



D7.6 DATA MANAGEMENT PLAN

Authors Neofytos Gerosavva (8BELLS), Ioannis Giannoulakis (8BELLS), Michalis Tzifas (8BELLS), Christos Anagnostopoulos(ISI), Apostolos Fournaris(ISI), Aris S.Lalos (ISI), Petros Kapsalas (PASEU), Gianmarco Genchi (CRF), Beatriz Gallego-Nicasio (ATOS), Ruben Trapero (ATOS), Nikolaos Voros(UoP), Georgios Keramidas(UoP), Christos Antonopoulos(UoP), Christos Panagiotou (UoP)

Work Package WP7 – Industry Driven Trial and Evaluation

Abstract

This report constitutes an output of the Task 7.1 “Dissemination planning and activities” and represents the D7.6 that is the Open Data Management Plan (DMP). This document is a public report that describes how the various CPSoSaware datasets will be collected, under what circumstances, and how these datasets will be monitored and stored within the project framework and after the completion of the project. Therefore, the overall aim of this document is to support the data management life cycle for all project datasets that will be collected, processed or generated by the project partners.



Deliverable Information

<i>Work Package</i>	WP7
<i>Task</i>	T7.1 Dissemination planning and activities
<i>Deliverable title</i>	Data Management plan
<i>Type</i>	ORDP: Open Research Data Pilot
<i>Dissemination Level</i>	PU
<i>Status</i>	D: Draft
<i>Version Number</i>	1.0
<i>Due date</i>	M6

Project Information

<i>Project start and duration</i>	01/01/2020 – 31/12/2023, 36 months
<i>Project Coordinator</i>	Industrial Systems Institute, ATHENA Research and Innovation Center 26504, Rio-Patras, Greece
<i>Partners</i>	<ol style="list-style-type: none"> 1. ATHINA-EREVNIKIKO KENTRO KAINOTOMIAS STIS TECHNOLOGIES TIS PLIROFORIAS, TON EPIKOINONION KAI TIS GNOSIS (ISI) the Coordinator 2. FUNDACIO PRIVADA I2CAT, INTERNET I INNOVACIO DIGITAL A CATALUNYA (I2CAT), 3. IBM ISRAEL - SCIENCE AND TECHNOLOGY LTD (IBM ISRAEL) 4. ATOS SPAIN SA (ATOS), 5. PANASONIC AUTOMOTIVE SYSTEMS EUROPE GMBH (PASEU) 6. EIGHT BELLS LTD (8BELLS) 7. UNIVERSITA DELLA SVIZZERA ITALIANA (USI), 8. TAMPEREEN KORKEAKOULUSAATIO SR (TAU) 9. UNIVERSITY OF PELOPONNESE (UoP) 10. CATALINK LIMITED (CATALINK) 11. ROBOTEC.AI SPOLKA Z OGRANICZONA ODPOWIEDZIALNOSCIA (RTC) 12. CENTRO RICERCHE FIAT SCPA (CRF) 13. PANEPISTIMIO PATRON (UPAT)
<i>Website</i>	www.cpsosaware.eu

Control Sheet

VERSION	DATE	SUMMARY OF CHANGES	AUTHOR
0.1	15/3/2020	Initial ToC defined and circulated to the consortium for approval	Neofytos Gerosavva
0.2	25/04/2020	Initial Draft circulated to the Consortium	Neofytos Gerosavva, Ioannis Giannoulakis, Michalis Tzifas (8BELLS)
0.3	15/05/2020	ATOS, ISI, PASEU, UoP, CRF	Christos Anagnostopoulos, Apostolos Fournaris, Aris S. Lalos (ISI), Petros Kapsalas (PASEU), Gianmarco Genchi (CRF), Beatriz Gallego- Nicasio, Ruben Trapero (ATOS), Nikolaos Voros, Georgios Keramidas, Christos Antonopoulos, Christos Panagiotou (UoP)
0.4	25/05/2020	EIGHT BELLS drafted next version incorporating partners input	Neofytos Gerosavva, Ioannis Giannoulakis, Michalis Tzifas (8BELLS)
0.5	01/06/2020	Pre-final version released	Neofytos Gerosavva, (8BELLS)
0.6	22/06/2020	Review by I2CAT, UPAT	Javier Fernandez Hidalgo (I2CAT) Gerasimos Arvanitis (UPAT)

Data Management Plan

0.7	23/06/2020	Review comments addressed; new version produced	Neofytos Gerosavva, (8BELLS)
1.0	30/06/2020	Final version - Document ready for submission to the European Commission	Apostolos Fournaris, Aris Lalos (ISI)

	NAME
Prepared by	8BELLS
Reviewed by	I2CAT, UPAT
Authorised by	8BELLS

DATE	RECIPIENT
	Project Consortium
	European Commission

Table of contents

Executive Summary.....9

1 Introduction.....10

 1.1 Purpose and scope.....10

 1.2 Open Research data pilot10

 1.3 Structure of Document11

 1.4 Approach11

2 Personal data protection: EU regulations15

 2.1 Legal framework.....15

 2.2 CPSoSaware Ethical Procedures.....19

3 Open access requirements21

 3.1 Open access in H2020.....21

 3.2 Open Access to Scientific Publications.....21

4 CPSoSaware data management plan.....23

 4.1 CPSoSaware datasets25

 4.2 Data stakeholders in CPSoSaware40

 4.3 Process to provide the Open Data43

 4.4 Process of data storage, preservation and security44

5 Conclusion.....46

6 References47

7 ANNEXES49

ANNEX I: DATA MANAGEMENT QUESTIONNAIRE.....50

ANNEX II: INFORMATION SHEET58

ANNEX III: XL-SIEM JSON Data Format.....59

List of Tables

Table 1: FAIR Data Management at a glance: issues to cover in your Horizon 2020 DMP [7]	14
Table 2 : Intra-communication network performance data.....	25
Table 3: simulating and prototyping hardware designs	26
Table 4: XL-SIEM dataset.....	28
Table 5: EnvModel4AD	29
Table 6 : Offline Variants of the Data Records	30
Table 7: LISATraffic Sign.....	31
Table 8: Mapillary Global.....	32
Table 9: The German Traffic Sign Recognition Benchmark (GTSRB)	33
Table 10: The German Traffic Sign Detection Benchmark (GTSDB) (ISI).....	34
Table 11: Lyft data	34
Table 12: Kitti data	35
Table 13: Motion Distorted Lidar.....	36
Table 14: Operator Heart Rate.....	37
Table 15: Layout of the manufacturing cell.....	37
Table 16: Strength of the operator	38
Table 17: Number of windshields	39
Table 18: Safety Zone violation	39
Table 19: Datasets to be generated during project lifetime	43

List of Figures

Figure 1 : Open Research Data in Horizon 2020- How it works [1][2]	14
Figure 2: Critical Information Data lifecycle	24
Figure 3: CPSoSaware Datasets per category.....	40

Executive Summary

As is being defined in the CPSoSWARE DoA, the research data gathered in CPSoSaware will be described in the current document that is the project Data Management Plan, following the guidelines set by the European Commission Bodies regarding Data Management in H2020. The Data Management Plan intends to be a living document in which information can be progressively elaborated through the project lifecycle and the current version will be updated accordingly. Additionally, the current document will furthermore provide an insight to ethical and data management guidelines, best practices for collecting data information, specific standards, laws applying, information on Open research data pilot approach etc. Moreover, CPSoSaware as a participant in the Open Research Data Pilot of the H2020 Programme, has to follow the recommendations by the European Commission and the FAIR Data Management Plan in order to make the research data Findable, Accessible, Interoperable and Re-usable. For this reason, a questionnaire was prepared in order to be answered by the consortium partners for collecting their specific datasets and understanding the type of generated data. Nonetheless, at this point it is not possible to collect all the available information regarding how to make the provided datasets FAIR as it is still very early in the project. Thus, at this document we tried to describe the various datasets collected as briefly as possible providing at the same time all the information available for the time being, along with partners' plans on making them open and available. **After the completion of the aforementioned questionnaire, the project has identified a list of 16 datasets.** The whole process of collecting information about partners' datasets will be described in this deliverable and the respective questionnaires as well as other source of related information can be found in the respective annexes.

1 Introduction

1.1 Purpose and scope

Generally, Data Management plans are created in order to ensure that a proper strategy is being followed for the preservation of the data management life cycle principles within an H2020 project.

The CPSoSaware project as member of the Horizon 2020 Open Research Data Pilot [1], is required to define a data management plan that will describe the collected datasets expected to be generated all through the project duration, along with the related stakeholders and the specific characteristics that are attached to these datasets. Moreover it will provide information on what a Data Management Plan contains, how the consortium partners provided the datasets information, how we intent to maintain the data, make it openly free and available etc. Annexes contain the questionnaire template used for collecting the datasets and some additional information. Furthermore, the CPSoSaware Data Management Plan aims to provide the main elements of data management policy and methodology within the project, and additionally to be the point of reference and guide when accessing and/or managing data generated from project research activities. The Open Research Data Pilot (ORD) [2] [3], enables and maximizes open access, as well as reuse of the research data generated by EU-funded projects. According to OpenAIRE, benefits of taking an active approach to research data management include ease of access, efficiency, speed as well as improved quality and transparency of research.

There are two main pillars to the ORD Pilot:

1. **Developing a Data Management Plan (DMP)**
2. **Providing open access to research data, if possible.**

The FAIR [4] principles have been generated with the purpose of improving the best practices towards data monitoring in research projects. We intend to follow the guidelines principles provided for the **Open Research Data Pilot (ORD)** and the main goal is to make the data FAIR that is findable, accessible, interoperable and re-usable. Thus this report will also define the way for data handling, storing, archiving and sharing according to the “Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020” and in the “Guidelines on FAIR Data Management in Horizon 2020” [5][6].

1.2 Open Research data pilot

Due to the reason that CPSoSaware participates in the Pilot on Open Research Data in Horizon 2020, it will offer open access to scientific results reported in publications, to related scientific data and to data generated throughout the project duration. Open data as being defined in the Open Research pilot website is the data that is free to use, reusable, and redistributable. The basic aim of the Open Research Data Pilot is to make the research data generated by selected Horizon 2020 projects open. As mentioned previously, any project that chooses to be part of the pilot must develop a data management plan, store the generated data in a research data repository, ensure that third parties can freely access, mine, exploit, reproduce and disseminate it and furthermore define clearly the tools that will be needed to use the raw data for the validation of the research results (or provide the tools themselves). The Data Management Plan will be updated according to the project needs in order to be able to reflect important changes to the project.

Such changes can be for instance the generation and inclusion of new datasets, new consortium policies or other various external factors. For this reason, the status of Data Management plan will be monitored during project lifetime by ATHENA RESEARCH CENTER - Industrial Systems Institute (ISI) as the project coordinators and from EIGHT BELLS as the dissemination leaders.

1.3 Structure of Document

This document has been structured as follows:

- Section 2 refers to the **EU regulations and legal framework**
- Section 3 defines the **open access requirements**
- Section 4 is the most important section of the document as it defines the **CPSoSaware Data Management Plan and describes the datasets produced by the project**. Other issues described in this section refer to Data stakeholders in CPSoSaware, legal requirements and procedures prior of data collection, process of collecting, monitoring and handling the data –methodology to be applied, data management life cycle for all datasets, process to provide the Open Data, and way of preserving and curating the data
- Section 5 summarizes and concludes the deliverable
- ANNEX I contains the questionnaire that has been used for the collection of information from the CPSoSaware partners

1.4 Approach

Based on the Guidelines of FAIR Data Management in Horizon 2020 a set of questions was prepared (Annex I) and sent to the consortium partners in order to guide them on how to collect the necessary information regarding the datasets that will be generated throughout the project duration. Then each partner made an internal assessment and provided feedback with a level of detail appropriate to the project, indicating whether there will be data to which open access can be granted without adversely affecting legitimate interests, including IPRs. The feedback provided to the document editor (EIGHT BELLS) will be described in the respective section. Of course, given the very early project stage at the time being of writing it was not required to provide detailed answers to all the questions. According to the summary table 1 below, issues that should be addressed include data summary, purpose of the data collection, types and formats of data, FAIR data principles (making data findable, accessible, and interoperable), data storage and security, ethical aspects etc. Moreover, in the Figure 1 below we present the steps on how the Open Research Data pilot works within the H2020 projects.

DMP component	Issues to be addressed
	<p>State the purpose of the data collection/generation</p> <ul style="list-style-type: none"> • Explain the relation to the objectives of the project • Specify the types and formats of data generated/collected • Specify if existing data is being re-used (if any)

	<ul style="list-style-type: none">• Specify the origin of the data• State the expected size of the data (if known)• Outline the data utility: to whom will it be useful
2. FAIR Data 2.1. Making data findable, including provisions for metadata	Outline the discoverability of data (metadata provision) <ul style="list-style-type: none">• Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?• Outline naming conventions used• Outline the approach towards search keyword• Outline the approach for clear versioning• Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how
2.2 Making data openly accessible	Specify which data will be made openly available. If some data is kept closed provide rationale for doing so <ul style="list-style-type: none">• Specify how the data will be made available• Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?• Specify where the data and associated metadata, documentation and code are deposited• Specify how access will be provided in case there are any restrictions

<p>2.3. Making data interoperable</p>	<ul style="list-style-type: none"> • Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability. • Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies
<p>2.4. Increase data re-use (through clarifying licences)</p>	<ul style="list-style-type: none"> • Specify how the data will be licenced to permit the widest reuse possible the maximum reusability • Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed • Specify whether the data produced and/or used in the project is reusable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why • Describe data quality assurance processes • Specify the length of time for which the data will remain re-usable
<p>3. Allocation of resources</p>	<ul style="list-style-type: none"> • Estimate the costs for making your data FAIR. Describe how you intend to cover these costs • Clearly identify responsibilities for data management in your project • Describe costs and potential value of long term preservation
<p>4. Data security</p>	<ul style="list-style-type: none"> • Address data recovery as well as secure storage and transfer of sensitive data
<p>5. Ethical aspects</p>	<ul style="list-style-type: none"> • To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and

	related technical aspects if not covered by the former
6. Other	<ul style="list-style-type: none"> Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Table 1: FAIR Data Management at a glance: issues to cover in your Horizon 2020 DMP [7]



Figure 1 : Open Research Data in Horizon 2020- How it works [1][2]

2 Personal data protection: EU regulations

2.1 Legal framework

The research data generated or created within a research project may include publicly available datasets, statistics, reporting material such as deliverables, scientific publications, presentations, dissemination material, experimentation data, various kind of measurements, simulation results, technical datasets such as source code, libraries, surveys & interviews results etc.

Some of the expected CPSoSaware datasets are:

- Data coming from pilot sites, towards trials implementation e.g., questionnaires from focus groups, interviews, online questionnaires etc
- Software components and source code
- Publicly available datasets
- Simulation data
- Data coming from sensors

CPSoSaware consortium's priority is to stay within the framework structured by the joined provision of:

- a) The European Regulation 2016 regarding "Protection of natural persons with regard to the processing of personal data and on the free movement of such data" [8]
- b) Horizon 2020 Ethics guidelines [9]

The project consortium will furthermore follow professional ethical practice and comply with the Charter of Fundamental Rights of the European Union.

According to Guidelines on FAIR Data Management in Horizon 2020, the data generated during and after all projects should follow the FAIR data principles that require that data are Findable, Accessible, Interoperable and Reusable. According with the same document these requirements precede the implementation stage of a project and don't necessarily suggest any specific technology, standard, or implementation solution. **The CPSoSaware consortium as mentioned previously, will take all possible measures for addressing the principles, and criteria derived from the FAIR Guidelines, and starting from this document, we will describe all necessary steps for being successful on this issue.**

In order to facilitate the application of FAIR principles, the EC suggests various standards that can be taken into account towards the adoption of the FAIR principles.

The CPSoSaware Description of Work within the section 5 "ETHICS AND SECURITY" refers to relevant laws and directives as well as ethical issues and guidelines describing the rights any of the actors have on the IPRs involved in the project. Some of them are listed below.

Relevant Laws and Directives (mentioned on DoW, section 5.1, page 128)

- General Data Protection Regulation (GDPR) (EU) 2016/679, of the European Parliament and of the

<p>Council of 27 April 2016: on protection of natural persons with regard to the processing of personal data and on the free movement of such data</p>
<p>- Directive 2002/58/EC: Processing of personal data and the protection of privacy in the electronic communications sector</p>
<p>- Charter of Fundamental Rights of the European Union</p> <p>The Charter sets the starting point for any research or action conducted within the context of the European Union and contains 54 articles divided in seven titles: dignity, freedoms, equality, solidarity, citizens' rights, justice and general provisions governing the interpretation and application of the Charter. This Charter must be taken into consideration by Member States when applying European Union law.</p> <p>Any action taken within CPSoSaware project needs to be compliant with all fundamental rights preserved in this Charter.</p>
<p>European Code of Conduct for Research Integrity</p> <p>This document serves the European research community as a framework for self-regulation across all scientific and scholarly disciplines and for all research settings. The European Commission recognises the Code as the reference document for research integrity for all EU-funded research projects and as a model for organisations and researchers across Europe. Fundamental principles of research integrity include reliability honesty respect and accountability</p>
<p>- Directive 83/570/EEC, Regulation or administrative action relating to proprietary medicinal products - EGE (European Group on Ethics): Opinion N° 21 (2007) on the ethical aspects of nanomedicine</p>
<p>- Public participation: transparency</p>
<p>- Directive 99/5/EC: Radio equipment and telecommunications terminal equipment and the mutual recognition of their conformity</p>
<p>- Directive 97/66/EC on Data Protection in the Telecommunications Sector</p>
<p>- Art. 29 - Data Protection Working Party: Working Document on Privacy on the Internet</p>

- WMA: Declaration of Helsinki (2004);
- UNESCO: Universal Declaration on Bioethics and Human Rights (2005);
- CoE: (1) Convention on Human Rights and Biomedicine (Convention of Oviedo), Articles 15-18 ETS n°164 (1997); and (2) Additional Protocol on Biomedical Research, CETS n°195 (2005).
- Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (1985);

German Legislations and Guidance: (this is relevant with the trials within the WP6 framework that will take place in Germany from PASEU)

- Act to Adapt Data Protection Law to Regulation (EU) 2016/679 and to Implement Directive (EU) 2016/680 (July 5, 2017).
- The National Council for Ethics: http://www.ethikrat.org/ ("Deutscher Ethikrat"), as established by legislation of the "Deutscher Bundestag" on April 26, 2007 may give recommendations which are not binding.
- The Central Ethics Committee of the German Medical Association ("Zentrale Ethikkommission zur Wahrung ethischer Grundsätze in der Medizin und ihren Grenzgebieten bei der Bundesärztekammer [ZEKO]") which gives opinions on general ethical issues and which may give advice to the Ethics Committees of the Medical Associations at their request. This advice is not binding on these committees.
- X-ray regulation (Röntgenverordnung – RöV) (especially §28g)
- Regulation on the Protection from Ionising Radiance (Strahlenschutzverordnung – StrSchV) (especially §§23, 24, 92)

- The Act on Medical Devices (Medical Devices Act) (Medizinproduktegesetz – MPG) (especially §§19-24)
- Medicinal Products Act / Pharmaceutical Products Act (Drug Law) (Arzneimittelgesetz – AMG) (especially §§40-42, §77)

Italian Legislations and Guidance: (This is relevant with the trials within the WP6 framework that will take place in Italy from CRF)

- DECRETO LEGISLATIVO 10 agosto 2018, n. 101: Disposizioni per l'adeguamento della normativa nazionale alle disposizioni del regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati). (18G00129) (GU Serie Generale n.205 del 04-09-2018)
- Implementation of the European Union guidelines on good clinical practice for trials on medicinal products (Health Ministry, Decree of 15 July 1997).
- Legislative Decree 24 June 2003, No. 211 transposing Directive 2001/20/EC concerning the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use. An unofficial English translation of the above mentioned Decree is available at: http://www.agenziafarmaco.gov.it/sites/default/files/decreto_24062003_inglese.pdf
- Decree of the Ministry of Health, December 21, 2007, Complete guide on the modalities to request the authorization of a clinical trial on a medicinal product for human use, the disclosure of substantial amendments and the final declaration to the competent authority.
- The Ministerial Decree 8 February 2013 stating the reorganization of Ethics Committees in Italy, member characteristics and background: so called DecretoBalduzzi.

Specific ethical issues addressed by the project (described in section 5.1.2 of the DoW –page 129)

Ethical need for technology solution to security and privacy

The Project Consortium will follow the relevant legal Acts that are in place, in each partner’s country and any applicable EU legislation regarding data protection and during the project length it will stipulate any conclusive needs within the consortium. This may refer to the temporality for storage of data, security of data transfer, relevant consent applications and relevant advertisement of the use of the data. **In addition**

to the checks made above with trial subjects, the technology should take into account the fact that each participant in the research will be given a unique identification code, rather than a name, and all data will be securely stored and preserved, both electronically and on paper. No unauthorised access can be allowed.

Informed consent

Informed Consent is the decision, which must be written, dated and signed, to take part in a research campaign, taken freely after being duly informed of its nature, significance, implications and risks and appropriately documented, by any person capable of giving consent or, where the person is not capable of giving consent, by his or her legal representative; if the person concerned is unable to write, oral consent in the presence of at least one witness may be given in exceptional cases, as provided for in national legislation. The European Commission - Research Directorate-General provides guidance on informed consent and this will be respected.

An informed consent will be used towards the execution trials within the WP6 “Industry Driven Trial and Evaluation” framework as well as towards the implementation of other tasks such as for instance T2.1 Analysis of user skills/factors, virtual cognitive user/environment models and metrics modelling

The research trial informed consent document will provide a summary of the trial, including:

- **its purpose,**
- **the treatment procedures and schedule**
- **potential risks and benefits**
- **participants’ rights**

Consent documentation will follow the Informed Consent Form Template for Clinical Studies provided by the World Health Organisation Research Ethics Review Committee. The research that is going to be undertaken requires persons able to freely understand and question the consent procedure.

2.2 CPSoSaware Ethical Procedures

The design and management of cyber-physical systems in industrial applications require serious consideration for critical ethical aspects such as:

- The layers of access to technology
- The difficulty of respecting privacy and confidentiality when third parties may have a strong interest in getting access to electronically recorded and stored personal data
- Informed consent
- Incidental finding the difficulty in ensuring the security of **shared personal health data**
- **Physiological, behaviour and cognitive load data**
- **Transparency of the collected data.**

The proper management of these issues will be carefully investigated by the CPSoSaware Ethics Board, while their implementation in CPSoSaware, will be described on the current document

As mentioned in the DoA the CPSoSaware consortium commits to the following:

- i) the principles of the European Convention of Human Rights,
- ii) the rules of the Convention of the Council of Europe for the protection of individuals with regard to automatic processing of personal data and especially the General Data Protection Regulation (GDPR) (EU) 2016/679, of the European Parliament and of the Council of 27 April 2016, on protection of natural persons with regard to the processing of personal data and on the free movement of such data, will be strictly followed when addressing the ethical questions of CPSoSaware.

The CPSoSaware system is going to be tested in two different pilot sites (Germany and Italy) by performing trail scenarios to two different use cases.

- **The first pilot use case will be undertaken by Panasonic (PASEU).** The pilot will be focused on connected semiautonomous vehicles where we will perform trials focused on Human in the loop scenarios, like non predictable failures that may involve the human driver and how this affects the design operation and continuous support of the CPSoSaware solution as well as human situational awareness enhancement when using the CPSoSaware architecture. We will also use this use-case to access the cybersecurity mitigation strategies using the CPSoSaware architecture and its response to cyberattacks.

-**The second use case will be undertaken by CRF.** The pilot will be focused on Human-Robot Collaboration (HRC) in the manufacturing environment and will involve trials that challenge the MOOD CPSoSaware concept and trials on accidents/failures as well as cybersecurity attacks that challenge the collaborative control mechanism and the autonomic decentralized operation of the CPSoSaware solution as well as the Design operation continuum support in the presence of cybersecurity attacks.

Project trials will not intensify the daily living conditions of participants, will not cause any other negative side-effects or prolong the duration of any tasks.

Data protection issues

All the researchers to be involved in **CPSoSaware** will comply with and follow the principles outlined by the General Data Protection Regulation (GDPR) (EU) 2016/679, of the European Parliament and of the Council of 27 April 2016, on protection of natural persons with regard to the processing of personal data and on the free movement of such data.

3 Open access requirements

3.1 Open access in H2020

The “H2020 Programme Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 Open access (OA) refers to the practice of providing online access to scientific information that is free of charge to the end-user and reusable.

“Scientific” refers to all academic disciplines. In the context of research and innovation, “scientific information” can mean:

- a) Peer-reviewed scientific research articles (published in scholarly journals), or
- b) Research data (data underlying publications, curated data and/or raw data).

CPSoSaware will offer open access to scientific results reported in publications, to relevant scientific data and to data generated throughout the project lifecycle, under the provision that they are anonymized and fully respecting national and EU privacy regulations. As it is well known for projects that are participating in the Pilot on Open Research Data in Horizon 2020, the EC suggests the [FAIR Data Management Plan Template](#). As described in the Guidelines of FAIR Data Management in Horizon 2020, this template is a set of questions that the partners should answer with a level of detail appropriate to the project.

The OpenAIRE portal & Zenodo

The OpenAIRE portal [10] is based on the European OA repository infrastructure and provides tools for publishing (depositing) EC funded publications. The goal is to make as much European funded research output as possible available, to promote open scholarship and substantially improve the discoverability and reusability of research publications and data. In order to create workflows which will enable an interoperable network of repositories and easy upload into an open all-purpose repository called Zenodo. Zenodo [11] is used to host the publications/datasets/etc and allows for directly linking the publications to the project’s profile on OpenAIRE.

3.2 Open Access to Scientific Publications

Open access (OA) refers to the practice of providing, free of charge to any user, online access to all peer - reviewed scientific information and all the research data. The CPSoSaware Consortium strongly believes in the necessity of transparency of the scientific process, in particular for science driven by public funds. **The Open Access Model aims to ensure free, without barriers access to scientific literature for readers.**

As stated in the Participant Portal H2020 Online Manual in the Open access & Data management section [12], beneficiaries can freely choose between two options towards open access for their publications:

A. Self-archiving also referred to 'green' open access means that the author, or a representative, makes accessible the published article or the final peer-reviewed manuscript in an online repository before, at the same time as, or after publication.

B. Open access publishing also referred as 'gold' open access means that the publisher make their articles immediately available in open access mode and no later than six months on the publisher /journal website (12 months for articles in the fields of social sciences and humanities) after publication.

Access to Presentations and Peer-Reviewed Publications: The consortium recognises the need to ensure that research publications and related research data are made widely and publically available via open access paradigm. In order increase the visibility of project results and promote them to the scientific community, **CPSoSaware consortium intends to follow the open access publication model and use open access repositories connected to the tools suggested by the EU Commission.**

The way to do this will be via:

- i) creating CPSoSaware repositories on the pre-mentioned EC OpenAIRE and CERN's ZENODO and populating them with open access articles, chapters and public deliverables of the project and
- ii) making all project publications available through the official project website of CPSoSaware <http://cpsosaware.eu/>. In order to achieve this, CPSoSaware consortium will pursue publications in high quality, open access journals and conferences

4 CPSoSaware data management plan

This section will provide a summary on how data will be produced, monitored, published and preserved throughout the project, providing the necessary clarifications for any restrictions which apply to any of these steps. During the three-year project lifecycle various datasets from different consortium members, representing different domains, will be generated.

The CPSoSaware data management plan presents within this chapter a sufficiently detailed description of:

- The datasets that will be produced
- The data controllers
- Their origin (if there is a primary source)
- Data specific types
- The tasks that these datasets will be generated within the project
- How the data will be made available & procedures for doing so, related software and tools
- Whether there will be restriction to any groups and why
- Ways of storage, archiving and preservation
- Timeline indication
- Reference to existing standards adopted, and other related information

Of course as already mentioned, according to the way the project will evolve and the related progress in the various work packages, datasets maybe subject to changes or updates in terms of the types, formats, and origins of the data so acting accordingly the Data Management plan is considered a document that will change dynamically addressing these potential changes.

We could describe with six steps the life cycle of Critical Information Data:

- 1) Plan and Create
- 2) Store
- 3) Use
- 4) Share
- 5) Archive
- 6) Delete

The Figure 2 illustrates these steps:

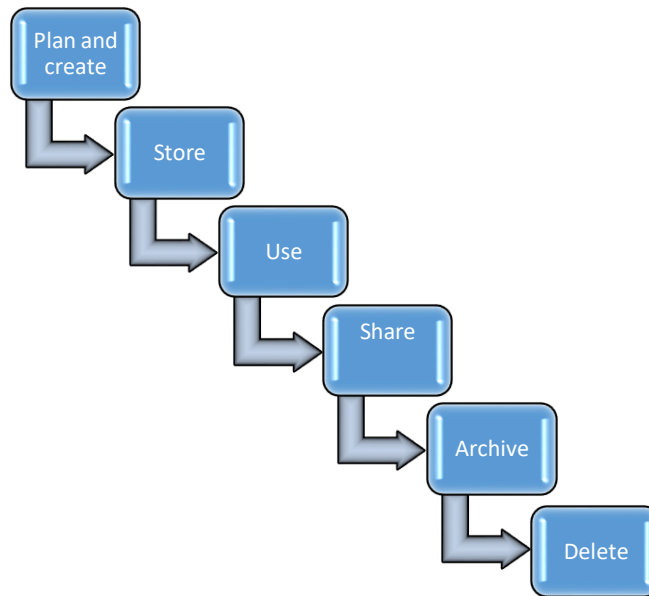


Figure 2: Critical Information Data lifecycle

The processes, which will be implemented in relation to data protection, are divided into the following categories:

- Storage of digital data
- Storage of physical data
- Sharing of data
- Data disposal, deletion and destruction

CPSoSaware Datasets will be described by:

- (i) Dataset Name
- (ii) Time period covered
- (iii) Purpose
- (iv) Format
- (v) Data origin
- (vi) Dataset Ownership
- (vii) Dataset visibility and sharing,
- (viii) Storage

The datasets to be generated are listed in the next section in table format but it is possible, as the project evolves these tables will be modified with additions or removals of datasets.

4.1 CPSoSaware datasets

DATASET 1: Intra-communication network performance data / University of Peloponnese (UoP)

Name of the dataset	The dataset is “Intra-communication network performance” coming from NS3 simulations
Time period covered	From M1 to M28 of the project lifecycle
Purpose	The purpose of the data is mainly to identify the optimum network configuration operation points with respect to specific application scenarios and respective network requirements extracted.
Format	Simulated - The input will be network requirements extracted from the application use cases, the output will be network performance. Indicative metrics will include packet delay, delay jitter, throughput, packet loss, power consumption.
Data origin	Simulated data
Ownership	University of Peloponnese (UoP)
Visibility and Sharing	As a general principle UoP intends to make data freely available, although the ownership will always belong to UoP. The way to do so will be through UoP Cloud infrastructure following specific data structure to be defined later in the project
Description	The data comes from NS3 [13] massive number of simulations and is generated by computational models where model and metadata are equally important to output data - i.e. climate models, economic models, materials models. This data will be generated initially within the work of the tasks T2.2 and mainly T4.2
Storage	Data will be stored in UoP cloud facilities and will be exposed as required by the project. Data are of low security importance, e.g. no personal data or medical data are involved.

Table 2 : Intra-communication network performance data

DATASET 2: Simulating and prototyping hardware designs / University of Peloponnese

Name of the dataset	The dataset is for simulating and prototyping hardware designs (gem5 and vitis)
----------------------------	---

Time period covered	From M1 -M28 of the project lifecycle
Purpose	The purpose of the data is mainly to simulate and prototype new reliable and low power hardware accelerators
Format	Simulated
Data origin	The data comes from gem5 and vitis (FPGA) [14][15] tools
Ownership	University of Peloponnese
Visibility and Sharing	As a general principle University of Peloponnese intends to make data freely available, although the ownership will always belong to the university. The way to do so will be through UoP Cloud infrastructure following specific data structure to be defined later in the project
Description	The data is: simulation statistics from open source architectural simulators. This data will be generated initially within the T3.1 and mainly T4.1 framework
Storage	Data will be stored in UoP cloud facilities and will be exposed as required by the project. Data are of low security importance, e.g. no personal data or medical data are involved.

Table 3: simulating and prototyping hardware designs

DATASET 3: XL-SIEM dataset /ATOS

Name of the dataset	XL-SIEM dataset
Time period covered	The time period covered by the dataset is configurable at the XL-SIEM [16].

<p>Purpose</p>	<p>The purpose of the dataset are:</p> <ul style="list-style-type: none"> - to receive and maintain event information gathered from different security probes with different type of security information depending on the nature of the security probe sending the data - to generate and maintain security alarms derived from the correlation of the received security events.
<p>Format</p>	<p>Both security events and security alarms use the same data format, represented by JSON formatted data, containing information related to the event or alarm as text.</p> <p>(See also Annex IV at the end of the document.)</p>
<p>Data origin</p>	<ul style="list-style-type: none"> • Observational → Security events are data captured in real time directly from external security probes. • Derived compiled → Security alarms are data coming from the correlation of events.
<p>Ownership</p>	<p>Security events are generated by sensors and probes deployed in the CPS infrastructure. The data encapsulated in security events can be of different nature and related to network traffic, application usage, and device observations. Therefore, the data is owned by the users of the CPS infrastructure. The Data controller is the organisation that owns and controls the CPS infrastructure (e.g. Use case partners).The XL-SIEM processes security events generated in the CPS infrastructure. If the XL-SIEM is installed in the same CPS infrastructure, the organisation that owns and controls the CPS is also a data processor. If the XL-SIEM is installed in a different environment and security events are transferred from the CPS to this other environment, then the owner of the XL-SIEM environment is data processor.</p>
<p>Visibility and Sharing</p>	<p>Security events and alarms may contain personal data, e.g. e-mail addresses, IP addresses. It is also possible to implement profiling techniques to identify behaviours of individuals, e.g. usage of the network. On the other hand, it is possible to implement anonymisation / pseudonymisation to de-identify any personal data included in security alarms, before sharing them. Templates for informed consent form and privacy policy can be developed.</p>

Description	Probes will be monitoring the different elements of the CPS at different levels: device, network, application, controller. These probes generate security events that will be used as input by the XL-SIEM. The XL-SIEM collects these security events, processes and correlates them using security directives developed to detect threats and attacks. The XL-SIEM generates security alarms as output. These alarms can be used by a) security operators/system administrators to investigate a potential security incident and respond if the incident is confirmed; b) calculate security indicators.
Storage	<p>Events are received to the XL-SIEM through a secured rsyslog channel using TLSv2 certificates. Transfer exported using RabbitMQ queues which access is secured using also TLSv2.</p> <p>Events and alarms are stored in a database that implements authentication and authorization controls but does not implement any anonymisation controls. For viewing data at rest, data is stored in a MariaDB. Any software capable of connecting to a MySQL based database would be able to read the data. For exported security alarms a RabbitMQ client, configured properly with suitable TLS certificates is required</p>

Table 4: XL-SIEM dataset

DATASET 4 : EnvModel4AD (PASEU)

Name of the dataset	EnvModel4AD (Environmental modelling for Autonomous driving)
Time period covered	From M06 -M36 of the project lifecycle
Purpose	<p>The purpose of the data is mainly to:</p> <ul style="list-style-type: none"> ● Provide dataset, which will be used for training the AI modules consisting the Ecosystem of CPSoSaware. ● Contribute the scenes from which textural and appearance elements will be used for data augmentation in generating simulated visual content. ● Provide the scenes and scenarios that will be used for calculating the Key Performance Indicators for the Autonomous use case.
Format	The data is : Audio-visual + Text + Numerical + Reused

Data origin	<p>Sensor readings:</p> <ol style="list-style-type: none"> 1. Cameras 2. Lidar 3. Differential GPS 4. Ultrasonic Sensors 5. Clock Measurements <p>Wheel tick/ Flex Ray Measurements</p>
Ownership	PASEU
Visibility and Sharing	<p>The Data will be shared on a network location, accessible by the related partners in accordance to the GDPR legislation.</p> <p>Personal Data will not be shared. All the involved parties will provide their written consent for participating in the data capturing.</p>
Description	<p>The CPSoSaware consortium will define the test-use-cases and the parameters related to the data characteristics (e.g., lighting conditions, weathering conditions, vehicle speed, scene structure, etc) and will employ a traceable document structure where all the params above will be considered. Subsequently, PASEU will deploy the equipment contributed to the project in order the data to be recorded. An information sheet will be distributed along with a letter of consent. Moreover, an NDA process will be needed. For protecting the data PASEU will make use of anonymisation techniques and hide distinctive data of scene structural elements (car plates...)</p>
Storage	This data will be deposited in PASEU’s server.

Table 5: EnvModel4AD

ATHENA RESEARCH CENTER- ISI DATASETS

Data Collection/Selection Methodology

ISI will use public available datasets for training that will be based on 3 major pillars:

- (a) synthetic dataset
- (b) real dataset
- (c) augmented data

Throughout the following subsections the criteria for structuring the dataset will be extensively discussed.

Synthetic Dataset for Traffic Signs and environmental scenes

The following Table 6 describes the offline variants of the data records that can be created by customizing the simulation tool.

The headings for each column are described as follows:

- Attack Class: type of attack
- Attack Variation: manipulation methods used to perform attack
- Location: refers to various settings and conditions e.g. location, world, junction, straight etc
- Time of Day: day or night, can also be expanded to include dusk
- Weather: different weather conditions
- Signs/Duration: presents the duration in number of frames and the amount of possible signs used for each dataset

Note: It is expected to have around 10K samples for each data class.

Attack class	Attack Variation	Location	Time of Day	Weather	Signs/ Duration
Normal (no attack)	None	>3 Locations	Day/ Night	Sun/ Cloudy / Rain	20signs for 30 frames
Attacked with Noise	Gaussian Noise, Masking / Obstruction	>3 locations	Day/ Night	Sun / Cloudy / Rain	20signs for 30 frames
Adversarial Attack	ML generated attack	>3 locations	Day/ Night	Sun/ Cloudy / Rain	20signs for 30 frames

Table 6 : Offline Variants of the Data Records

Additionally, the data generation method should support different types of cameras (e.g. wide field of view) and customizable cameras so as to resemble the target position of the camera on the real vehicle.

Below the expected datasets to be used from ISI during the project duration :

DATASET 5: LISATraffic Sign (ISI)

Name of the dataset	LISATraffic Sign[17]
Time period covered	WP3 duration M6-M32
Purpose	These datasets will be used to train the machine learning models and to generate the attacks described in the section above.
Format	Videos and annotated frames.
Data origin	Visual datasets on traffic signs
Ownership	Publicly available datasets for traffic signs.
Visibility and Sharing	Publicly available datasets
Description	<p>The LISA Traffic Sign dataset is a US traffic signs dataset that contains set of videos and annotated frames.</p> <ul style="list-style-type: none"> • 47 US sign types • 7855 annotations on 6610 frames. • Sign sizes from 6x6 to 167x168 pixels. • Images obtained from different cameras. <p>Image sizes vary from 640x480 to 1024x522 pixels.</p> <ul style="list-style-type: none"> • Some images in colour and some in grayscale. • Full version of the dataset includes videos for all annotated signs.
Storage	Redmine repository

Table 7: LISATraffic Sign

DATASET 6: Mapillary Global (ISI)

Name of the dataset	Mapillary Global [18]
----------------------------	-----------------------

Time period covered	WP3 duration M6-M32
Purpose	These datasets will be used to train the machine learning models and to generate the attacks described in the section above.
Format	Street-level image with bounding box annotations for detecting and classifying traffic signs around the world
Data origin	Visual datasets on traffic signs
Ownership	Publicly available datasets for traffic signs
Visibility and Sharing	Publicly available datasets
Description	<p>The dataset contains a diverse street-level image with bounding box annotations for detecting and classifying traffic signs around the world.</p> <ul style="list-style-type: none"> • 100,000 high-resolution images (52,000 fully annotated, 48,000 partially annotated) • Over 300 traffic sign classes with bounding box annotations • Global geographic reach of images and traffic sign classes, covering 6 continents • Variety of weather, season, time of day,
Storage	Redmine repository

Table 8: Mapillary Global

DATASET 7: The German Traffic Sign Recognition Benchmark (ISI)

Name of the dataset	The German Traffic Sign Recognition Benchmark (GTSRB) [19]
Time period covered	WP3 duration M6-M32
Purpose	These datasets will be used to train the machine learning models and to generate the attacks described in the section above.
Format	Images
Data origin	Visual datasets on traffic signs
Ownership	Publicly available datasets for traffic signs

Visibility and Sharing	Publicly available datasets
Description	<p>GTSRB is a multi-class, single-image classification challenge.</p> <ul style="list-style-type: none"> • Single-image, multi-class classification problem • More than 40 classes • More than 50,000 images in total • Large, lifelike database • Reliable ground-truth data due to semi-automatic annotation • Physical traffic sign instances are unique within the dataset (i.e., each real-world traffic sign only occurs once)
Storage	Redmine repository

Table 9: The German Traffic Sign Recognition Benchmark (GTSRB)

DATASET 8: The German Traffic Sign Detection Benchmark (GTSDB) (ISI)

Name of the dataset	The German Traffic Sign Detection Benchmark (GTSDB) [20]
Time period covered	WP3 duration M6-M32
Purpose	These datasets will be used to train the machine learning models and to generate the attacks described in the section above.
Format	Images
Data origin	Visual datasets on traffic signs
Ownership	Publicly available datasets for traffic signs
Visibility and Sharing	Publicly available datasets
Description	<p>GTSDB is a single-image detection assessment for researchers with interest in the fields of computer vision, pattern recognition and image-based driver assistance.</p> <ul style="list-style-type: none"> • a single-image detection problem • 900 images (divided in 600 training images and 300 evaluation images) • division into three categories that suit the properties of various detection approaches with different properties

	<ul style="list-style-type: none"> an online evaluation system with immediate analysis and ranking of the submitted results
Storage	Redmine repository

Table 10: The German Traffic Sign Detection Benchmark (GTSDB) (ISI)

DATASET 9: Lyft data (ISI)

Name of the dataset	Lyft data [21]
Time period covered	WP3 duration M6-M32
Purpose	These datasets will be used to train the expected machine learning models and to generate the attacks described in the section above
Format	It includes over 55,000 human-labelled 3D annotated frames, a drivable surface map, and an underlying HD spatial semantic map to contextualize the data
Data origin	Visual datasets on traffic signs, publicly available datasets for raw sensor camera and LiDAR data
Ownership	Ownership to Lyft
Visibility and Sharing	Publicly available datasets/ available datasets produced by the automotive sensing community, which apart from the camera, incorporate other sensors as well
Description	<p>Mainfeatures:</p> <ul style="list-style-type: none"> Up to 7 cameras Up to 3 lidars Over 55,000 3D annotated frames A drivable surface map An HD spatial semantic map 4,000 Lane segments 197 Crosswalks
Storage	Redmine repository

Table 11: Lyft data

DATASET 10: Kitti data (ISI)

Name of the dataset	Kitti data [22]
Time period covered	WP3 duration M6-M32
Purpose	These datasets will be used to train the expected machine learning models and to generate the attacks described in the section above
Format	Stereo, odometry, 3D object detection, 3D tracking datasets
Data origin	Publicly available datasets for raw sensor camera and LiDAR data - Kitti
Ownership	Ownership to Kitti
Visibility and Sharing	Publicly available datasets/ available datasets produced by the automotive sensing community, which apart from the camera, incorporate other sensors as well.
Description	Kitti contains a suite of vision tasks built using an autonomous driving platform. The full benchmark contains many tasks such as stereo, odometry, 3D object detection, 3D tracking, etc.
Storage	Redmine repository

Table 12: Kitti data

DATASET 11: Motion Distorted Lidar data (ISI)

Name of the dataset	Motion Distorted Lidar
Time period covered	WP3 duration M6-M32
Purpose	These datasets will be used to train the expected machine learning models and to generate the attacks described in the section above
Format	In order to produce datasets embodying information from both camera and LiDAR sensors, we can use CARLA [23] simulator that generates synchronized LIDAR and camera data with object annotations reflecting real-life sensor arrays.
Data origin	Simulated data coming from CARLA simulator
Ownership	Publicly available datasets for raw sensor camera and LiDAR data.

	Ownership is to CARLA: An Open Urban Driving Simulator, Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, Vladlen Koltun; PMLR 78:1-16
Visibility and Sharing	Publicly available datasets /available datasets produced by the automotive sensing community, which apart from the camera, incorporate other sensors as well.
Description	<p>This is a large-scale dataset generated using CARLA, aiming to dynamic object detection.</p> <p>Main features:</p> <ul style="list-style-type: none"> • Town 1 (2.9 km of drivable roads with 90 vehicles) • Town 2 (1.9 km of drivable roads with 60 vehicles)
Storage	Redmine repository

Table 13: **Motion Distorted Lidar****DATASET 12: Operator Heart Rate (CRF)**

Name of the dataset	Operator Heart Rate
Time period covered	Covering all the trials period of the project
Purpose	The purpose of the data is mainly to check the operator's health and if an accident has occurred with the robot
Format	Operator Heart Rate: Text - .csv File
Data origin	<p>The data comes from:</p> <p>Observational → Sensors reading – Smartwatch</p>
Ownership	The owner of the data will be CRF, but they will be openly shared
Visibility and Sharing	There are ethical aspects, but it is possible to resolve by anonymizing the operators. We will give operators an information sheet and sign their consent for the use of data for the project. Use of pseudonyms for operators

Description	The data will be acquired during all the manufacturing use cases scenarios that will be implemented by CRF. If the operator were to suffer an injury from the heart rate, we could become aware of it and start the emergency procedures
Storage	Redmine and FCA private cloud/ Office Computer and private Cloud storage (FCA Google Drive). Data will made available through email or Redmine

Table 14: Operator Heart Rate

DATASET 13: Layout of the manufacturing cell/ CRF

Name of the dataset	Layout/CAD
Time period covered	From M12
Purpose	The aim is to share the cell layout with partners in a technically acceptable format
Format	Simulated - The CAD of the cell will be provided - JT (Jupiter Tessellation)
Data origin	Simulated data
Ownership	CRF
Visibility and Sharing	It will be shared with all project partners. For external sharing it is necessary to discuss it first
Description	JT is an ISO-standardized 3D data format. A .jt file will be shared which will contain the 3D Layout of the cell. CAD can be viewed with free software called JT2GO [24]
Storage	The data will be stored on a PC in CRF, on a private cloud of FCA and on Redmine

Table 15: Layout of the manufacturing cell

DATASET 14: Strength of the operator / CRF

Name of the dataset	Strength
Time period covered	From M24

Purpose	The aim is to understand the operator's fatigue when handling the Robot + Gripper
Format	Real Time - The format is not yet defined
Data origin	Real data coming from sensors
Ownership	CRF
Visibility and Sharing	It will be shared with all project partners. For external sharing it is necessary to discuss it first
Description	The gripper handles are sensorized to allow guided movement, we could use this same data to understand the effort that the operator is carrying out, it could help us understand the ergonomic index of the cell
Storage	The data will be stored on a PC in CRF, on a private cloud of FCA and on Redmine

Table 16: Strength of the operator

DATASET 15: Number of windshields/ CRF

Name of the dataset	Windshield camera
Time period covered	From M24
Purpose	The camera will be used to understand the number of windshields present in the container
Format	Real Time - The format will be defined as soon as the hardware is consolidated
Data origin	Real data
Ownership	CRF
Visibility and Sharing	It will be shared with all project partners. For external sharing it is necessary to discuss it first

Description	The camera will be aimed at the windshield container and an algorithm will be able to count the number of the windshield, when this is lower than a predetermined threshold an alarm will be activated to the PLC
Storage	The data will be stored on a PC in CRF, on a private cloud of FCA and on Redmine

Table 17: Number of windshields

DATASET 16: Safety Zone violation/ CRF

Name of the dataset	SafetyEYE/Light Curtain
Time period covered	From M24
Purpose	A SafetyEye and a Light Curtain will be used to ensure the safety of the work area
Format	Real Time - The format will be defined as soon as the hardware is consolidated
Data origin	Real data
Ownership	CRF
Visibility and Sharing	It will be shared with all project partners. For external sharing it is necessary to discuss it first
Description	The two systems will create two different datasets that will have the same purpose, that of identifying any violations of the safety zone.
Storage	The data will be stored on a PC in CRF, on a private cloud of FCA and on Redmine

Table 18: Safety Zone violation

In the Figure 3 below we provide a statistical 3D-pie that presents the number of CPSoSaware datasets per category

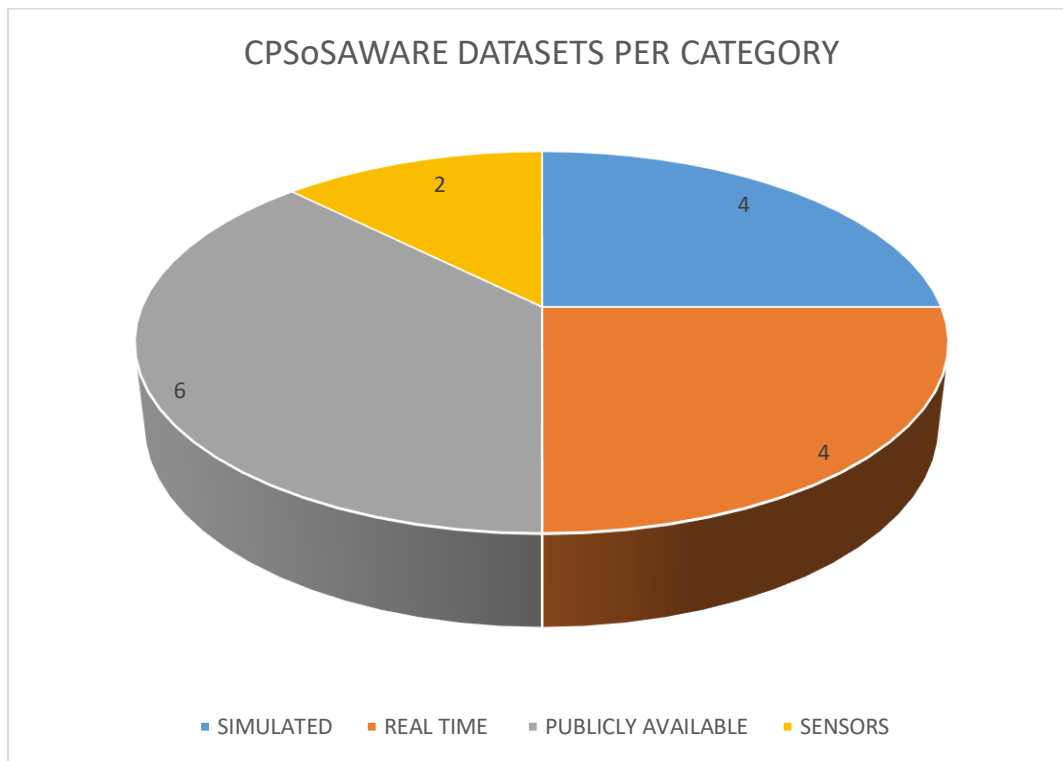


Figure 3: CPSoSaware Datasets per category

4.2 Data stakeholders in CPSoSaware

Data stakeholders within the project can come from many different backgrounds having different objectives and intentions **and some of the most important data stakeholders are:**

- Project partners that will generate these datasets such as
 - CRF
 - PASEU
 - ATOS
 - University of Peloponnese
 - ATHENA RESEARCH CENTER- ISI
- Consortium partners that will probably interact with these datasets within the framework of common tasks completion
- EIGHT BELLS as the dissemination leader
- European Commission
- Participants in the CPSoSaware trials
- Research communities – academia and industry based
- Standardisation organisations

Below the summary of datasets provided from the various project partners and described in detail in the previous section 4.1:

SUMMARY OF DATASETS

Name of the Dataset	Partner	Type	Accessibility
Intra-communication network performance	UoP	Simulated	As a general principle University of Peloponnese intends to make data freely available, although the ownership will always belong to the university.
Simulating and prototyping hardware designs	UoP	Simulated	As a general principle University of Peloponnese intends to make data freely available, although the ownership will always belong to the university.
XL-SIEM dataset	ATOS	Security events and security alarms - Real Time data	Security events and alarms may contain personal data, e.g. e-mail addresses, IP addresses. It is possible to implement anonymisation / pseudonymisation to de-identify any personal data included in security alarms, before sharing them.
EnvModel4AD (Environmental modelling for Autonomous driving)	PASEU	Sensor (cameras, lidar, GPS, Ultrasonic, Clock) readings And Wheel tick/ Flex Ray Measurements	<ul style="list-style-type: none"> • The Data will be shared on a network location, accessible by the related partners in accordance to the GDPR legislation. • Any personal Data will not be shared. • Data cannot be openly available unless another agreement is reached. • Data will be available for the duration of the project. Any usage beyond this date is dependent on a separate contract. • The data cannot be used after the end of the project.

Data Management Plan

			<ul style="list-style-type: none"> Any re-usage of the data out of the scope of CPSoSaware needs to be defined by a separate contract. The data are liable to Intellectual Property Ownership
LISA Traffic Sign	ISI	Traffic signs	Publicly available
Mapillary Global	ISI	Traffic signs	Publicly available
The German Traffic Sign Recognition Benchmark (GTSRB)	ISI	Traffic signs	Publicly available
The German Traffic Sign Detection Benchmark (GTSDDB)	ISI	Traffic signs	Publicly available
Lyft	ISI	Raw Sensor Camera and LiDAR Data	Publicly available
Kitti	ISI	Raw Sensor Camera and LiDAR Data	Publicly available
Motion Distorted Lidar	ISI	Raw Sensor Camera and LiDAR Data	Publicly available
Operator Heart Rate	CRF	Sensors reading – Smartwatch	Will be made publicly available
Layout/CAD	CRF	Simulated data	It will be shared with all project partners. For external sharing it is necessary to discuss it first
Strength	CRF	Real time data	It will be shared with all project partners. For external sharing it is necessary to discuss it first

Windshield camera	CRF	Real time data	It will be shared with all project partners. For external sharing it is necessary to discuss it first
SafetyEYE/Light Curtain	CRF	Real time data	It will be shared with all project partners. For external sharing it is necessary to discuss it first

Table 19: Datasets to be generated during project lifetime

4.3 Process to provide the Open Data

The project as an OPEN RESEARCH Data pilot has the objective of making data available publicly as much as possible. Of course, any potential data sharing through dissemination activities etc, will be evaluated taking into account the possibility of exploiting the data, and the ethical considerations described previously. As mentioned above the project will use Zenodo as the platform for storing and managing the data generated and OpenAire for linking the databases and publications. In addition, the project will disseminate public information through its website (<http://cpsosaware.eu/>) and the social media. Data that will be shared only within the consortium will be deposited in the Redmine [25] private area. Having defined the data management plan and the way to collect categorize and monitor the data, the dissemination leaders EIGHT BELLS along with the Project coordinator ATHENA RESEARCH CENTRE- ISI will ensure that this plan will be followed accurately and at a later stage when datasets become available a decision will be made on the choosing of potential publishing technologies that will be used for making available the open data at the related data repositories

In this regard, the project will progressively distinguish between:

- Internal datasets that cannot be shared publicly. Certain specific dataset can be shared only within the consortium towards deliverables and task completion.
- Open datasets that will be made open and available through OpenAire/ Zenodo, through the website or other means, in order to be accessible from third parties. The dissemination leader will be responsible to make available the various datasets generated through the project duration. Furthermore, open data that will be made available on the CPSoSaware website will be posted the respective sections such as the ones below:

<http://cpsosaware.eu/deliverables/>

<http://cpsosaware.eu/publications/>

The steps to be executed for providing and managing the open data sets will include:

1. Definition of the open datasets (through the questionnaires)
2. Select appropriate repository of the open CPSoSaware data (OpenAIRE),

3. Define additional storage places that the open data could be shared and disseminated and ensure that the open data will be available after the project lifecycle ending
4. Collect the CPSoSaware open data sets, and relevant metadata from all the partners that will be generating data in order to be uploaded to the hosting platform(s)
5. Upload the collected open data on the aforementioned selected storage platforms and the project's website
6. Make it possible for third parties to openly and freely access, exploit, reproduce and disseminate data
7. Provide information on the tools and instruments needed to validate the results (or provide the tools)

4.4 Process of data storage, preservation and security

This section documents the standards that will be used to ensure the integrity of the datasets, and how these datasets will be preserved and kept accessible during the project lifetime and afterwards. An efficient method of storage along with the necessary levels of access, are important considerations for comprehensive protection.

The Guidelines on Data Management in Horizon 2020 require defining procedures that will be put in place for long-term storage preservation of the project data and any other associated information. Data that will be put in storage must be protected from potential unauthorized access of any kind and as such datasets should be stored in repositories that provide a unique and persistent identification (an identifier). **The Zenodo repository possesses these archiving capabilities including backup and will be used to archive and preserve the project datasets and additionally other data such as software code, reports, articles, presentations etc.** The open data will be stored on the project website <http://cpsosaware.eu> and will be equipped with same security measures implemented for the website.

Furthermore, the following measures will be taken to ensure data security at each study site:

Security of electronic records:

- Electronic data will be stored on secure servers which are backed up daily.
- Data encryption – use of end-to-end encryption standards like AES and TLS/SSL technologies, with client-side key management is a popular mechanism that will be utilised and employed by CPSoSaware and will ensure strongest data protection.

Permission management will allow granular access to data, while activities related to files management will be monitored (who opened files, activity logs). The project will also enable back up options like: **deleted file recovery, and device control tools (such as access revoke, remote wipe)**. Furthermore, in order to prevent unauthorized access, of the project's data, the following data **security** measures will be realized:

- Data will be stored in online repositories which are password protected and/or grant access only upon correct identification.
- use of different layers of security and multi-factor authentication
- use extensive data control and governance features like permission management, security policies and access revoke

- For possible execution of focus groups sessions, consent has to be given by the participants in Germany and Italy, to use the data within the project. The collected data will be fully anonymized and any possible measures to protect the participants and avoid any privacy issues will be taken

The Redmine repository

Redmine was selected from the project coordinator ATHENA RESEARCH CENTER- ISI as the private area platform. Redmine is an open source project management web application that offers to the users various features such as flexible issue tracking system, Gantt chart and calendar, news, documents & files management etc. Written using the Ruby on Rails framework, it is cross-platform and cross-database. Through the Redmine repository, the CPSoSaware website provides to the consortium partners the ability to access a common project collaboration workspace, which represents a password protected repository that will be regularly updated with project documentation, and will be the area for accessing important project management information (reports, minutes, presentations, working documents, deliverables etc). As it contains confidential information, it is accessible only to registered consortium members and can be accessed through the “private area link” presented on the website main menu. The private area is organized per work package in order to stimulate collaboration between the consortium members, efficient production of deliverables and easy access to information. User accounts and privileges are managed by the project coordinator and there is no register procedure offered in the login form.

5 Conclusion

The aim of a data management plan is to prevent any unauthorized data disclosure, which can occur in many various forms either through dissemination activities, publications, press releases and in electronic, oral or written way.

This document presented the Data Management Plan for the CPSoSaware project following the Guidelines on Data Management in Horizon 2020. This document provides the OPEN RESEARCH DATA pilot principles and guidelines, ethical aspects, legal framework that will be taken into account towards the implementation of the project objectives. Moreover, it describes the data management life cycle for all datasets that will be collected, processed or generated by CPSoSaware project partners through the project duration. Sixteen different datasets have been identified and presented, provided from the project partners through the attached questionnaire. Some of these datasets refer to publicly available datasets while other refer to simulated data and data coming from raw sensor camera and lidar etc generated or used from partners such as ATOS, ISI, UoP, PASEU, CRF.

The Data Management Plan is a deliverable directly connected to the upcoming trials and pilot executions (WP6 activities) as well as to the ethical aspects and laws that rule the data privacy assessment and monitoring within the EU funding projects. For any data collected and generated during the project trials all security and ethical guidelines and standards will be applied. In addition, data owners will decide the way that will handle data visibility and sharing limitations taking into account the current document. Additionally any ethical considerations, especially about data protection, privacy and security will be discussed among the respective partners that will be generating the data, the coordinator, the dissemination leader and the project ethical board.


6 References

- [1] Open Research Data Pilot in H2020 <https://www.openaire.eu/item/open-research-data-pilot-in-h2020>
- [2] What is the EC Open Research Data Pilot? <https://www.openaire.eu/what-is-the-open-research-data-pilot>
- [3] Open Research Data Pilot in Horizon 2020. How can OpenAIRE help? [What is the Open Research Data Pilot? https://utlib.ut.ee/sites/default/files/openaire/OpenAIRE2020_FactSheet_DataPilot.pdf](https://utlib.ut.ee/sites/default/files/openaire/OpenAIRE2020_FactSheet_DataPilot.pdf)
- [4] H2020 Programme Guidelines on FAIR Data Management in Horizon 2020 https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- [5] FAIR Principles <https://www.go-fair.org/fair-principles/>
- [6] Implementing FAIR Data Principles: The Role of Libraries <https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf>
- [7] SUMMARY TABLE 1 FAIR Data Management at a glance: issues to cover in your Horizon 2020 DMP https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf#page=10
- [8] 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 (General Data Protection Regulation) (Text with EEA relevance) <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- [9] HORIZON 2020 ETHICS <https://ec.europa.eu/programmes/horizon2020/en/h2020-section/ethics>
- [10] OpenAire <https://www.openaire.eu/openaire-portal>
- [11] About Zenodo <https://about.zenodo.org/>
- [12] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm
- [13] NS3 Simulator <https://www.nsnam.org/>
- [14] VITIS <https://www.xilinx.com/products/design-tools/vitis.html>
- [15] gem5 Simulator <https://www.gem5.org/>
- [16] Cross-Layer SIEM (XL-SIEM) for Finance <https://finsecurity.eu/solutions/cross-layer-siem-for-finance/>

- [17] Mogelmoose, A., Trivedi, M. M., & Moeslund, T. B. (2012). Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Transactions on Intelligent Transportation Systems*, 13(4), 1484-1497.
- [18] Neuhold, G., Ollmann, T., Rota Bulò, S., & Kotschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4990-4999).
- [19] Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2011, July). The German traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks* (pp. 1453-1460). IEEE.
- [20] The German Traffic Sign Detection Benchmark
<http://benchmark.ini.rub.de/dev/index.php?section=gtsdb&subsection=news>
- [21] Lyft <https://level5.lyft.com/dataset/>
- [22] The KITTI Vision Benchmark Suite <http://www.cvlibs.net/datasets/kitti/>
- [23] The CARLA Simulator <https://carla.org/>
- [24] JT2Go <https://www.plm.automation.siemens.com/global/it/products/plm-components/jt2go.html>
- [25] Redmine platform <https://www.redmine.org/>

7 ANNEXES

ANNEX I: DATA MANAGEMENT QUESTIONNAIRE

Project Acronym	DATA MANAGEMENT QUESTIONNAIRE
CPSOSAWARE	
Description: <i>The intention of this questionnaire is to be used towards the writing of the CPSoSAWARE Data Management Plan that represents the D7.6 of the project</i>	
Name of the Dataset	<p><i>Instructions: Please provide a meaningful name so that we can refer to it unambiguously in the future</i></p> <p><i>The dataset is</i></p>
Purpose of the Dataset	<p><i>State the purpose of the data collection/generation, indicating the relation with the objectives of the project. Add additional objectives if necessary</i></p> <p><i>The purpose of the data is mainly to:</i></p>
Types and Format of data	<p>Describe the type of data used or generated within the project, specifying the form and format of the data.</p> <p>For instance is it Numerical, text, audiovisual simulated, reused?</p> <ul style="list-style-type: none"> • Text • Numerical • Audiovisual • Simulated • Reused <p>The data is :</p>
Reuse of existing data	YES/ NO

<p>Data origin</p>	<p>Define and describe the origin/source of your data. For instance data can be gathered from different sources.</p> <ul style="list-style-type: none"> • Observational → Data captured in real time - often not reproducible i.e. sensor readings, images, telemetries, sample data... • Experimental → Data from lab equipment, often reproducible, but with high costs - i.e. chromatograms, magnetic fields readings... • Simulation → Data generated by computational models where model and metadata are equally important to output data - i.e. climate models, economic models, materials models .. • Derived compiled → Data coming from analysis or compilation. Reproducible but with high costs - i.e. the results of text and data mining, compiled databases. • Reference of canonical → Collection or conglomeration of smaller (peer-reviewed) datasets published and curated - i.e. chemical structures, gene sequence databanks, spatial data portals. <p>The data comes from</p>
<p>Dataset is</p>	<p>Fixed: never change after being collected or generated.</p> <p>Growing: new data may be added, but the old data is never changed or deleted.</p> <p>Revisable: new data may be added, and old data may be changed or deleted.</p>
<p>Task – use case scenario</p>	<p>Purpose: to indicate task/subtask where Dataset was generated</p> <p><i>Instructions: Please provide a short overview of the use cases scenarios or the task that the data will be generated in and clarify how things will really happen during pilots, who will be involved, etc</i></p>
<p>Data owner/controller</p>	<p>Who is the data owner/ controller?</p>
<p>Time period covered by the Dataset</p>	<p>Purpose: Start and end date of the period covered by the Dataset</p>

Expected size /Quantity	Quantity IN MB/ GB In case of not just digital archiving is required, indicated quantities of other form of storage.
Data security and storage	<p>What provisions are in place for data security?</p> <p>Instructions: Please indicate any methods considered for secure Data storage and transfer of sensitive Data. Will be the data safely stored in certified repositories for long term preservation and curation?</p> <p>Types of storage can include (i.e. Office computer, Hard Drive, Tape back-up system, Institute network drive, Institute Central Data storage, private Cloud storage ...), briefly describing the data security policy applied.</p>
Necessary software	Purpose: Names of any special-purpose software packages required to create, view, analyse, or otherwise use the Data
Ethical aspects / Protection of Personal Data notification of processing operations	Are there any ethical or legal issues that can have an impact on Data sharing?
Details on the procedures for obtaining informed consent	<i>Instructions:</i> Please give details on the procedures for obtaining informed consent from the Data subjects (e.g. providing an information sheet together with the consent form).
NDA process	Is an NDA process necessary? If yes how you plan to set up (necessary steps) the process?
Protective measures	<i>Please indicate any such protective measures (e.g. use of anonymisation techniques, use of pseudonyms, non-disclosure of audio-visual materials, voice records, etc.)</i>

2 FAIR Data
2.1 Making Data findable (Dataset description: metadata, persistent and unique identifiers e.g.)
<p>2.1.1 Are the Data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism?</p> <p>Explain how data are documented and if metadata are provided, listing the information made available/discoverable.</p>
<p>2.1.2 Naming conventions used</p> <p>What naming conventions do you follow? Describe the system used to name and structure electronic files and folders. Refer also to any file renaming procedure or tools used</p>
<p>2.1.3 Search keywords approach</p> <p>Will search keywords be provided that optimize possibilities for re-use? Indicate the approach to keywords generation, indexing and tagging.</p>
<p>2.1.4 Do you provide clear version numbers?</p> <p>Describe the versioning and traceability approach used (especially if the dataset is growing or revisable).</p>
<p>2.1.5 What metadata will be created?</p>
<p>2.2 Making Data openly Accessible</p> <p><i>Instructions: which Data will be made openly available and if some Datasets remain closed, the reasons for not giving access; where the Data and associated metadata, documentation and code are deposited (repository?); how the Data can be accessed (are relevant software tools/methods provided)?</i></p>

<p>2.2.1 Data openly available</p> <p>Which Data produced and/or used in the project will be made openly available as the default?</p> <p>Indicate ownership of the data, if it is openly available or can be made openly available.</p>
<p>2.2.2 Data kept closed</p> <p>Indicate if data access is restricted, to what users, and explain the reasons.</p>
<p>2.2.3 How will the Data be made available? Indicate how you intend to make data available</p>
<p>2.2.4 Methods or software (SW) tools for data access</p> <p>What methods or software tools are needed to access the Data?</p> <p>Indicate methods and SW tools needed to access the data. Clarify if the relevant software (e.g. in open source code) is included in the data set.</p>
<p>2.2.5 SW documentation and other information needed</p> <p>Is documentation about the software needed to access the Data included?</p> <p>Indicate any specific SW documentation that is needed to access the data, or additional information that is needed to understand the data (i.e. abbreviations, supplementary notes).</p>
<p>2.2.6 Is it possible to include the relevant software?</p>
<p>2.2.7 Repository for deposit of data, metadata, documentation and code</p> <p>Where will the Data and associated metadata, documentation and code be deposited?</p> <p>Indicate the (open or private) repositories in which the data, metadata, documentation and code are restored and/or those in which they will be stored in the future.</p>

<p>2.2.8 Access restrictions</p> <p>Have you explored appropriate arrangements with the identified repository?</p> <p>IF YES</p> <p>Indicate if there are limitations and restrictions to access the data, and if they are linked to a specific timeframe. Explain how access will be provided after these restrictions are lifted.</p>
<p>2.2.9 Is there a need for a Data access committee?</p> <p>YES / NO</p>
<p>2.2.10 Are the conditions for access well described conditions?</p>
<p>2.2.11 How will the identity of the person accessing the Data be ascertained?</p> <p>For instance Login and password?</p>
<p>2.3 Making Data Interoperable</p> <p><i>Instructions: which standard or field-specific Data and metadata vocabularies and methods will be used</i></p>
<p>2.3.1 Are the Data produced in the project interoperable?</p> <p>Assess the level of interoperability of the dataset.</p>
<p>2.3.2 Standard vocabulary or mapping to commonly used ontologies</p> <p>What Data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?</p> <p>Refer to commonly used ontologies to map the dataset, considering also the use of existing common platforms and tools – e.g.: EMMO, BFO, MatONTO, Materials Ontology.</p>
<p>2.3.3 Data licensing for wide reuse</p> <p>If applicable, define data licensing approach for the dataset wide reuse. Indicate the chosen licenses tools.</p>

<p>2.4 Increase Data re-use</p> <p><i>Instructions: which Data will remain re-usable and for how long, is embargo foreseen; how the Data is licensed; Data quality assurance procedures</i></p>
<p>2.4.1 Timing of data availability for re-use (incl. indications on embargo)</p> <p>When will the Data be made available for re-use? If applicable, define the timeframe the data will be available for re-use..</p>
<p>2.4.2 Data usability by Third Parties (after the end of the project)</p> <p>Are the Data produced and/or used in the project useable by third parties, in particular after the end of the project? Indicate any limitation to the use of the data by Third Parties, after the end of the project.</p>
<p>2.4.3 Restrictions to data re-use</p> <p>Indicate and explain any restriction to the re-use of data (i.e. confidentiality agreements, other issues).</p>
<p>2.4.4 Length of time of data reusability</p> <p>How long is it intended that the Data remains re-usable? Indicate the time limit for the data reusability, if any.</p>
<p>2.4.5 Quality assurance process</p> <p>Are Data quality assurance processes described? If applicable please explain how quality of the data is assured, how the consistency and quality of data collection are controlled and documented.</p>
<p>3 Allocation of resources</p>
<p>3.1 Cost estimates</p> <p>What are the costs for making Data FAIR in your project? Estimate the costs for making your data FAIR (findable, accessible, interoperable and reusable)</p>

3.2 Cost coverage

How will these be covered? If able describe how you intend to cover these costs (i.e. institute dedicated resources, dedicated part of the project budget ...).

3.3 Data management responsibilities

Who will be responsible for Data management in your project? Identify responsibilities for data management of this dataset (with in your research group and institute, and within the project if applicable).

THANK YOU FOR COMPLETING THIS QUESTIONNAIRE!

ANNEX II: INFORMATION SHEET



INFORMATION SHEET

Project: CPSoSAWARE

Aim of the study

Description of the study and incentive

Privacy and anonymity

- All data gathered in our projects will be processed anonymously and only be used within this project. All participants personal info will be coded (for example using pseudonyms) in the analysis and reporting of the data. This means that your name will not be linked to the gathered information.

ANNEX III: XL-SIEM JSON Data Format

```
{
  "AlarmEvent": {
    "DST_IP_HOSTNAME": <string>,
    "RELATED_EVENTS": <string>,
    "DST_IP": <string>,
    "PLUGIN_NAME": <string>,
    "SRC_IP": <string>,
    "PRIORITY": <integer>,
    "RELIABILITY": <integer>,
    "SUBCATEGORY": <string>,
    "USERDATA3": <string>,
    "USERDATA4": <string>,
    "PLUGIN_SID": <string>,
    "USERDATA1": <string>,
    "USERDATA2": <string>,
    "ORGANIZATION": <string>,
    "CATEGORY": <string>,
    "PLUGIN_ID": <string>,
    "USERNAME": <string>,
    "FILENAME": <string>,
    "BACKLOG_ID": <string>,
    "RELATED_EVENTS_INFO": {
      "a": {
        "date": <string>, ◊ long format
        "plugin_id": <integer>,
        "log": <string>,
        "interface": <string>,

```

```
"dst_ip": <string>,
"src_ip": <string>,
"userdata7": <string>,
"fdate": <string>, ◇ YYYY-mm-dd HH:MM:SS
"userdata8": <string>,
"userdata5": <string>,
"userdata6": <string>,
"userdata9": <string>,
"userdata3": <string>,
"userdata4": <string>,
"userdata1": <string>,
"userdata2": <string>,
"src_port": <string>,
"plugin_sid": <integer>,
"event_id": <string>,
"filename": <string>,
"organization": <string>,
"dst_port": <string>,
"tzzone": <string>,
"device": <string>,
"username": <string>
},
"b": {
  "date": <string>, ◇ long format
  "plugin_id": <integer>,
  "log": <string>,
  "interface": <string>,
  "dst_ip": <string>,
  "src_ip": <string>,
```

```
    "userdata7": <string>,
    "fdate": <string>, ◊ YYYY-mm-dd HH:MM:SS
    "userdata8": <string>,
    "userdata5": <string>,
    "userdata6": <string>,
    "userdata9": <string>,
    "userdata3": <string>,
    "userdata4": <string>,
    "userdata1": <string>,
    "userdata2": <string>,
    "src_port": <string>,
    "plugin_sid": <integer>,
    "event_id": <string>,
    "filename": <string>,
    "organization": <string>,
    "dst_port": <string>,
    "tzone": <string>,
    "device": <string>,
    "username": <string>
  }
},
"PROTOCOL": <integer>,
"RISK": <integer>,
"SRC_PORT": <integer>,
"SENSOR": <string>,
"SRC_IP_HOSTNAME": <string>,
"SID_NAME": <string>,
"USERDATA7": <string>,
"DATE": <string>, ◊ YYYY-mm-dd HH:MM:SS,
```

```
"USERDATA8": <string>,  
"USERDATA5": <string>,  
"USERDATA6": <string>,  
"PASSWORD": <string>,  
"USERDATA9": <string>,  
"DST_PORT": <integer>,  
"EVENT_ID": <string>  
}  
}
```