	Description
V0.0.1	20/12/2022	Erkuden Rios	TECNALIA	TOC prepared
V0.1	13/01/2023	Erkuden Rios	TECNALIA	TOC agreed and initial contents provided. Dataset template defined.
V0.2	17/02/2023	Erkuden Rios	TECNALIA	Internal review version.
V0.3	21/02/2023	Panagiotis Radoglou	UOWM	Reviewed version.
V0.4	23/02/2023	Jason Mansell	TECNALIA	Reviewed version.
V1.0	28/02/2023	Erkuden Rios	TECNALIA	Final version.

1.7 Change log

No change log entries.

2 AI4CYBER Data Management Plan

This section describes the overall Data Management Plan of the project. This plan is the initial version which shall be adjusted to project needs in the future. The current plan first identifies the data types that will be generated in the project and outlines how the FAIR principles should be kept in the open data. The plan describes the data security and data privacy measures adopted in the project to ensure no privacy-sensitive and no security-sensitive information is disclosed in the project data. Ethics aspects in relation to the data are also described. The roles and responsibilities of the partners and key participants in the project in the data management plan are clarified, providing the necessary contact data. Finally, the timeline for monitoring and updating the DMP is presented, and the procedure is explained.

2.1 Data summary

The AI4CYBER research activities will require the handling of diverse data types, including both data used in the project and data generated in the project.

In this section we describe the types of data resulting from project activities, together with the description of their purpose and format. These data are open data that will be shared among the participants in AI4CYBER as well as with audiences outside the project.

It is worth noting that the project participants are expected to also use similar open data types from other projects and open public sources. The purpose of the use of such open data is multiple, including reviewing state-of-the-art approaches to problems tackled by AI4CYBER, learning about events organised by other projects that may be interesting for AI4CYBER research, seeking collaboration opportunities with external experts in the themes studied by the project, etc.

2.1.1 Project deliverables

AI4CYBER produces a set of reports that summarize the main project activities, research methodologies, results, etc., which are contractual deliverables as per the Grant Agreement. From all these deliverables only those of public nature (marked as PU) are shared outside the Consortium as open data. The project public deliverables will be publicly released, when it is prescribed in the description of the work, only after the acceptance from the European Commission. The publication of the deliverables will be made both in the OpenAIRE repository (https://www.openaire.eu/) (through automatic publication once the reports are validated by the EC) and on the project website (https://ai4cyber.eu/).

Due to the cybersecurity relevant and cutting-edge research of AI4CYBER in the areas of understanding and investigation of advanced and AI-powered cyber-attacks, detection and protection means against them, their simulation, etc. a limited number of deliverables are initially considered as information that shall be made publicly available early in the project. Please note that the results of research work in AI4CYBER will be mostly published in open access scientific publications, as described in next section.

Table 1 lists the public deliverables planned in AI4CYBER.

Table 1: Public deliverables in AI4CYBER

Del. No.	Deliverable Name	Description	Leader	Type	Due Month
D1.2	Data Management Plan - Initial version	This document details the established procedures and criteria to guarantee ethics and privacy in all project activities and outcomes, as well as the data management strategy of the project.	TECNALIA	DMP	M6

D1.4	Data Management Plan - Final version	This document includes a revision, to be performed at end of the project, of the established procedures and criteria to guarantee ethics and privacy in all project activities and deliverables.	TECNALIA	DMP	M36
D3.1	AI-driven self-testing and automatic error correction for robustness - Initial version	The first version of the description of methodology and prototype implementation for automatic correction of robustness related weaknesses and AI boosted symbolic execution-based vulnerability identification algorithms.	SLAB	OTHER	M17
D3.3	AI-driven self-testing and automatic error correction for robustness - Final version	The final version of the description of methodology and prototype implementation for automatic correction of robustness related weaknesses and AI boosted symbolic execution-based vulnerability identification algorithms.	SLAB	OTHER	M34
D8.1	AI4CYBER Communication Tools	This deliverable will cover the set-up of social media channels, templates, flyers, and all communication supporting materials. The initial version of the project website will include at least project objectives and contact details. AI4CYBER website will be set-up by TECNALIA and continuously enhanced by all partners to include public downloadable results and links to related news and initiatives.	TECNALIA	DEC	M4

As it can be seen in the table, two public deliverables (of type OTHER) will include not only a report, but some other type of result associated with it. These other types of results may come in form of software, models, algorithms, and workflows. Furthermore, as initially planned, some of the non-public deliverables in the project do also include software that will be released with a dual licence, i.e., an open source software version and a privative version.

The open source software results of the project, including the OSS models, software PoCs and tools from the project, will be uploaded to widely used open repositories such as GitLab. Following the principle *'as open as possible as closed as necessary'*, when possible, the Creative Commons licenses CCBY or CC-0 for project outcomes (for those not commercially IP-protected outcomes) will be used to ensure that they are shared with minimal restrictions, aside from attribution to the authors or creators.

2.1.2 Scientific publications

AICYBER Consortium adheres to the open science principles of the Horizon Europe program promoting and facilitating open cooperative work among project participants, with other related projects and initiatives, and with the scientific community in general. The communication and dissemination strategy of the project will address the systematic sharing of knowledge and tools from AI4CYBER as early and widely as possible.

In this line, the aim of the partners is to promote and offer free-of-charge scientific information whenever possible. Therefore, research results published during the project will be open access whenever the publication license authorizes it and the commercial interests of the partners align with it. For early works, we may use the pre-registration and open peer-review service offered by the European site for open research [2].

Open access to the peer reviewed scientific publications of the project will be provided with the highest standards when possible. The members of the Consortium will give 'golden' or "green" open access. These types of access imply that a published scientific article or the final peer-reviewed manuscript will be immediately or after a delay (embargo) of 6-12 months as maximum, provided in open access mode by the publisher or the authors, respectively.

2.1.3 Other publications

AI4CYBER will also produce or reuse data for its dissemination, communication, networking, and exploitation activities under WP8. The list below describes the relevant data and the formats that are foreseen to be used by WP8:

- Videos of AI4CYBER solutions (mp4)
- Recordings of AI4CYBER webinars (mp4)
- Press releases (docx, pdf)
- Brochures (docx, pdf)
- Newsletters (docx, pdf)
- List of events in which AI4CYBER was represented (docx)
- Project presentations (pptx, ppsx, pdf)
- List of website publications and posts (docx)
- Partners blog posts for the website (docx)
- Statistical data from webpages and social media pages (xlsx)

All these data will be made publicly accessible through the project website (https://ai4cyber.eu/) whenever possible. In case that videos and recordings containing personal data of participants are published, data subjects will be asked to sign an informed consent prior the publication, as it will be explained in Section 2.3.1.1. In case any of the participants objects to publicly share the data, it may undergo a curation process to anonymize the information if possible or be otherwise kept private.

2.1.4 Research datasets

The AI4CYBER project will use and generate datasets from project research activities carried out in the study and development of the AI4CYBER solution components.

Regarding open access to research data, those data sets behind the research results that will be considered open will be made available in trusted European open data pools such as Zenodo, AI4EU platform or new emerging ones dedicated to cybersecurity. This contribution aims at enhancing the collective knowledge regarding advanced and AI-powered cyber-attacks in the cyber security community. This will facilitate the research, design, and development of robust solutions to prevent and react against such attacks.

While aiming at making research datasets as openly available as possible, the Consortium will seek to protect the commercial interests of participating beneficiaries and will respect European and national privacy and security regulations.

When handling the datasets in the project, it is necessary to provide details of the public datasets used in the research as well as describe those datasets produced in the project, particularly those that will be made available to the public. Therefore, a template for describing the datasets was defined in

AI4CYBER, based on the template defined in H2020 ELECTRON project [3], which in turn was based on the HORIZON 2020 DMP Template version 2.0 [4].

2.2 FAIR Data

AI4CYBER is committed to adhere to the four FAIR principles as defined by the EC guidelines [7], which refer to the findability, accessibility, interoperability, and reusability of the scientific research data. FAIR principles are particularly addressing the management of research datasets developed in the project, since they support open sharing, distribution, and reuse of data.

FAIR stands for the following four key characteristics of the open (meta)data so as it can be useful to its users:

- Findability: open (meta)data shall be easily searchable and locatable, which relates to the fact that it is necessary to include identifiers, keywords, version numbers and metadata to maximise the possibilities for finding and re-using the data.
- Accessibility: openness of the (meta)data shall be guaranteed, and the means of access clarified and enabled. Besides the details of the repository where the data will be deposited, the means and protocols to access the need to be clarified together with any potential restrictions.
- Interoperability: refers to the ability of the (meta)data to integrate or work together with minimal effort with other (meta)data and tools. Therefore, to enable (meta)data exchange, re-use, and interoperation the vocabularies, standards, formats, and methodologies that will be used need to be clarified.
- Reusability: the ability of the (meta)data to be utilized in other contexts including accurate and relevant attributes, clear and accessible (meta)data usage licenses, meeting domain relevant standards and identifying provenance.

2.2.1 Making data findable, including provisions for metadata

In order to ensure the data can be easily discovered and identified, AI4CYBER project will adopt standard naming conventions, search keywords, versioning numbering, and metadata, as follows.

AI4CYBER plans to use the **standard identification** mechanism, namely Digital Object Identifiers (DOI) to ensure persistent and unique identifiers in the data. DataCite[1] for data identifiers will also be evaluated.

Metadata is data about the research data itself. It allows other researchers to find the data and is essential for the reuse of the data. The richer the metadata, the easier to find and reuse the data by other researchers. Metadata (data type, location, etc.) will be loaded in a standardized way, and, simultaneously, it will be kept separate from the original research data.

With the aim to characterise the data, metadata can include multiple types of information, including DOI, title, date of creation, date of publication, version number, author, publisher, copyright, license, keywords, grant agreement number, project acronym, data format, etc.

The following bibliographic metadata will be used to identify the open scientific publications while providing access to them

- Funding grant: EC Horizon Europe
- AI4CYBER, Grant Agreement No 101070450
- Publication date, length of the embargo period (if applicable), persistent identifier

[1] https://datacite.org/

Keywords shall also be provided that optimize possibilities for searching and re-use. These keywords can also be part of the metadata.

The datasets generated in the project will undergo different stages where versioning of the datasets will be needed. The original dataset along with all its copies and versions will be kept by the dataset owner, together with the corresponding metadata set associated to each dataset. The project will maintain a catalogue to enable version control of the datasets created in the project in order to track the updates. The partners are requested to register in the catalogue all the datasets produced and maintained by them.

2.2.2 Making data openly accessible

The AI4CYBER project is committed to follow the European Commission's guidelines for open access to research data [6]. The plan of the project is to rely on a widely used open repository to deposit the project generated datasets. Among the candidate open access repositories for the project datasets ZENODO (https://zenodo.org/) open repository is the one favoured by the Consortium. ZENODO was developed as a joint result by OpenAIRE and CERN. The repository is securely hosted and operated by CERN, and it integrates with OpenAIRE platform, which already will store the public data of AI4CYBER project as well as all public deliverables of the project approved by the EC. This repository is the first candidate considered in the project since it is a trusted repository widely used to store open data from EU-funded projects. In order to publish all the partners' generated data under the common umbrella of the project, the Project Coordinator would set up a dedicated ZENODO community for the project that will be used by the partners as the main repository of the open datasets from the research.

Prior to the publication of the open datasets, the dataset owner shall identify whether the dataset may contain private or sensitive information of any kind. The curation of the data will be carried out by the dataset owner prior to the dataset publication, including all necessary measures for the protection of personal data (see section 2.3.1), protection of security sensitive information (see section 2.3.2), and protection of confidential data of the owner.

The datasets and associated metadata and documentation will be deposited in ZENODO, while the OSS code will be deposited in either ZENODO or another code repository in use by the partners that offers open access. In this case, the integration with ZENODO will be seek.

The open data will include a machine-readable license whenever possible.

Every partner will be responsible for providing an open access option to every scientific article and paper published by them. The methodology for making scientific publications open will be the following one:

1. The research group publishes the article in the journal, conference, or publisher of their choice.
2. The hosting partner of the main author adds the final peer-reviewed manuscript to an open access repository. There will be different options:
   a. Open Access journal or conference (gold open access)
   b. Self-archiving (green open access), in case the host institution of the author has an institutional repository for open access publications, integrated with OpenAIRE platform.
   c. Archiving resort (Zenodo, preferably) if the previous options cannot be used.
3. Registration of the publication in OpenAIRE portal linking it with AI4CYBER project.

In addition, all other open publications and public project deliverables will be published in the project website.

2.2.3 Making data interoperable

In order to facilitate the reuse of open data from external sources, as well as AI4CYBER open data be reused outside the project, AI4CYBER will promote and use generic format standards compatible with freely available software programs.

The usage of common terminology across all data types is a facilitator of data interoperability. The AI4CYBER project is developing a well-defined taxonomy of attacks and other data types which will be complemented with project glossary and vocabulary mappings.

Overall, considering data interoperability is a crucial aspect of effective data management and sharing. During the project, it may become necessary to identify new data that requires interoperability measures, such as the use of standardized data vocabularies or established methodologies to facilitate seamless integration. Such information can be included in subsequent versions of the DMP to ensure that data interoperability is appropriately addressed.

2.2.4 Increase data re-use (through clarifying licences)

The European Commission defines open access as the provision of scientific information online, which is free of charge to users and can be reused. This approach is widely acknowledged as a means of enhancing science, innovation, and efficiency across both public and private sectors.

By depositing the data into an open access repository like ZENODO, the AI4CYBER project will facilitate the reusability of its data.

The research datasets generated in the project use cases may be linked to findings that the partners may further exploit following commercial interests. Therefore, granting the availability of such data for third parties re-use will be studied on a case-by-case basis in accordance with the interests of the partners in AI4CYBER.

Open Access licenses shall be granted in the data wherever possible. Other types of licensing are also possible in case no open access is needed. However, the principle ‘as open as possible, as closed as necessary’ shall be applied. Licenses shall be used by dataset owners to offer access to their data stating the permissions over the data. data. It is worth noticing that in most cases (e.g., Creative Commons licenses) licenses exclusively relate to copyright and copyright-related issues, while personal information protection is not covered. The Creative Commons Attribution 4.0 International (CC BY 4.0) is usually the best option for scientific publications and open source software. The CC0 license absolves the user of all responsibility for the content of the data.

2.3 Data security and privacy

This section describes two important procedures established in AI4CYBER around the protection of data handled in the project, as follows:

- Personal data protection management, including:
  - the measures that will be implemented to safeguard the rights and freedoms of the data subjects, both research participants in AI4CYBER and other stakeholders, in relation to the protection of their privacy.
  - the privacy enhancing techniques that will be implemented in the data handled by the project.
  - the security measures that will be implemented to prevent unauthorised access to personal data or the equipment used for processing,
- The security measures that will be implemented to prevent that no security sensitive information is disclosed in the project deliverables and data.

2.3.1 Personal data protection management

As required in the Grant Agreement, the AICYBER consortium partners are committed to process personal data in compliance with the applicable EU, international and national law on data protection (in particular, the GDPR [9]). The beneficiaries are required by the Grant Agreement and the Consortium Agreement to limit their personnel access to personal data unless it is strictly needed for the purpose of implementing, managing, or monitoring the project activities.

In compliance with the GDPR, the partners processing personal data agreed in the Consortium Agreement to implement the required data security measures to protect the data, including the logical and physical access control measures to prevent unauthorised access to personal data and to the data processing systems, timely updates of information systems, deployment of network security measures, backup and recovery systems, etc.

The schema adopted for partners processing personal data is the joint controllership as described Section 2.3.1.3.

The legal basis for processing that data would be informed consent. In all project activities the collection of personal data shall be limited to the strictly needed amount and types, and whenever personal data is about to be collected the data subjects involves shall be duly informed and requested to sign the Consent form designed in the project, cf. Appendix B.

In internal project activities such as technical meetings and plenary meetings, either in person or online, in case personal data is required, the participants shall be informed, and their consent granted through signing the consent form. In WP8 dissemination activities, as part of the event registration all stakeholders will be asked to sign a consent form before having their pictures taken during AI4CYBER events, webinars, and meetings. The personal data requested to stakeholders during the sign up for subscription to project newsletters will only be used with the purpose to verify the subscription and send the newsletter.

2.3.1.1 Personal data in AI4CYBER

In general, the technical research focus of the project, i.e., the design and development of AI-enabled cybersecurity solutions, is not related to personal data processing, and the research carried out in the project does not require the processing of personal data, but data of the networks and systems under test in the use cases of the project, which are emulated critical infrastructures, so no real data is contained therein.

During the project activities personal data (such as telephone numbers, group pictures, etc.) may be processed for contacting the project participants to participate in project tasks, arrange project meetings, organise events or disseminate them.

In particular, the project meetings related files may contain:

- Project meetings recordings stored for minutes purposes and for clarifying technical descriptions and decisions, which may contain presenters' images and voices, as per informed consent.

As part of WP7 validation, the following personal data items may be processed:

- Personal opinions of respondents to AI4CYBER solution validation questionnaires made online through the EU Survey tool, as per informed consent.
- Videos showing the validation of AI4CYBER solutions (mp4) that may contain presenters' images and voices, as per informed consent.

In addition, as part of WP8 work, the following personal data items may be processed:

- Personal data of the targeted stakeholders in dissemination, communication, networking, and exploitation activities (first and last name, email addresses, country, type of organisation) (xlsx), as per informed consent.
- Personal data of stakeholders and partners joining project dissemination events (images) (jpg), as per informed consent.
- Videos showing the use of AI4CYBER solutions (mp4) that may contain presenters' images and voices, as per informed consent.
- Recordings of AI4CYBER webinars (mp4) that may contain images and voices of both presenters and attendees, as per the informed consent of both.

As mentioned above, all beneficiaries are required by the Grant Agreement to limit their personnel access to such personal data unless it is strictly needed for the purpose of implementing, managing, or monitoring the project activities, and to ensure that their personnel is obliged to confidentiality clauses in the Grant Agreement and the Consortium Agreement of the project.

2.3.1.2 Personal data in AI4CYBER research datasets

The project does not foresee that real personal data is contained in the datasets used or generated in the project. However, in the event that personal data is contained in the datasets from the use cases, prior to the sharing of the dataset, appropriate mechanisms of anonymization shall be applied. The dataset owner shall be responsible for implementing the anonymization, evaluating its appropriateness, and ensuring the quality of the resulting dataset.

The Project Coordinator shall assist the dataset owner in selecting the anonymization technique and tool in each case, if necessary. The PC shall oversee that the anonymization has been successfully applied.

2.3.1.3 Personal data processing and DPOs

The AI4CYBER project has adopted a Joint Controller schema where all partners in the consortium are data controllers who jointly determine the purposes and means of personal data processing and are responsible for any processing by third parties from outside the consortium acting as data processors on their behalf. Therefore, the joint controllership basis, the joint-controllers shall implement the appropriate technical and organisational measures and data protection policies so as to fulfil the provisions of Article 26 of the GDPR [9] and all other applicable national or European Union legislation on the matter.

In this line, the partners have appointed contact representatives who will oversee handling all data protection aspects of the project.

The DPOs appointed for AI4CYBER are the following:

1. TECNALIA: Javier Lerma, Programs Director, email:javier.lerma@tecnalia.com
2. UOWM: Dionysios Kalogeras, DPO, dpo@uowm.gr
3. MI: Edgardo Montes de Oca, CEO Montimage, edgardo.montesdeoca@montimage.com
4. TSG: Pascal Bisson, R&D supervisor, pascal.bisson@thalesgroup.com
5. SLAB: Ákos Milánkovich, researcher, akos.milankovich@search-lab.hu
6. FEA: Tibor Bakota, CEO, tibor.bakota@frontendart.com
7. EOS: Elodie Reuge, Senior Project Manager, elodie.reuge@eos-eu.com
8. PDMFC: Francisco Correia Loureiro, francisco.loureiro@pdmfc.com
9. ITTI: Lidia Samp, lidia.samp@itti.com.pl
10. PPC: Spyropoulos Ioannis, dpo.office@dei.gr
11. HES: Paula Correia, DPO, pcorreia@hevora.min-saude.pt
12. CXB: Esther García Encinas, Manager in Innovation & Privacy's Department, esther.garcia@caixabank.com
13. MINDS: Anastasia Kazakli, Director and DPO of MetaMind Innovations P.C., nkazakli@metamind.gr

The partners' DPO contacts are responsible for acting as primary reference for their organisations with respect to data protection aspects in AI4CYBER project. They are willing to collaborate in ensuring that appropriate data protection measures are taken by their organisations and in monitoring any data protection issue that may arise from the project activities. In case of issues, they will duly contribute to solving them in timely manner.

2.3.2 Security management

The security management in AI4CYBER refers to the protection of security sensitive information.

There are two main bodies set-up in the project to the security risks of the deliverables and data handled in the project: The Project Security Officer (PSO) and the Security Advisory Board (SAB).

The AI4CYBER PSO, appointed at moth three (M3) of the project, is in charge of monitoring the project deliverables and prevent any security issue arising from them, as well of ensuring the correct protection and handling of the EU classified information (EUCI) in case it is necessary.

The PSO is supported by the Security Advisory Board (SAB) of the project which was set up at M3 too. The SAB mission is to assist the PSO in the design and implementation of the relevant measures to manage, mitigate and effectively respond to security concerns when they arise.

Therefore, the PSO together with the SAB will promote a pro-active culture of security amongst the AI4CYBER consortium and ensure timely management of security risks.

The following four experts on cybersecurity compose the AI4CYBER SAB, chaired by the PSO. All of them have excellent background in cybersecurity and together they provide good knowledge of the sectors tackled during the three use cases of the project, as indicated.

- **Project Security Officer (PSO)**: Prof. Panagiotis Sarigiannidis, UOWM, psarigiannidis@uowm.gr is the Director of the ITHACA lab (https://ithaca.ece.uowm.gr/), co-founder of the 1st spin-off of the University of Western Macedonia: MetaMind Innovations P.C. (https://metamind.gr), and Associate Professor in the Department of Electrical and Computer Engineering in the University of Western Macedonia, Kozani, Greece. He received the B.Sc. and Ph.D. degrees in computer science from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2001 and 2007, respectively. He has published over 260 papers in international journals, conferences and book chapters, including IEEE Communications Surveys and Tutorials, IEEE Transactions on Communications, IEEE Internet of Things, IEEE Transactions on Broadcasting, IEEE Systems Journal, IEEE Wireless Communications Magazine, IEEE Open Journal of the Communications Society, IEEE/OSA Journal of Lightwave Technology, IEEE Transactions on Industrial Informatics, IEEE Access, and Computer Networks. He received 5 best paper awards. He has been involved in several national, European and international projects. He is currently the project coordinator of three H2020 projects, namely a) H2020-DS-SC7-2017 (DS-07-2017), SPEAR: Secure and PrivatE smArt gRid, b) H2020-LC-SC3-EE-2020-1 (LC-SC3-EC-4-2020), EVIDENT: bEhaVioral Insgihts anD Effective eNergy policy acTions, and c) H2020-ICT-2020-1 (ICT-56-2020), TERMINET: nexT gEneRation sMart INterconnectEd ioT, while he coordinates the Operational Program MARS: sMart fArming with dRoneS (Competitiveness, Entrepreneurship, and Innovation) and the Erasmus+ KA2 ARRANGE-ICT: SmartROOT: Smart faRming innOvatiOn Training. He also serves as a principal investigator in the H2020-SU-DS-2018 (SU-DS04-2018), SDN-microSENSE: SDN-microgrid reSilient Electrical eNergy SystEm and in three Erasmus+ KA2: a) ARRANGE-ICT: pArtneRship foR AddressiNG mEgatrends in ICT, b) JAUNTY: Joint undergAduate coUrses for smart eNergy managemenT sYstems, and c) STRONG: advanced firST RespONders traininG (Cooperation for Innovation and the Exchange of Good Practices). His research interests include telecommunication networks, internet of things and network security. It is worth mentioning that Panagiotis Sarigiannidis is a SAB member of various projects like SPEAR, SDN-microSENSE and ELECTRON. Moreover, he was the PSO of SPEAR. Finally, he is an IEEE member and participates in the Editorial Boards of various journals.
- **Three SAB members** corresponding to the three sectors addressed by the project use cases, as follows:
  - Energy sector: Alexios Lekedikis, Assistant Director at EU Programs Coordination Department, PPC, A.Lekidis@dei.gr.

- Health sector: Ricardo Cabecinha, Information Security Officer, HES, rjcabecinha@hevora.min-saude.pt.
- Banking sector: Mario Maawad, Security Innovation and Digital Transformation Manager, CXB, mmaawad@caixabank.com.

The AI4CYBER project does not work with EU classified information but most deliverables have limited dissemination because, while they are not classified (no EU restricted), still their content may be security sensitive.

The security management bodies will review all technical deliverables of the project as they are delivered, and they will identify any aspect within the reports and on in the information/data handled to develop them that should be protected. In case their review reveals that some security-sensitive information is contained in the document or data, the PSO and the SAB shall timely collaborate to define the measures and guidelines for the partners on how to prevent the information be disclosed.

As part of security procedures, the PSO and SAB will consult with the data owner to ensure that no security issues arise from the data when sharing it internally or externally to the project, i.e., assessing that the data may not reveal any security or privacy sensitive information.

The PSO and the SAB will meet every 3 months to discuss the security reviews' results and the activities during the period identifying any potential security risks. The meetings shall be web-meetings chaired by the PSO.

Both the PSO and the SAB shall organise every 3 months web-meetings with the Executive Board to report to the Project coordinator and WP leaders the summary of their activities and discuss with them about upcoming technical deliverables and activities. The goal of the meetings is to early identify potential security issues foreseeing by the participants in the research tasks as well as in deliverables production and data generation.

Finally, seeking that all the Consortium is get informed about the progress of the security management activity, the summary of the security monitoring activities, decisions taken and results, shall be reported by the PSO at the General Assembly meetings (2 per year). Instructions to the partners on how to prevent the disclose of security sensitive information shall also be delivered during these meetings.

2.4 Ethics aspects

No harmful ethics impact is expected from the project activities on the environment and society. On the contrary, AI4CYBER works towards improved protection of critical services in Europe, which will benefit the whole society.

Moreover, the WP6 of the project researches on how AI4CYBER solution shall respect the principles of responsible Artificial Intelligence, including explainability, fairness and security of the AI employed in the AI4CYBER services to be developed in the project.

The project manager is in charge of monitoring the ethics are respected in all project activities and deliverables, to ensure their compliance with relevant ethical standards, particularly the Responsible Research and Innovation (RRI) framework. In addition, the ethics principles in the ALLEA (All European Academies)'s European Code of Conduct for Research Integrity [9] shall be respected when conducting the research. All project participants shall collaborate with the project manager and notify of any misconduct, misuse, or issue that they may identify in the project.

The Informed consent designed in AI4CYBER to guarantee all project participants, including the partners' staff as well as External Advisory Board are informed about the purpose of the project research as well as about their rights when participating in the project activities is shown in Appendix B.

In addition, with the aim to prevent any negative impact from the datasets shared in the project, the DMP reviews will make sure no real personal information is contained in the datasets handled or produced in the project, and to do so the project manager will coordinate the implementation of the necessary anonymisation techniques if needed.

Ethical aspects shall also be considered when describing the datasets used and generated in the project. The dataset characterization shall include a description of any relevant ethical and legal aspect that should be taken into account, for example the existence of any bias (of gender, race, etc.) already identified in the data.

2.5 Data Management Roles and Responsibilities

Figure 1 presents the data management responsibility structure of the project. As it can be seen in the figure, all partners in AI4CYBER consortium are joint data controllers since they jointly participate in the decisions about personal data processing in the project.

As introduced in Section 2.3.1, each partner has appointed a Data Protection Officer (DPO) who is the primary contact for the personal data protection issues in the project. Being the project coordinator, TECNALIA is responsible for appointing the AI4CYBER project DPO. The acting DPO of TECNALIA is the reference for handling any inquire or request that may arise from both internal and external data subjects around the processing of their personal data. The project DPO oversees the managing the data protection issues together with the partners' appointed DPOs.

As explained in Section 2.3.2, there are four partners that have appointed a security responsible representative. Together with UOWM that has appointed the PSO, the partners PPC, HES and CXB have appointed a Security Advisory Board member who will support the PSO in security aspects management.

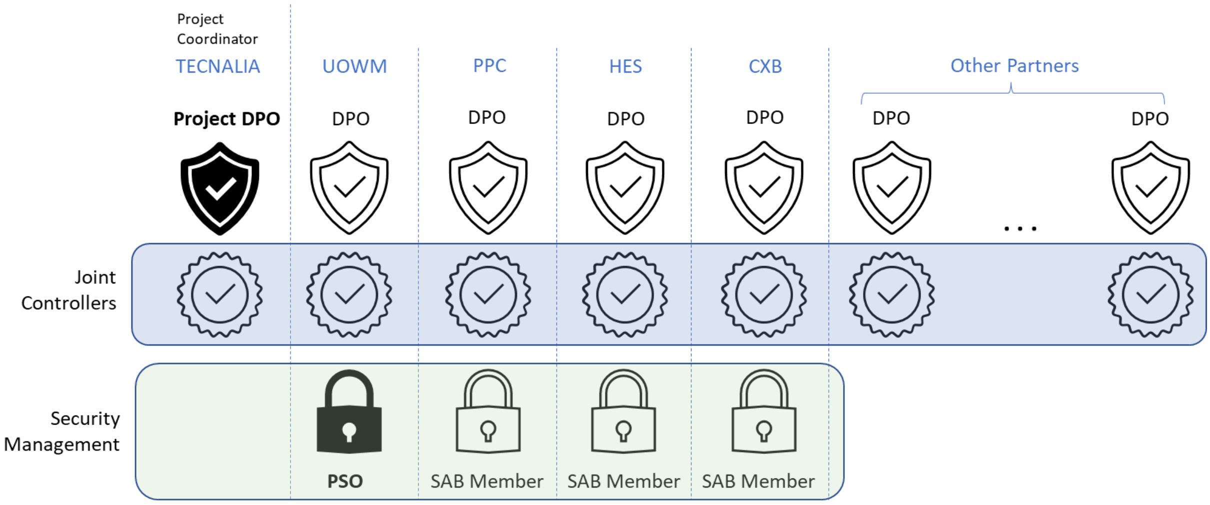


Figure 1: Data Management structure in AI4CYBER

Table 2 summarizes the roles and responsibilities related to the Data Management Plan.

Table 2: Data Management Roles in AI4CYBER

Title	Role	Partner	Person & contact
Project DPO	Responsible for reviewing the Data Management procedures and providing guidance as necessary.	TECNALIA	Javier Lerma (Javier.lerma@tecnalia.com)
Partner DPO	Responsible for reviewing the Data Management procedures and supporting the project DPO in handing privacy aspects in the project.	All Partners	See contact data in Section 2.3.1.

Joint Controller	Responsible for reviewing and following the Data Management Plan (DMP) and making sure the fulfilment of the partner responsibilities with respect to GDPR and other applicable regulation compliance.	All Partners	Appointed contact in the Consortium Agreement.
Quality Assurance Manager	Responsible for monitoring the overall project quality, including the data types generated in the project.	TECNALIA	Dr. Erkuden Rios (erkuden.rios@tecnalia.com)
Project Security Officer (PSO)	Responsible for guaranteeing that the rules on the handling of confidential or EU classified information and applicable security procedures are respected and avoid security issues in project deliverables and other data types.	UOWM	Prof. Panagiotis Sarigiannidis (psarigiannidis@uowm.gr) See section 2.3.2
Security Advisory Board Member	Responsible for advising the PSO, implementing security procedures and defining measures to protect the project deliverables and any other information/data where security issues are identified.	PPC, HES, CXB	See Members in Section 2.3.2

2.6 Monitoring and update of the DMP

The DMP plan presented herein is the initial version of the plan, which will be periodically reviewed by the AICYBER DPO and updated by all partners as necessary as the project progresses.

The DMP shall be updated in case any of the aspects in the plan need to be improved. Therefore, following updates may be required during the project lifetime as the project outcomes and data types are generated:

- Updates to the defined data types
- Updates to include additional data types
- Updates to research datasets currently identified
- Updates to include additional datasets used in the project
- Updates to include datasets produced in the project
- Updates to data security management procedures or related responsibilities
- Updates to data privacy management or related responsibilities

When revisiting the DMP, the following versions of the plan document will clearly identify the updates undergone.

In order to enable continuous monitoring and smooth contributions of the partners to the DMP, the interim version of the DMP will be made accessible in the project SharePoint so as all the partners can contribute to the datasets handled by them and can propose updates in form of comments and update suggestions in the document.

The periodic meetings of the consortium will allow to review the updates of the DMP and more formally, the status of the datasets will be reviewed during the plenary meetings of the project.

Figure 2 illustrates the timeline of the initially planned updates of the DMP. As it can be seen in the figure, the DMP will be regularly checked during the plenary meetings of the project, marked in the figure as General Assemblies (GA) in yellow. A formal major evaluation check point of the interim DMP is planned to be made during the GA of M26, once the initial integrated version of AI4CYBER solution has been demonstrated and the complete evaluation analysis finished. Please note that deliverable D1.4 corresponds to the final version of the DMP to be delivered at the end of the project.

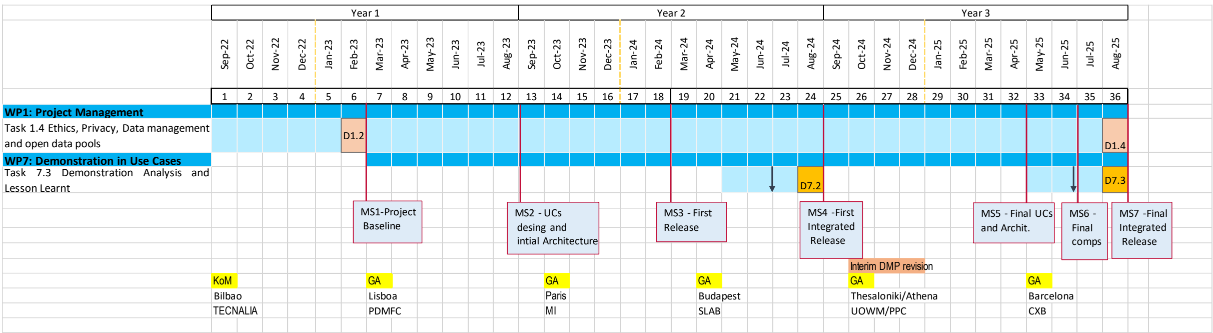


Figure 2: Data Management Plan updates

3 AI4CYBER research datasets

In this section we provide the list of research datasets that the AI4CYBER partners are using or plan to use in the project. The current list includes all datasets initially identified by the partners as useful in their research. Nevertheless, additional datasets may endorse the list as the project research progresses.

While in the present D1.2 only the description of currently available research datasets that are expected to be used in the project is included, in the final version of the DMP, i.e., the D1.4, due by the end of the project, the current list may be updated, and the description of datasets generated in AI4CYBER research activities will also be added. The rationale for this is that at the time of writing D1.2 the scientific research work in AI4CYBER is focused on the analysis of state-of-the-art and design of the AI4CYBER solution, while the solution development is not started yet.

To facilitate describing the datasets in a standard way, the project has defined a common template in form of a table that allows to identify and characterise the datasets. Appendix A provides the legend for the fields in the template.

3.1 Dataset for AIFix4SecCode

Dataset Name	**Dataset4AIFix4SecCode**
Dataset Summary	
Dataset Description	The Dataset for AIFix4SecCode contains bug-fix pairs for the following warnings: SonarQube: • S1444 • S2384 Spotbugs: • EI_EXPOSE_REP • EI_EXPOSE_REP2 • FI_PUBLIC_SHOULD_BE_PROTECTED • MS_EXPOSE_REP • MS_MUTABLE_ARRAY • MS_MUTABLE_COLLECTION • MS_MUTABLE_COLLECTION_PKGPROTECT • MS_SHOULD_BE_FINAL • NP_NULL_ON_SOME_PATH • NP_NULL_ON_SOME_PATH_EXCEPTION • NP_NULL_PARAM_DEREF • NP_NULL_PARAM_DEREF_ALL_TARGETS_DANGEROUS • NP_NULL_PARAM_DEREF_NONVIRTUAL • SQL_NONCONSTANT_STRING_PASSED_TO_EXECUTE • XSS_REQUEST_PARAMETER_TO_SERVLET_WRITER
Dataset Purpose	Provide learning set for bug fixing
Dataset Type/Format	JSON metadata + source code (bug-fix pairs)
Re-use of existing data	Yes, as an ongoing process we are extending this dataset.
Dataset Origin	First created during the AssureMOSS project
Dataset Collection	Manually validated from automatically searched content
Dataset Size	15,7 MB, containing 1204 bug-fix pairs
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP3
Relevant Tasks	T3.1, T3.2
Relevant Deliverables	D3.1, D3.3
Relevant Use Cases	UC2, UC3
Partners Services and Responsibilities	

Partner Owner	SLAB
Partners responsible for data collecting	SLAB
Partners responsible for data analysis	SLAB
Partners responsible for data storage	SLAB
FAIR Principles	
FAIR Data - findability	
Discoverable and Identifiable Data	Yes: uploaded to Zenodo
Naming Convention	The naming convention adopted for the Dataset.
Versioning	Yes, Semantic Versioning 2.0.0
Search Keywords	Yes, spotbugs, sonarqube, patch, bug, fix, manual
Metadata	No
FAIR Data – accessibility	
Dataset Openly Accessible	Yes
Methods/Tools for accessing the data	Zenodo, Github
Access Restrictions	Open access
Repository	https://github.com/assuremoss/Dataset4AIFix4SecCode https://zenodo.org/record/6669965
FAIR Data - interoperability	
Interoperability	Machine processible via the JSON metadata
Standards	-
FAIR Data – reusability	
Licence	Apache 2.0
Data sharing terms	-
Dataset sharing after the end of the project	-
Preservation and update of the dataset after the project	Will be extended during the project
Re-use timeframe	Relevant for a longer time frame
Allocation of Resources	
Cost for making the Dataset FAIR	No additional cost, EU-funded
Costs for the preservation (including backup) and the update of the Dataset	No cost
Data Security and privacy	
Security Measures	The dataset does not refer to any actual entity participating in a real industrial environment, presents open source data from Github.
Privacy Measures	The dataset does not refer to any actual entity participating in a real industrial environment, presents open source data from Github.
Ethics	
Ethical and Legal Aspects	-
Other Issues	-

3.2 JiraMiner dataset

Dataset Name	**JiraMiner**
Dataset Summary	
Dataset Description	The JiraMiner dataset gathers labelled data from Jira for Apache projects and connects them with the commits associated in Github.
Dataset Purpose	Provide learning set for bug fixing
Dataset Type/Format	JSON metadata + source code
Re-use of existing data	No, this is new to the project
Dataset Origin	First created during the AI4Cyber project
Dataset Collection	Collected with automated mining
Dataset Size	TBD
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP3
Relevant Tasks	T3.1, T3.2
Relevant Deliverables	D3.1, D3.3
Relevant Use Cases	UC2, UC3
Partners Services and Responsibilities	
Partner Owner	SLAB
Partners responsible for data collecting	SLAB
Partners responsible for data analysis	SLAB
Partners responsible for data storage	SLAB
FAIR Principles	
FAIR Data - findability	
Discoverable and Identifiable Data	Yes: will be uploaded to Zenodo
Naming Convention	The naming convention adopted for the Dataset.
Versioning	Yes, Semantic Versioning 2.0.0
Search Keywords	Yes, jira, github, issues, security, bug, fix
Metadata	No
FAIR Data – accessibility	
Dataset Openly Accessible	Yes
Methods/Tools for accessing the data	Zenodo, Github
Access Restrictions	Open access
Repository	TBD
FAIR Data - interoperability	
Interoperability	Machine processible via the JSON metadata
Standards	-
FAIR Data – reusability	
Licence	Apache 2.0
Data sharing terms	-
Dataset sharing after the end of the project	-
Preservation and update of the dataset after the project	Will be extended during the project
Re-use timeframe	Relevant for a longer time frame
Allocation of Resources	

Cost for making the Dataset FAIR	No additional cost, EU-funded
Costs for the preservation (including backup) and the update of the Dataset	No cost
Data Security and privacy	
Security Measures	The dataset does not refer to any actual entity participating in a real industrial environment, presents open source data from Github.
Privacy Measures	The dataset does not refer to any actual entity participating in a real industrial environment, presents open source data from Github.
Ethics	
Ethical and Legal Aspects	-
Other Issues	-

3.3 IEC 60870-5-104 Intrusion Detection Dataset

Dataset Name	**IEC 60870-5-104 Intrusion Detection Dataset**
Dataset Summary	
Dataset Description	The IEC 60870-5-104 Intrusion Detection Dataset includes various pcap and Common Separate Value (CSV) files related to IEC 60870-5-104 attacks. The CSV files refer to (a) Transmission Control Protocol/Internet Protocol (TCP/IP) and (b) IEC 60870-5-104 Intrusion Detection Dataset flow statistics related to the IEC 60870-5-104 packets of the pcap files. Each network flow is labelled as normal or the corresponding cyberattack.
Dataset Purpose	Artificial Intelligence (AI)-based IEC 60870-5-104 intrusion/anomaly detection
Dataset Type/Format	.7zip, .pcap/.pcapng, .csv, .txt
Re-use of existing data	No
Dataset Origin	This dataset was implemented by UOWM in the context of SDN-microSENSE project.
Dataset Collection	This dataset was collected in a real testbed consisting of multiple physical and virtual industrial devices, such as Remote Terminal Units (RTUs) and Programmable Logic Controllers (PLCs) using the IEC 60870-5-104 protocol.
Dataset Size	8GB
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP4, WP6
Relevant Tasks	Task 4.1: Deep Anomaly Detection Task 4.2: Federated Learning-enhanced Detection Task 4.3: Root Cause Analysis and Alert Triage Task 6.1: xAI of AI4Cyber Services
Relevant Deliverables	D4.1: Cyber Intelligence and Detection of advanced and AI-powered attacks - Initial version D4.2: Cyber Intelligence and Detection of advanced and AI-powered attacks - Final version D6.1: Models and methods for Trustworthiness of AI4CYBER services – Initial version

	D6.2: Models and methods for Trustworthiness of AI4CYBER services – Final version
Relevant Use Cases	Use Case 1: Detection and Mitigation of AI-powered Attacks against the Energy Sector
Partners Services and Responsibilities	
Partner Owner	UOWM
Partners responsible for data collecting	UOWM
Partners responsible for data analysis	Information on this is available to IEEE Dataport and Zenodo
Partners responsible for data storage	UOWM
FAIR Principles	
FAIR Data - findability	
Discoverable and Identifiable Data	Information on this is available to IEEE Dataport and Zenodo
Naming Convention	IEC 60870-5-104 Intrusion Detection Dataset
Versioning	V1
Search Keywords	Anomaly Detection, ICS, IEC 60870-5-104, IEC 104, Intrusion Detection, SCADA
Metadata	• **Relevant Publication**: P. Radoglou-Grammatikis, K. Rompolos, P. Sarigiannidis, V. Argyriou, T. Lagkas, A. Sarigiannidis, S. Goudos and S. Wan, "Modeling, Detecting, and Mitigating Threats Against Industrial Healthcare Systems: A Combined Software Defined Networking and Reinforcement Learning Approach", in IEEE Transactions on Industrial Informatics, vol. 18, no. 3, pp. 2041-2052, March 2022 • **Last Update**: Wed, 01/04/2023 - 13:05 • **DOI**: 10.21227/fj7s-f281 • **Data Format** *.csv, *.pcap, *.zip
FAIR Data – accessibility	
Dataset Openly Accessible	Yes, available in Zenodo
Methods/Tools for accessing the data	Tcpdump, Wireshark, Tshark, Pandas,
Access Restrictions	No – It will be publicly available
Repository	IEEE Dataport and Zenodo
FAIR Data - interoperability	
Interoperability	
Standards	Relevant Standards: IEC 60870-5-104
FAIR Data – reusability	
Licence	Creative Commons Attribution
Data sharing terms	The users of this dataset are kindly asked to cite the following paper: *P. Radoglou-Grammatikis, K. Rompolos, P. Sarigiannidis, V. Argyriou, T. Lagkas, A. Sarigiannidis, S. Goudos and S. Wan, "Modeling, Detecting, and Mitigating Threats Against Industrial Healthcare Systems: A Combined Software Defined Networking and Reinforcement Learning Approach", in IEEE Transactions on Industrial Informatics, vol. 18, no. 3, pp. 2041-2052, March 2022, doi: 10.1109/TII.2021.3093905.*
Dataset sharing after the end of the project	Yes

Preservation and update of the dataset after the project	No
Re-use timeframe	-
Allocation of Resources	
Cost for making the Dataset FAIR	No Cost
Costs for the preservation (including backup) and the update of the Dataset	No Cost
Data Security and privacy	
Security Measures	The dataset does not refer to any actual entity participating in a real industrial environment.
Privacy Measures	The dataset does not refer to any actual entity participating in a real industrial environment.
Ethics	
Ethical and Legal Aspects	No ethical and legal aspects. The dataset does not refer to any actual entity participating in a real industrial environment.
Other Issues	No

3.4 OCPP_DDOS_FakeEVs_Dataset

Dataset Name	**OCPP_DDOS_FakeEVs_Dataset**
Dataset Summary	
Dataset Description	This dataset contains samples of a Distributed Denial of Service attack what was conducted in the premises of the Internet of Things (IoT) laboratory of the PPC Innovation Hub. The topology consists of a Charging Station Management Platform (CSMS) based on the SteVe platform (https://github.com/steve-community/steve) and a Virtual Machine (VM) which represents the attacker. The attacker utilises a Python script and the `websockets` library in order to launch multiple requests (which follow the OCPP v1.6-J standard) that imitate multiple Electric Vehicle Charging Stations (EVCSs) which are trying to register to the CSMS. The attack aims to exhaust the available resources of the CSMS, hence, rendering it unresponsive and unavailable for legitimate EVCSs.
Dataset Purpose	Purpose of this dataset is to support the research concerning the study and detection of DDoS attacks against CSMSs under the OCPP v1.6-J standard.
Dataset Type/Format	pcap
Re-use of existing data	Yes
Dataset Origin	H2020 ELECTRON project (https://cordis.europa.eu/project/id/101021936)
Dataset Collection	The dataset was collected by using the `tshark` software on the VM that hosted CSMS.
Dataset Size	1.5 GB
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP3, WP4
Relevant Tasks	T3.4, T4.1, T4.2
Relevant Deliverables	D3.4, D4.1, D4.2
Relevant Use Cases	UC1: Detection and Mitigation of AI-powered Attacks against the Energy Sector
Partners Services and Responsibilities	

Partner Owner	PPC
Partners responsible for data collecting	PPC
Partners responsible for data analysis	UOWM, PPC
Partners responsible for data storage	PPC
FAIR Principles	
FAIR Data - findability	
Discoverable and Identifiable Data	No
Naming Convention	Each file inside the dataset will follow the following convention:<ISO_Date>_<Network_Protocol>_<Category>_<Sub category>_<Part> • ISO_Date: The date when the data capture took place, in the following format: YYYMMDD. • Network_Protocol: Specifies the application-layer network protocol that is relevant to the attack. For example, OCPP. • Category: The category of the attack, for example: DDoS, FDI, or DOS. • Subcategory: Specifies the subcategory of the attack, for example “FakeEVCS”. • Part: If the dataset consists of multiple files, this is the sequence number of the individual file.
Versioning	No
Search Keywords	ocpp, websockets, ocpp1.6, evcs, cpms, csms, dos, ddos
Metadata	No
FAIR Data – accessibility	
Dataset Openly Accessible	Yes
Methods/Tools for accessing the data	Any file archiever (e.g., 7Zip), Wireshark/tshark, tcpdump, programing libraries that can process pcap(ng) files (e.g., Python Scapy)
Access Restrictions	N/A
Repository	IEEE DataPort, Zenodo
FAIR Data - interoperability	
Interoperability	The Packet Capture (pcap) file format is adopted for the dataset, allowing interoperability with multiple tools and software that read and process pcap files.
Standards	N/A
FAIR Data – reusability	
Licence	CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/)
Data sharing terms	The sharing terms are specified by the adopted license.
Dataset sharing after the end of the project	The sharing terms are specified by the adopted license.
Preservation and update of the dataset after the project	N/A
Re-use timeframe	N/A
Allocation of Resources	
Cost for making the Dataset FAIR	N/A
Costs for the preservation (including backup) and the update of the Dataset	N/A
Data Security and privacy	
Security Measures	N/A
Privacy Measures	IP addresses and potential Personal Identifiable Information (PII) are anonymised.

Ethics	
Ethical and Legal Aspects	N/A
Other Issues	N/A

3.5 AI4AppSec dataset

Dataset Name	**AI4AppSec**
Dataset Summary	
Dataset Description	The AI4AppSec dataset is a comprehensive collection of labelled data obtained from the application logs of various systems such as SQL, MongoDB, Apache, NGINX, and similar technologies that are commonly used in the healthcare industry. The logs were gathered from a simulated Healthcare Environment (HES) and have been meticulously labelled to facilitate efficient analysis and modelling.
Dataset Purpose	Provide learning set for security, system and application logs within a system (e.g., server, client, services)
Dataset Type/Format	CSV, JSON, XLS
Re-use of existing data	No, this is new to the project
Dataset Origin	First created during the AI4Cyber project
Dataset Collection	Collected with automated mining using open-source libraries, PDMFC agents for the collection and PDMFC rules for labelling the data.
Dataset Size	TBD
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP3, WP4, WP5
Relevant Tasks	T3.1, T3.2, T4.1, T4.2, T5.4, T5.5
Relevant Deliverables	D3.1, D3.3, D4.1, D4.2, D5.1, D5.2
Relevant Use Cases	UC2, UC3
Partners Services and Responsibilities	
Partner Owner	HES
Partners responsible for data collecting	PDMFC
Partners responsible for data analysis	TBD
Partners responsible for data storage	PDMFC/HES
FAIR Principles	
FAIR Data - findability	
Discoverable and Identifiable Data	Yes: will be uploaded to GitHub, Zenodo or Kaggle
Naming Convention	The naming convention adopted for the Dataset.
Versioning	Yes, Semantic Versioning 2.0.0
Search Keywords	Yes, authentication logs, IDS, host IDS, ML HIDS, Linux, Servers, Application Security
Metadata	No
FAIR Data – accessibility	
Dataset Openly Accessible	Yes, TBD
Methods/Tools for accessing the data	Python, Ruby or similar, common SIEM tools or data analysis.
Access Restrictions	Open upon request.
Repository	TBD
FAIR Data - interoperability	

Interoperability	Machine processible via the CSV, or JSON metadata
Standards	-
FAIR Data – reusability	
Licence	Apache Licence 2.0
Data sharing terms	Reference to the project and to the involved partners.
Dataset sharing after the end of the project	Yes
Preservation and update of the dataset after the project	Will be extended during the project
Re-use timeframe	Relevant for a longer time frame
Allocation of Resources	
Cost for making the Dataset FAIR	No additional cost, EU-funded
Costs for the preservation (including backup) and the update of the Dataset	No cost
Data Security and privacy	
Security Measures	The dataset does not refer to any actual entity participating in a real industrial environment or is anonymized.
Privacy Measures	The dataset does not refer to any actual entity participating in a real industrial environment.
Ethics	
Ethical and Legal Aspects	-
Other Issues	-

3.6 AI4AuthLog dataset

Dataset Name	**AI4AuthLog**
Dataset Summary	
Dataset Description	The AuthLog dataset collects labelled data from the host systems of the simulated Healthcare environment (HES). PDMFC can host the procedure on other pilots after communication.
Dataset Purpose	Provide learning set for authentications within a system (e.g., server, client, services)
Dataset Type/Format	CSV, JSON or PCAP files
Re-use of existing data	No, this is new to the project
Dataset Origin	First created during the AI4Cyber project
Dataset Collection	Collected with automated mining using open-source libraries, PDMFC agents for the collection and PDMFC rules for labelling the data.
Dataset Size	TBD
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP3, WP4, WP5
Relevant Tasks	T3.1, T3.2, T4.1, T4.2, T5.4, T5.5
Relevant Deliverables	D3.1, D3.3, D4.1, D4.2, D5.1, D5.2
Relevant Use Cases	UC2, UC3
Partners Services and Responsibilities	
Partner Owner	HES
Partners responsible for data collecting	PDMFC
Partners responsible for data analysis	TBD
Partners responsible for data storage	PDMFC/HES

FAIR Principles	
FAIR Data - findability	
Discoverable and Identifiable Data	Yes: will be uploaded to GitHub, Zenodo or Kaggle
Naming Convention	The naming convention adopted for the Dataset.
Versioning	Yes, Semantic Versioning 2.0.0
Search Keywords	Yes, authentication logs, IDS, host IDS, ML HIDS, Linux, Servers
Metadata	No
FAIR Data – accessibility	
Dataset Openly Accessible	Yes, TBD
Methods/Tools for accessing the data	Python, Ruby or similar, common SIEM tools or data analysis.
Access Restrictions	Open after request.
Repository	TBD
FAIR Data - interoperability	
Interoperability	Machine processible via the CSV, or JSON metadata
Standards	-
FAIR Data – reusability	
Licence	Apache Licence 2.0
Data sharing terms	Reference to the project and to the involved partners.
Dataset sharing after the end of the project	Yes
Preservation and update of the dataset after the project	Will be extended during the project
Re-use timeframe	Relevant for a longer time frame
Allocation of Resources	
Cost for making the Dataset FAIR	No additional cost, EU-funded
Costs for the preservation (including backup) and the update of the Dataset	No cost
Data Security and privacy	
Security Measures	The dataset does not refer to any actual entity participating in a real industrial environment or is anonymized.
Privacy Measures	The dataset does not refer to any actual entity participating in a real industrial environment.
Ethics	
Ethical and Legal Aspects	-
Other Issues	-

3.7 AI4NetHealth dataset

Dataset Name	**AI4NetHealth**
Dataset Summary	
Dataset Description	The AI4NetHealth dataset collects labelled data from the network traffic of the simulated Healthcare environment (HES).
Dataset Purpose	Provide learning set for network anomalies
Dataset Type/Format	CSV, JSON or PCAP files
Re-use of existing data	No, this is new to the project
Dataset Origin	First created during the AI4Cyber project

Dataset Collection	Collected with automated mining using open-source libraries
Dataset Size	TBD
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP3, WP4, WP5
Relevant Tasks	T3.1, T3.2, T4.1, T4.2, T5.4, T5.5
Relevant Deliverables	D3.1, D3.3, D4.1, D4.2, D5.1, D5.2
Relevant Use Cases	UC2, UC3
Partners Services and Responsibilities	
Partner Owner	HES
Partners responsible for data collecting	PDMFC
Partners responsible for data analysis	UOWM
Partners responsible for data storage	PDMFC/HES
FAIR Principles	
FAIR Data - findability	
Discoverable and Identifiable Data	Yes: will be uploaded to GitHub, Zenodo or Kaggle
Naming Convention	The naming convention adopted for the Dataset.
Versioning	Yes, Semantic Versioning 2.0.0
Search Keywords	Yes, network traffic, pcap, IDS, ML IDS, network behaviour
Metadata	No
FAIR Data – accessibility	
Dataset Openly Accessible	Yes, TBD
Methods/Tools for accessing the data	Python, Ruby or similar, common SIEM tools or data analysis.
Access Restrictions	Open after request.
Repository	TBD
FAIR Data - interoperability	
Interoperability	Machine processible via the CSV, or JSON metadata
Standards	-
FAIR Data – reusability	
Licence	Apache Licence 2.0
Data sharing terms	Reference to the project and to the involved partners.
Dataset sharing after the end of the project	Yes
Preservation and update of the dataset after the project	Will be extended during the project
Re-use timeframe	Relevant for a longer time frame
Allocation of Resources	
Cost for making the Dataset FAIR	No additional cost, EU-funded
Costs for the preservation (including backup) and the update of the Dataset	No cost
Data Security and privacy	
Security Measures	The dataset does not refer to any actual entity participating in a real industrial environment or is anonymized.
Privacy Measures	The dataset does not refer to any actual entity participating in a real industrial environment.
Ethics	
Ethical and Legal Aspects	-

Other Issues	-

3.8 AI4WinEvent dataset

Dataset Name	AI4WinEvent
Dataset Summary	
Dataset Description	The AI4WinEvent dataset collects labelled data from Windows host systems of the simulated Healthcare environment (HES). PDMFC can host the procedure on other pilots after communication.
Dataset Purpose	Provide learning set for security, system and application logs within a system (e.g., server, client, services)
Dataset Type/Format	CSV, JSON, XLS
Re-use of existing data	No, this is new to the project
Dataset Origin	First created during the AI4Cyber project
Dataset Collection	Collected with automated mining using open-source libraries, PDMFC agents for the collection and PDMFC rules for labelling the data.
Dataset Size	TBD
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP3, WP4, WP5
Relevant Tasks	T3.1, T3.2, T4.1, T4.2, T5.4, T5.5
Relevant Deliverables	D3.1, D3.3, D4.1, D4.2, D5.1, D5.2
Relevant Use Cases	UC2, UC3
Partners Services and Responsibilities	
Partner Owner	HES
Partners responsible for data collecting	PDMFC
Partners responsible for data analysis	TBD
Partners responsible for data storage	PDMFC/HES
FAIR Principles	
FAIR Data - findability	
Discoverable and Identifiable Data	Yes: will be uploaded to GitHub, Zenodo or Kaggle
Naming Convention	The naming convention adopted for the Dataset.
Versioning	Yes, Semantic Versioning 2.0.0
Search Keywords	Yes, authentication logs, IDS, host IDS, ML HIDS, Linux, Servers
Metadata	No
FAIR Data – accessibility	
Dataset Openly Accessible	Yes, TBD
Methods/Tools for accessing the data	Python, Ruby or similar, common SIEM tools or data analysis.
Access Restrictions	Open upon request.
Repository	TBD
FAIR Data - interoperability	
Interoperability	Machine processible via the CSV, or JSON metadata
Standards	-
FAIR Data – reusability	

Licence	Apache Licence 2.0
Data sharing terms	Reference to the project and to the involved partners.
Dataset sharing after the end of the project	Yes
Preservation and update of the dataset after the project	Will be extended during the project
Re-use timeframe	Relevant for a longer time frame
Allocation of Resources	
Cost for making the Dataset FAIR	No additional cost, EU-funded
Costs for the preservation (including backup) and the update of the Dataset	No cost
Data Security and privacy	
Security Measures	The dataset does not refer to any actual entity participating in a real industrial environment or is anonymized.
Privacy Measures	The dataset does not refer to any actual entity participating in a real industrial environment.
Ethics	
Ethical and Legal Aspects	-
Other Issues	-

3.9 CXB-InsiderThreat-AzureAD dataset

Dataset Name	CXB-InsiderThreat-AzureAD
Dataset Summary	
Dataset Description	This dataset contains samples of the information collected by CXB's cloud-based access and directory management solution. It contains information about the accesses of 3rd party providers to CXB applications. The dataset is based on logs that are automatically generated by the solution in a time-stamped file that provides an audit trail for sign-in and sign-out events that it monitors. The most relevant information components those logs contain are "who", the identity of the user doing the sign-in, "how", the client/application used for the access, and "what", the target/resource accessed by the identity.
Dataset Purpose	Purpose of this dataset is to identify any behaviour not normal from 3rd-party providers' users that are accessing CXB infrastructure and applications, that can lead to identify any provider impersonation by malicious users and its network intrusion, which indeed could derive into additional critical threats like a ransomware or sensitive data leakages.
Dataset Type/Format	Log files (JSON/CSV)
Re-use of existing data	Yes
Dataset Origin	CXB internal systems
Dataset Collection	The dataset will be collected by CXB, by means of its internal tools.
Dataset Size	TBD
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP4, WP5
Relevant Tasks	T4.1, T4.2, T4.3, T5.1, T5.2
Relevant Deliverables	D4.1, D4.2, D5.1, D5.2
Relevant Use Cases	UC2: Robustness and autonomous adaptation of Banking applications to face AI-powered attacks.

Partners Services and Responsibilities	
Partner Owner	CXB
Partners responsible for data collecting	CXB
Partners responsible for data analysis	UOWM, TECNALIA, MI, CXB
Partners responsible for data storage	CXB
FAIR Principles	
FAIR Data – findability	
Discoverable and Identifiable Data	No
Naming Convention	Active Directory Audit Logs
Versioning	No
Search Keywords	Active directory, access, logs, intrusion, malicious user, insider threat
Metadata	No
FAIR Data – accessibility	
Dataset Openly Accessible	No
Methods/Tools for accessing the data	Through CXB systems.
Access Restrictions	Only accessible to previously authorized users.
Repository	CXB repository.
FAIR Data - interoperability	
Interoperability	Collected dataset will follow standard logs format
Standards	RFC5424, RFC5425
FAIR Data – reusability	
Licence	TBD
Data sharing terms	TBD
Dataset sharing after the end of the project	Anonymised dataset can be considered at the end of the project.
Preservation and update of the dataset after the project	N/A
Re-use timeframe	N/A
Allocation of Resources	
Cost for making the Dataset FAIR	N/A
Costs for the preservation (including backup) and the update of the Dataset	N/A
Data Security and privacy	
Security Measures	Original dataset is only accessible through accessing CXB premises and passing all the standard information security controls established by CXB.
Privacy Measures	IP addresses and potential Personal Identifiable Information (PII) should be anonymised.
Ethics	
Ethical and Legal Aspects	N/A
Other Issues	N/A

3.10 CXB-InsiderThreat-CyberArk dataset

Dataset Name	**CXB-InsiderThreat-CyberArk**
Dataset Summary	

Dataset Description	This dataset contains samples of the information collected by CXB's PAM (Privileged Access Management) solution, used to control, and strengthen the access of 3rd party providers to some CXB applications and environments that require additional security measures. The dataset is based on logs that are automatically generated by the solution in a time-stamped file that provides an audit trail for accesses to the environment.
Dataset Purpose	Purpose of this dataset is to identify any behaviour not normal from 3rd-party providers' users that are accessing CXB infrastructure and more concretely those ones that are accessing the SWIFT (Society for Worldwide Interbank Financial Telecommunications) environment (client to connect to a secure network and exchange financial information with other banks and financial institutions). This environment is highly secured and any anomaly on its users' activity should be supervised and controlled.
Dataset Type/Format	Log files (JSON)
Re-use of existing data	Yes
Dataset Origin	CXB internal systems
Dataset Collection	The dataset will be collected by CXB, by means of its internal tools.
Dataset Size	TBD
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP4, WP5
Relevant Tasks	T4.1, T4.2, T4.3, T5.1, T5.2
Relevant Deliverables	D4.1, D4.2, D5.1, D5.2
Relevant Use Cases	UC2: Robustness and autonomous adaptation of Banking applications to face AI-powered attacks.
Partners Services and Responsibilities	
Partner Owner	CXB
Partners responsible for data collecting	CXB
Partners responsible for data analysis	UOWM, TECNALIA, MI, CXB
Partners responsible for data storage	CXB
FAIR Principles	
FAIR Data – findability	
Discoverable and Identifiable Data	No
Naming Convention	Syslog format.
Versioning	No
Search Keywords	PAM, cyberark, access, logs, intrusion, malicious user, insider threat
Metadata	No
FAIR Data – accessibility	
Dataset Openly Accessible	No
Methods/Tools for accessing the data	Through CXB systems.
Access Restrictions	Only accessible to previously authorized users.
Repository	CXB repository.
FAIR Data - interoperability	
Interoperability	Collected dataset will follow standard logs format
Standards	RFC5424, RFC5425
FAIR Data – reusability	
Licence	TBD

Data sharing terms	TBD
Dataset sharing after the end of the project	Anonymised dataset can be considered at the end of the project.
Preservation and update of the dataset after the project	N/A
Re-use timeframe	N/A
Allocation of Resources	
Cost for making the Dataset FAIR	N/A
Costs for the preservation (including backup) and the update of the Dataset	N/A
Data Security and privacy	
Security Measures	Original dataset is only accessible through accessing CXB premises and passing all the standard information security controls established by CXB.
Privacy Measures	IP addresses and potential Personal Identifiable Information (PII) should be anonymised.
Ethics	
Ethical and Legal Aspects	N/A
Other Issues	N/A

3.11 CXB-InsiderThreat-Prisma dataset

Dataset Name	CXB-InsiderThreat-Prisma
Dataset Summary	
Dataset Description	This dataset contains samples of the information collected by CXB SASE solution, focusing on the accesses of 3rd party providers to CXB infrastructure and applications. The dataset is based on logs that are automatically generated by the solution in a time-stamped file that provides an audit trail for system events or network traffic events that it monitors. Log entries contain artifacts, which are properties, activities, or behaviors associated with the logged event, such as the application type or the IP address of the user.
Dataset Purpose	Purpose of this dataset is to identify any behaviour not normal from 3rd-party providers' users that are accessing CXB infrastructure and more concretely those ones that are accessing the Financial Terminal environment (application that provides most of the services that branch offices employees need to perform their day-to-day activities with clients).
Dataset Type/Format	Log files (CSV)
Re-use of existing data	Yes
Dataset Origin	CXB internal systems
Dataset Collection	The dataset will be collected by CXB, by means of its internal tools.
Dataset Size	TBD
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP4, WP5
Relevant Tasks	T4.1, T4.2, T4.3, T5.1, T5.2
Relevant Deliverables	D4.1, D4.2, D5.1, D5.2
Relevant Use Cases	UC2: Robustness and autonomous adaptation of Banking applications to face AI-powered attacks.
Partners Services and Responsibilities	
Partner Owner	CXB

Partners responsible for data collecting	CXB
Partners responsible for data analysis	UOWM, TECNALIA, MI, CXB
Partners responsible for data storage	CXB
FAIR Principles	
FAIR Data – findability	
Discoverable and Identifiable Data	No
Naming Convention	Syslog format. Traffic, System, User-ID, Authentication logs.
Versioning	No
Search Keywords	SASE, access, logs, intrusion, malicious user, insider threat
Metadata	No
FAIR Data – accessibility	
Dataset Openly Accessible	No
Methods/Tools for accessing the data	Through CXB systems.
Access Restrictions	Only accessible to previously authorized users.
Repository	CXB repository.
FAIR Data - interoperability	
Interoperability	Collected dataset will follow standard logs format
Standards	RFC5424, RFC5425
FAIR Data – reusability	
Licence	TBD
Data sharing terms	TBD
Dataset sharing after the end of the project	Anonymised dataset can be considered at the end of the project.
Preservation and update of the dataset after the project	N/A
Re-use timeframe	N/A
Allocation of Resources	
Cost for making the Dataset FAIR	N/A
Costs for the preservation (including backup) and the update of the Dataset	N/A
Data Security and privacy	
Security Measures	Original dataset is only accessible through accessing CXB premises and passing all the standard information security controls established by CXB.
Privacy Measures	IP addresses and potential Personal Identifiable Information (PII) should be anonymised.
Ethics	
Ethical and Legal Aspects	N/A
Other Issues	N/A

3.12 SIMARGL2021 dataset

Dataset Name	**SIMARGL2021**
Dataset Summary	
Dataset Description	Network Intrusion Detection Dataset
Dataset Purpose	Facilitating machine-learning-based network intrusion detection
Dataset Type/Format	Netflows in CSV format

Dataset Origin	Realistic traffic gathered in the H2020 SIMARGL project
Dataset Collection	Described in the paper available at: https://www.mdpi.com/1424-8220/21/13/4319
Dataset Size	Described in the paper available at: https://www.mdpi.com/1424-8220/21/13/4319
Frequency of the dataset collection	Described in the paper available at: https://www.mdpi.com/1424-8220/21/13/4319
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP6
Relevant Tasks	T6.1
Relevant Deliverables	D6.1
Relevant Use Cases	Explainability of AI-powered network intrusion detection
Partners Services and Responsibilities	
Partner Owner / Contact person	-
Partners responsible for data collecting	-
Partners responsible for data analysis	-
Partners responsible for data storage	-
FAIR Principles	
FAIR Data – accessibility	
Dataset Openly Accessible	Dataset is public:
Methods/Tools for accessing the data	https://www.kaggle.com/datasets/h2020simargl/simargl2021-network-intrusion-detection-dataset
Access Restrictions	-
Repository	https://www.kaggle.com/datasets/h2020simargl/simargl2021-network-intrusion-detection-dataset
FAIR Data – reusability	
Licence	CC BY 4.0
Data sharing terms	-
Dataset sharing after the end of the project	-
Preservation and update of the dataset after the project	-
Re-use timeframe	-
Allocation of Resources	
Cost for making the Dataset FAIR	N/A
Costs for the preservation (including backup) and the update of the Dataset	N/A
Data Security and privacy	
Data Security and privacy	
Security Measures	N/A
Privacy Measures	
Ethics	
Ethical and Legal Aspects	N/A
Other Issues	N/A

3.13 French COVID19 study

Dataset Name	French COVID19 study
Dataset Summary	
Dataset Description	The French COVID19 study Dataset includes various Common Separate Value (CSV) files related to the dynamics of COVID19 at local level in France during the beginning of 2021. Each CSV files contain time series at Department level in France about incidence rates, hospitalization rates, etc. The datasets will be used to demonstrate the application of xAI methods on ML model trained on time series
Dataset Purpose	Artificial Intelligence (AI)-Detect abnormal times series in a corpus of (multivariate) time series.
Dataset Type/Format	.csv
Re-use of existing data	No
Dataset Origin	This dataset was collected from available Open Data
Dataset Collection	This dataset was collected according to available Open Data
Dataset Size	~15 Mo
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP6
Relevant Tasks	Task 6.1: xAI of AI4Cyber Services
Relevant Deliverables	D6.1: Models and methods for Trustworthiness of AI4CYBER services – Initial version D6.2: Models and methods for Trustworthiness of AI4CYBER services – Final version
Relevant Use Cases	
Partners Services and Responsibilities	
Partner Owner	Thales
Partners responsible for data collecting	Thales
Partners responsible for data analysis	Thales
Partners responsible for data storage	Thales
FAIR Principles	
FAIR Data - findability	
Discoverable and Identifiable Data	
Naming Convention	
Versioning	V1
Search Keywords	Anomaly Detection, time series
Metadata	• Row: departments, columns; interest values at each time stamp • Data Format *.csv
FAIR Data – accessibility	
Dataset Openly Accessible	Yes, at the demand
Methods/Tools for accessing the data	Pandas,
Access Restrictions	No – It will be accessible at the demand
Repository	

FAIR Data - interoperability	
Interoperability	
Standards	
FAIR Data – reusability	
Licence	
Data sharing terms	
Dataset sharing after the end of the project	Yes
Preservation and update of the dataset after the project	No
Re-use timeframe	-
Allocation of Resources	
Cost for making the Dataset FAIR	
Costs for the preservation (including backup) and the update of the Dataset	
Data Security and privacy	
Security Measures	Data coming from Open Data information
Privacy Measures	Data coming from Open Data information
Ethics	
Ethical and Legal Aspects	No ethical and legal aspects.
Other Issues	No

4 Conclusion

This document presents the Data Management Plan of AI4CYBER which shall be us by project partners as guidance to fulfil the requirements and that the project should follow with regards to the management of data.

The *Ethics, Privacy, Data management and open data pools* task has defined the initial version of the DMP presented in this document. The plan describes all the data types handled by the project including deliverables, scientific publications, other publications, and research datasets. The plan establishes the procedures and criteria to address the relevant aspects of making project generated data FAIR (Findable, Accessible, Interoperable and Re-usable), including data accessibility conditions for data verification and re-use, and data curation and preservation measures. Furthermore, it describes the procedures to ensure adherence to relevant ethical standards (e.g., the RRI framework) and compliance with data protection legislation (i.e., the General Data Protection Regulation) in data management.

With the aim to increase the impact and enabling open science, the project will adhere to the principle of ‘as open as possible, as closed as necessary’. The partners aim at using and contributing to certified open repositories of data such as Zenodo and Open AIRE, where project open data will be deposited. Whenever possible, the green model for open access will be favoured for project scientific publications which will be stored also in OpenAIRE. In addition, all open data generated in the project will be published or referenced in the project website. For those not commercially IP protected datasets or other type of project outcomes, Creative Commons CCBY or CC-0 licenses will be favoured.

As part of the plan a template for describing the research datasets was defined, and a total of thirteen datasets that will be used in the project were already identified and described.

The DMP will be a living plan along the course of the project, and it will be reviewed in a regular basis, with a major review in M26. A final version of the plan will be delivered by the project (M36). It is envisioned that the final version will include the description of all the open datasets used and produced in the project research.

References

[1] "AI4CYBER," Ai4cyber.eu. [Online]. Available: https://ai4cyber.eu/. [Accessed: 28-02-2023].

[2] Open Research Europe, Available: https://open-research-europe.ec.europa.eu [Accessed: 28-02-2023].

[3] H2020 ELECTRON project, Available: https://cordis.europa.eu/project/id/101021936 [Accessed: 28-02-2023].

[4] European Commission, Template Horizon 2020 Data Management Plan (DMP) version 2.0, Available: https://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf [Accessed: 28-02-2023].

[5] European Commission, H2020 Programme - Guidelines on FAIR Data Management in Horizon 2020, version 3.0, 2016, Available: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf [Accessed: 28-02-2023].

[6] European Commission, Guidelines on Open access to publications and research data in Horizon 2020, Available: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm [Accessed: 28-02-2023].

[7] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18 [Accessed: 28-02-2023].

[8] European Commission, Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Available: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf [Accessed: 28-02-2023].

[9] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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 ('GDPR')

[10] All European Academies, The European Code of Conduct for Research Integrity, Berlin 2017. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/european-code-of-conduct-for-research-integrity_horizon_en.pdf [Accessed: 28-02-2023].

Appendix A. Dataset description template

This appendix shows the template defined in AI4CYBER project to describe the datasets that will be used and produced in the project. The template intends to facilitate both the partners and the readers of the DMP to identify the characteristics of each dataset and its purpose in the context of the project work.

Therefore, the following table provides the template for the description of the partner’s datasets in both D1.2 and D1.4. The partners shall use this template for describing both the research datasets that they are using or will use in research activities of project, and the research datasets that they will generate in their research.

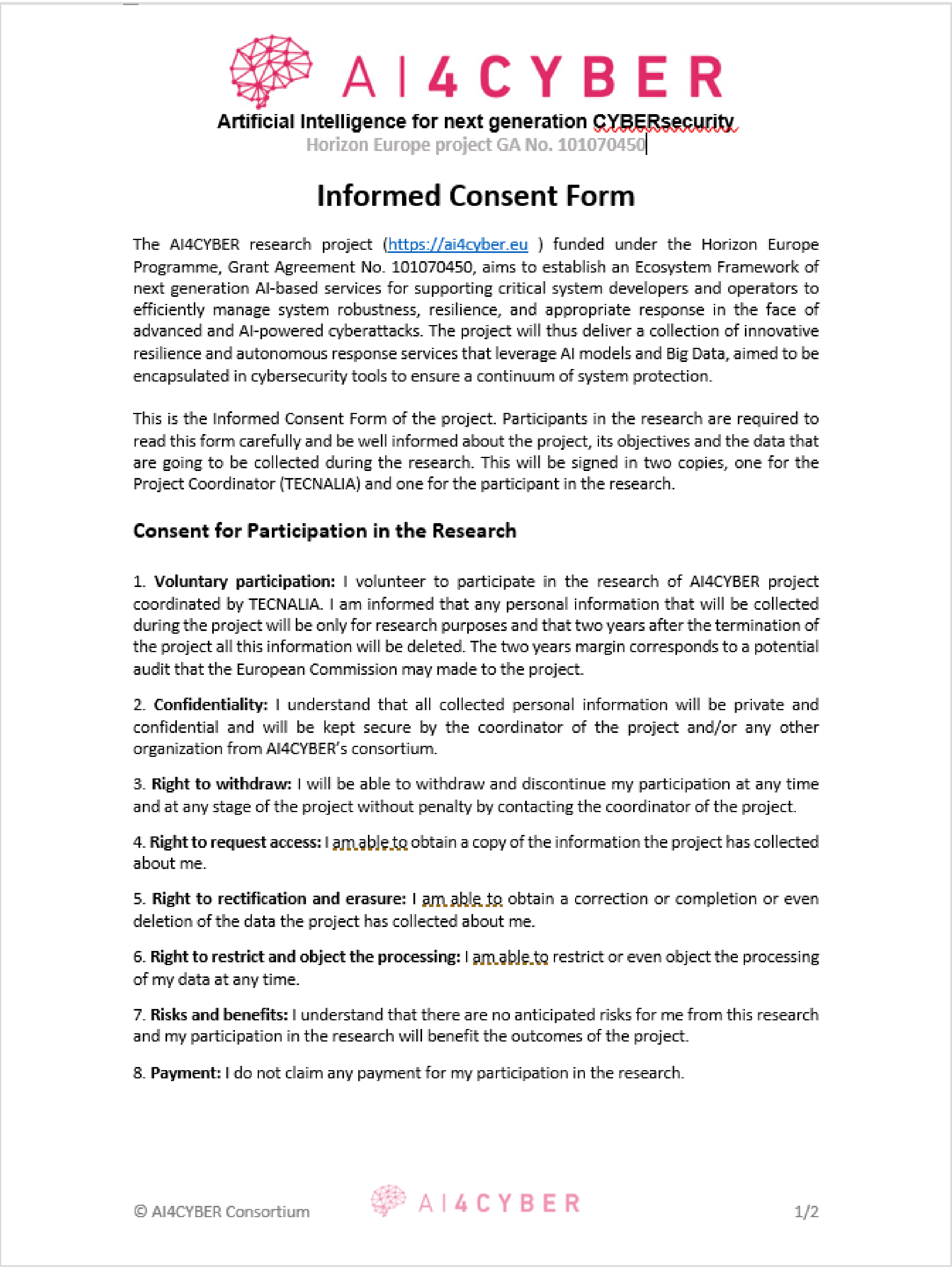
Please note that the template extends and adapts to AI4CYBER needs the template of the H2020 ELECTRON project (Grant agreement ID: 101021936, https://cordis.europa.eu/project/id/101021936).

Table A.1: AI4CYBER Dataset Template

Dataset Name	Name of the Dataset
Dataset Summary	
Dataset Description	Description of the Dataset.
Dataset Purpose	Research purpose of the Dataset.
Dataset Type/Format	Type/format of the Dataset (e.g. CSV file, MySQL db, pcap file, etc.).
Re-use of existing data	Whether the Dataset is expected to re-use previously collected data. Yes/No
Dataset Origin	Context and method used in the dataset generation (e.g. the name of the EU-funded project where it was created).
Dataset Collection	How the Dataset was collected.
Dataset Size	Actual or Estimated size of the Dataset.
AI4CYBER Work Packages, Tasks and Deliverables	
Relevant Work Packages	WP(s) where the Dataset will be utilized.
Relevant Tasks	Task(s) where the Dataset will be utilized.
Relevant Deliverables	Deliverable(s) where the Dataset will be utilized.
Relevant Use Cases	Use Case(s) where the Dataset will be utilized/generated (if any). Note that there are three use cases in AI4CYBER: • UC1: Detection and Mitigation of AI-powered Attacks against the Energy Sector • UC2: Robustness and autonomous adaptation of Banking applications to face AI-powered attacks • UC3: Resilient hospital services against advanced and AI-powered cyber-physical attacks
Partners Services and Responsibilities	
Partner Owner	Partner owner of the Dataset.
Partners responsible for data collecting	Partners who will be responsible for collecting the Dataset.
Partners responsible for data analysis	Partners who will be responsible for analysing the Dataset.
Partners responsible for data storage	Partners who will be responsible for storing the Dataset.
FAIR Principles	
FAIR Data - findability	
Discoverable and Identifiable Data	Whether the data within the Dataset are expected to be discoverable and identifiable. Yes (indicate means) / No
Naming Convention	The naming convention adopted for the Dataset.
Versioning	Whether the Dataset will follow a versioning policy and which. Yes (indicate policy) / No

Search Keywords	Whether the Dataset is expected to support Search Keywords. Yes (indicate keywords) / No
Metadata	Whether the Dataset is expected to keep metadata information. Yes (indicate types of metadata) / No
FAIR Data – accessibility	
Dataset Openly Accessible	Whether the Dataset is expected to be openly accessible. Yes / No.
Methods/Tools for accessing the data	Methods for accessing the data (if any).
Access Restrictions	Access restrictions (if any).
Repository	The repository or data pool where the Dataset will be stored (if any).
FAIR Data - interoperability	
Interoperability	Please describe if the Dataset will be interoperable.
Standards	Please provide any interoperability standards.
FAIR Data – reusability	
Licence	The license which protects the Dataset (if any).
Data sharing terms	Data sharing terms related to the Dataset (if any).
Dataset sharing after the end of the project	Data sharing terms related to the Dataset, upon the conclusion of the project (if any).
Preservation and update of the dataset after the project	The long-term support plan for the Dataset, upon the conclusion of the project (if any).
Re-use timeframe	Description of the timeframe for re-use (if any).
Allocation of Resources	
Cost for making the Dataset FAIR	The cost of ensuring the Dataset is FAIR.
Costs for the preservation (including backup) and the update of the Dataset	The cost of preservation and update of the Dataset.
Data Security and privacy	
Security Measures	The adopted security measures for the Dataset.
Privacy Measures	The adopted privacy measures for the Dataset.
Ethics	
Ethical and Legal Aspects	The Ethical and Legal Aspects of the Dataset
Other Issues	Please describe any other issues.

Appendix B. Informed consent form



AI4CYBER

Artificial Intelligence for next generation CYBERsecurity

Horizon Europe project GA No. 101070450

Informed Consent Form

The AI4CYBER research project (https://ai4cyber.eu) funded under the Horizon Europe Programme, Grant Agreement No. 101070450, aims to establish an Ecosystem Framework of next generation AI-based services for supporting critical system developers and operators to efficiently manage system robustness, resilience, and appropriate response in the face of advanced and AI-powered cyberattacks. The project will thus deliver a collection of innovative resilience and autonomous response services that leverage AI models and Big Data, aimed to be encapsulated in cybersecurity tools to ensure a continuum of system protection.

This is the Informed Consent Form of the project. Participants in the research are required to read this form carefully and be well informed about the project, its objectives and the data that are going to be collected during the research. This will be signed in two copies, one for the Project Coordinator (TECNALIA) and one for the participant in the research.

Consent for Participation in the Research

1. **Voluntary participation:** I volunteer to participate in the research of AI4CYBER project coordinated by TECNALIA. I am informed that any personal information that will be collected during the project will be only for research purposes and that two years after the termination of the project all this information will be deleted. The two years margin corresponds to a potential audit that the European Commission may made to the project.

2. **Confidentiality:** I understand that all collected personal information will be private and confidential and will be kept secure by the coordinator of the project and/or any other organization from AI4CYBER's consortium.

3. **Right to withdraw:** I will be able to withdraw and discontinue my participation at any time and at any stage of the project without penalty by contacting the coordinator of the project.

4. **Right to request access:** I am able to obtain a copy of the information the project has collected about me.

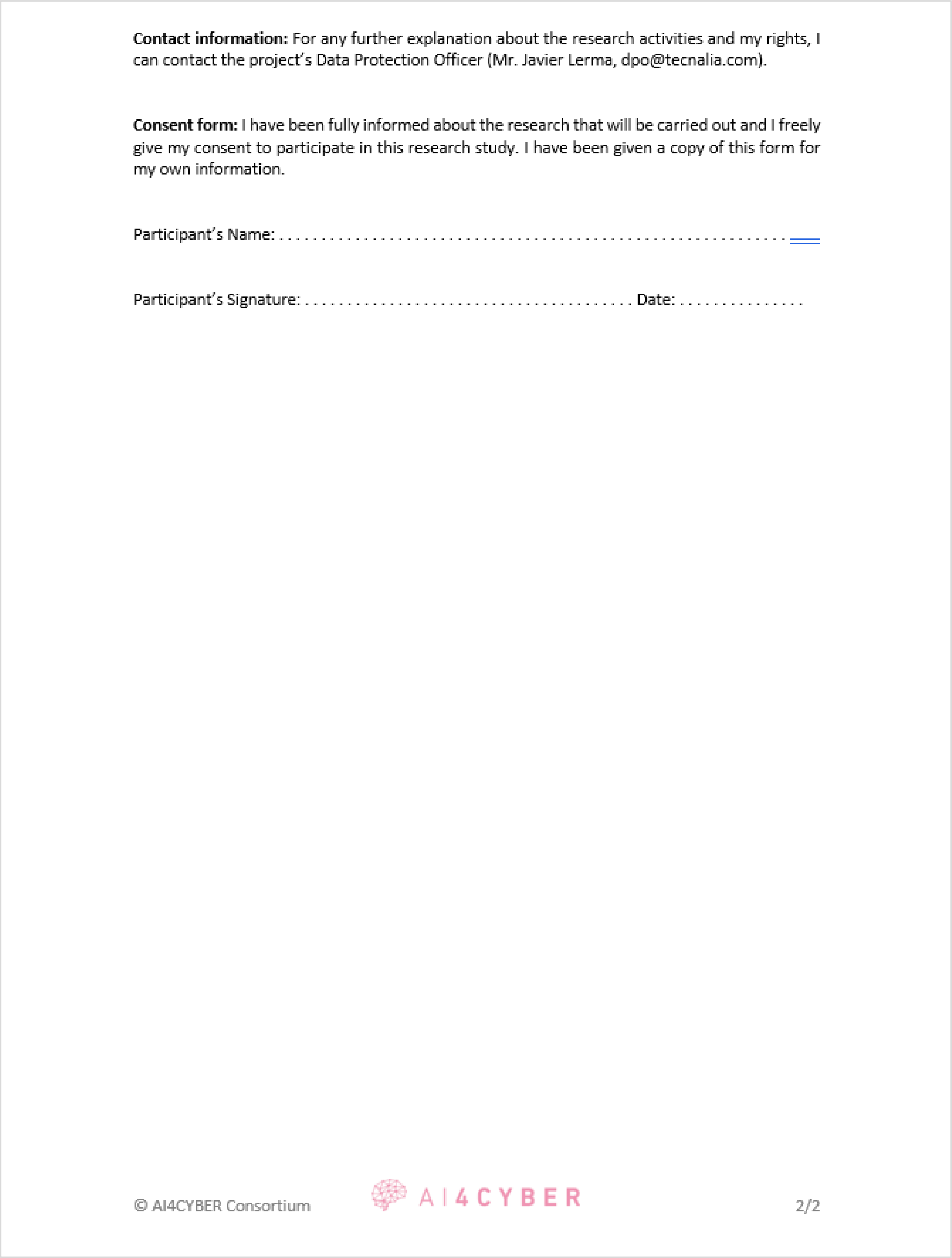
5. **Right to rectification and erasure:** I am able to obtain a correction or completion or even deletion of the data the project has collected about me.

6. **Right to restrict and object the processing:** I am able to restrict or even object the processing of my data at any time.

7. **Risks and benefits:** I understand that there are no anticipated risks for me from this research and my participation in the research will benefit the outcomes of the project.

8. **Payment:** I do not claim any payment for my participation in the research.

© AI4CYBER Consortium AI4CYBER 1/2



Contact information: For any further explanation about the research activities and my rights, I can contact the project's Data Protection Officer (Mr. Javier Lerma, dpo@tecnalia.com).

Consent form: I have been fully informed about the research that will be carried out and I freely give my consent to participate in this research study. I have been given a copy of this form for my own information.

Participant's Name: . . . .

Participant's Signature: . . . . Date: . . . .

© AI4CYBER Consortium 2/2