Grant Agreement N°: 101020259
Topic: SU-DS02-2020

ARCADIAN-IoT

Autonomous Trust, Security and Privacy Management Framework for IoT

D1.6: Data Management Plan

Revision: v.1.0

Work package	WP1
Task	1.4
Due date	31/7/2021
Submission date	30/07/2021
Deliverable lead	IPN
Version	1.0

Abstract

Data management is a crucial part of responsible conduct of research, as it ensures the value of the research results and assists in preserving that value for future years.

This deliverable establishes the Data Management Plan for ARCADIAN-IoT project, based on Horizon 2020's template[1]. The document characterizes how data will be managed in terms of collection, storage and backup, security, preservation and sharing (in applicable cases). It also discusses initial plans regarding open-source software release.

The Data Management Plan~~t~~ will be updated every time concrete and relevant changes regarding data management are decided. Changes will be made available on-line, and provided in annex to periodic reports.

Keywords: Data Management Plan, Datasets, FAIR

Document Revision History

Version	Date	Description of change	List of contributor(s)
v0.1	18/06/2021	Draft version of some sections of DMP	Sérgio Figueiredo (IPN)
V0.2	29/06/2021	Added dataset templates, other information	Sérgio Figueiredo (IPN)
V0.3	08/07/2021	Added dataset inputs for industrial IoT domain	Alexandru Gilga (BOX2M)
V0.4	08/07/2021	Added dataset inputs for surveillance IoT domain	Pedro Colarejo (LOAD)
V0.5	09/07/2021	Integration and initial revision	Sérgio Figueiredo (IPN)
V0.6	12/07/2021	Added dataset inputs for medical IoT domain	Ricardo Ruíz (RGB)
V1.0	28/07/2021	Final review resulting from SAB comments	Sérgio Figueiredo (IPN)

Disclaimer

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1-template

Project co-funded by the European Commission under SU-DS02-2020		
Nature of the deliverable:	R	
Dissemination Level		
PU	Public, fully open, e.g. web	√
CI	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to ARCADIAN-IoT project and Commission Services	

** R: Document, report (excluding the periodic and final reports)*
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
OTHER: Software, technical diagram, etc

EXECUTIVE SUMMARY

This deliverable defines a preliminary version of the Data Management Plan for ARCADIAN-IoT. It presents the summary regarding data being generated or collected within the project, and current status of the project with respect to data Findability, Accessibility, Interoperability and Reusability, providing available details regarding each of ARCADIAN-IoT domain's targeted datasets. Future versions of this document will be released when the specification of the project's use cases is complete and whenever relevant.

TABLE OF CONTENTS

LIST OF TABLES

ABBREVIATIONS

API	Application Programming Interface
CPU	Central Processing Unit
CSV	Comma Separated Values
DOI	Digital Object Identifier
DPM	Data Management Plan
FAIR	Findable, Accessible, Interoperable and Reusable
GAP	Grant Agreement Preparation
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
IoT	Internet of Things
JSON	JavaScript Object Notation
LoRa	Long Range
N/A	Not Available
ORD	Open Research Data
RAM	Random Access Memory
TBD	To be Defined
XML	eXtensible Markup Language

1 INTRODUCTION

This is the first version of the Data Management Plan. Considering the project's early stage, and particularly the fact that the use cases and architecture are still in an early specification phase, there are several aspects and decisions which will only take place at a later stage. As such, it is important to consider that this deliverable will be updated whenever relevant updates or changes regarding data management take place. This is expected at least in three moments:

- Close to the release of use cases and requirements specification (M8), e.g. updating the types and/or format of data being generated or collected;
- Near the release of ARCADIAN-IoT architecture (M12), e.g. as a result of interface specification update;
- During the integration activities to take place in WP5.

This document is organized as follows:

- In section 2, background information regarding the Open Research Data (ORD) pilot and FAIR data is presented;
- Section 3 presents the data summary, providing an overview of the purpose of data collection, types and formats of data to be collected, target data to be reused and made available for reusage by the consortium or externally, among others;
- Section 4 provides – at the moment, partial – answers to questions regarding data Findability, Accessibility, Interoperability and Reusability (FAIR) principles, and also presents the defined dataset description template and currently available information regarding the datasets for the three domains addressed in the project, namely Emergency and vigilance using drones and IoT (domain A), secured early monitoring of grid infrastructures (domain B), and Medical IoT (domain C);
- Section 5 provides information on resources allocation for data management;
- Section 6 describes the methods to be adopted for data security, namely regarding its storage, sharing rules, and data disposal, deletion and destruction;
- In section 7, ethical aspects are briefly summarized, pointing towards other relevant deliverables of the project;
- Finally, section 8 wraps-up the document, presenting main take-aways of the DMP at it current state.

2 BACKGROUND

2.1 Open Research Data (ORD) Pilot

Through the **Open Research Data (ORD) Pilot,** the European Union encourages projects funded under the European Union Framework Programme for Research and Innovation Horizon 2020 to provide open access (free of charge, online access for any user) to research data generated in the context of H2020 projects. The ORD pilot is mainly aimed at improving and maximizing access to and re-use of research data (from H2020 projects), while considering the need to balance openness and protection of scientific information, commercialisation, and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions. Addressing such balance, the possibility for a project to opt out the ORD pilot is available during application stage, during grant agreement preparation (GAP) stage or after the signature of the grant agreement, i.e. during project execution.

The ARCADIAN-IoT consortium has decided to opt out from the ORD pilot process considering the incompatibility with the need for confidentiality linked to security and other legitimate reasons.

2.2 FAIR data overview

The data generated during and after all projects should follow the FAIR data principles that require that data are Findable, Accessible, Interoperable and Reusable. These requirements don't affect implementation choices and don't necessarily suggest any specific technology, standard, or implementation solution. In this direction, H2020 projects shall adopt methodologies for data generation, collection and sharing to ensure that:

- data are **findable** due to the exploitation of metadata for convenient data discovery and of standard persistent and unique identifiers (such as DOIs);
- data are openly **accessible**, where this is possible; adequate justification is required to be provided if otherwise. Towards this, projects shall use methods and tools for providing access to data along with any required complementary pieces of information, such as guidelines for repository access and use;
- data are **interoperable** and allow for data exchange and re-use among researchers due to the extended and targeted exploitation of standardized data representation formats, vocabularies, etc. or mappings when the former is not possible.
- data **re-use** is promoted through clarifying licenses.

The FAIR concept implementation of each project is documented in a DMP, which is a key element of good data management. DMPs help shape the data management life cycle principles to be followed by an H2020 project. These are created during the first 6 months of a project, and they are appropriately refined through its course so as to fulfil evolving requirements. The ARCADIAN-IoT consortium is expected to adhere to the conditions laid out in the DMP below, in which all details related to management of ARCADIAN-IoT research data are specified.

3 DATA SUMMARY

3.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

Data pertaining to IoT devices, IoT applications and users (persons) will be generated and collected for the purpose of addressing and validating the objectives defined in the Grant Agreement, namely:

- **Objective 1**: To create a decentralized framework for IoT systems;
- **Objective 2**: To enable security and trust in the management of objects' identification;
- **Objective 3**: To enable distributed security and trust in management of persons' identification;
- **Objective 4**: To provide distributed and autonomous models for trust, security and privacy -- enablers of a Chain of Trust;
- **Objective 5**: To provide a hardened encryption with recovery ability;
- **Objective 6**: To provide self and coordinated healing with reduced human intervention;
- **Objective 7**: To enable proactive information sharing for trustable Cyber Threat Intelligence and IoT Security Observatory.

3.2 What types and formats of data will the project generate/collect?

The project is expected to generate and collect data characterizing the following entities:

- **Persons**: such as physical characteristics, identity validation data (picture, name), voice, image or video captures, or location. In the medical IoT domain, patient health data / vital signs will additionally be monitored (e.g. heart rate, temperature, SpO2);
- **Devices**: IoT device identifier (for smartphones, drones, grid components, wearables), operation data (e.g. CPU or RAM usage), battery state, or location;
- **Networks** (e.g. 4G, 5G, LoRa): link bandwidth real-time / average usage, packet characteristics (protocol, source / destination ports & addresses);
- **Applications / services**: service / application status obtained through monitoring (e.g. CPU utilization, uptime, downtime, etc), telemetry data from sensors.
- The exact data formats will be discussed among partners and specified later.

3.3 Will you re-use any existing data and how?

The project will aim at reusing existing data available for research. In case multiple datasets concerning a common type of data are targeted, and taking into consideration the importance of their harmonization, data templates and lists of pre-defined values may be defined.

The identified datasets to be reused will not include personal data. The currently targeted datasets are N-BaIoT[2] and UNSW Bot-IoT[3].

3.4 What is the origin of the data?

The origin of the data will be quite diverse. For domain A, it can be anticipated that data will be obtained in controlled, experimental lab settings. For domain B, will be obtained both from BOX2M

[2] https://archive.ics.uci.edu/ml/datasets/detection_of_IoT_botnet_attacks_N_BaIoT
[3] https://research.unsw.edu.au/projects/bot-iot-dataset

experimental laboratory, and from live deployment in customer infrastructure. Finally, for domain C, data will be collected from real patients and in real conditions, during treatments and collecting information from RGBs equipment at the Navarra University hospital.

3.5 What is the expected size of the data?

The data will be of variable size, according to type (e.g. telemetry, audio, video, biometric data) and format (e.g. raw, encoded, encrypted or compressed). Initially, each domain owner will be responsible for storing its datasets. Subject to discussions, some datasets are expected to be shared with the consortium, with different needs to be taken into account (e.g. anonymization in the case of Medical IoT). In those cases, a shared repository will be compiled to realise the research objectives, and is expected to be hosted through one of two options: 1) at IPN's data storage facilities, or 2) using the project's dedicated Microsoft Sharepoint instance, which is hosted in European servers (safeguarding GDPR requirements). Some specific, partial data, may be hosted by other partners (e.g. in cases data sharing among the consortium is not possible due to confidentiality reasons).

3.6 To whom might it be useful ('data utility')?

Most datasets which will come out of this project will be in the shared repository (to be defined), the main target group being the consortium partners themselves, which may leverage the datasets as benchmarks for future research.

Additionally, specific datasets which are decided to be made public, if any, will be stored in an open access data repository to be decided later in the project (e.g. Zenodo[4]). Those datasets will be aimed at ensuring that relevant research communities (e.g. addressing ehealth, industrial IoT or public safety) may benefit from the selected data sets.

[4] https://zenodo.org/

4 DATASETS HANDLING AND DESCRIPTION

4.1 FAIR data

4.1.1 Data findability

During the experiments, when applicable, metadata will be automatically attached, and collected via a graphical user interface by the researcher. Such metadata will then be uploaded to the project database with references to the raw data. Common metadata structures will be adopted among partners whenever possible.

For data agreed to be made public, curated datasets will be published on an open repository (e.g. Zenodo) making them accessible and discoverable, and indexed (e.g. a possibility is to use the EU Open Data Portal[5]). All published data sets will receive a DOI that will be referred to in any scientific publication that made use of this data set.

4.1.2 Data accessibility

Data

Regarding data agreed to be made public, the uploading of curated data sets will be done in an open format (e.g. CSV) to the agreed repository, under a licence such as MIT license[6], which has limited restriction on reuse, as it allows both non-commercial and commercial use. The metadata will be provided in a readable format (JSON, XML or CSV).

Software

As for the software developed during the project, the hosting location will still be defined, as well as the method for accessing and retrieving it. Regarding open-source software release plans, XLAB plans to continue working some existing libraries, like GoFE[7] and CiFEr[8]. Also, for the secure multi-party computation, some extensions for SCALE-MAMBA might need to be implemented[9]. RISE expects that all its respectively developed software will be released as open-source.

4.1.3 Data interoperability

Considering the project will collect data concerning multiple domains (e.g. medical IoT vs smart grids), different domain-specific standards may be considered – still to be discussed, as currently available dataset information is limited.

Data such as time series, will be stored in open formats, following a domain-specific standard, if any exists. If the need to use wider standards is identified, proper mappings will be developed during the project. To ensure machine readable or actionable data sets, CSV, JSON or XML will be used as formats for the meta data.

[5] http://data.europa.eu
[6] https://opensource.org/licenses/MIT
[7] https://github.com/fentec-project/gofe
[8] https://github.com/fentec-project/CiFEr
[9] https://github.com/KULeuven-COSIC/SCALE-MAMBA

4.1.4 Data reusability

In some specific cases data and software will be licensed (e.g. using MIT licenses), enabling both research and industry to re-use ARCADIAN-IoT data or software. During the project, datasets and software Identified as relevant and ready for release will be curated and published.

The setup of a data quality check inline through the development of continuous integration pipelines will be considered, helping ARCADIAN-IoT partners to validate datasets, ensure code correctness and increase re-usability. The published datasets will also be referenced in scientific publications, using their DOIs, and at conference presentations, to maximise re-use.

4.2 Dataset description template

In order to identify and collect relevant properties regarding the different datasets to be defined within the project, centred on FAIR data-related issues, a common template – based on the one from SecureIoT project[10] – was established.

This template is necessary for laying ground for future DMP versions, towards its completeness, as it represents a common – but adjusted to the project needs – format for characterizing which, how, when data will be used or produced, and involving what entities, among other relevant aspects; while a significant portion of the information is not available yet, this structure works as such as a beacon to be revisited and updated by the consortium.

We now present and provide details on the template defined for describing the datasets:

4.2.1 General information

- **Reference Number**: Sequence Number
- **Title**: Title of the Dataset
- **Version**: Dataset version
- **Description**: Brief description of what data represents
- **Data origin**: Whether data already exists or date it is expected to be released
- **Dataset availability**: Date of the dataset availability
- **Future revisions anticipated**: Whether future revisions are anticipated
- **Owner**: Provider of the datasets
- **Contact Person**: Person in charge of the release of the dataset and its inclusion in the ARCADIAN-IoT repository
- **Related Use Cases**: The set of ARCADIAN-IoT use cases that the dataset related. The description of the use cases is performed with reference to deliverable D2.1.
- **Utility / Potential Use**: Potential usages of interest by ARCADIAN-IoT target community. Potential examples include research / experimentation, service development / integration, training / education

4.2.2 Context

- **Observable human data (if applicable):** which human data will be collected (e.g. biometric data, vital signs)
- **Directly observable device**: devices for which data will be purposely collected (e.g. sensor, wearable, edge node, gateway)
- **Directly observable software**: software from which data will be purposely collected, e.g. IoT application, gateway software, cloud service application

[10] "Deliverable D1.3 – Data Management Plan", SecureIoT Project, June 2018

- **Indirectly observable device**: devices which are not directly monitored, be exhaustive to the extent possible. i.e., Sensor, robot, vehicle board, monitor device, edge node, gateway.
- **Indirectly observable software**: List the software which is observed indirectly.

4.2.3 Data access

Three possibilities are identified (with the first two potentially coinciding):

1. Data is already retrieved and stored as data files

2. Monitoring data can be retrieved through an interface

3. Data is present in SW/HW but no means exists yet to access them remotely (implies the need for a probe to be developed).

Data access has the following attributes:

- **Dataset provided as data file(s)**: Define if the dataset is provided as data file(s)
- **Remote accessibility**: Define if the data are remotely accessible and how (e.g. protocol, message format, pull/push, interface, etc).
- **In case data are not yet accessible, target retrieval mode**: Define the method which enables the data to be accessed in the future.

4.2.4 Data description

- **Data format**: e.g. NetFlow, pcap, syslog, json (when an interface is used, the format of embedded data should be described)
- **Encryption**: if and how the data are encrypted
- **Data format description**: syntax and semantics of data, particularly important for non-standard formats (e.g. describe the columns of a csv file, or the structure and semantics of what contains a JSON file)
- **For unusual format, tool to read it**: if their data format is not standard, the required tool/library to read the data
- **Dataset generation**: whether data was monitored in a system with real users. If not, describe how data have been generated
- **Attack**: whether the dataset contains attacks. If so, whether attacks are annotated, and what is the granularity of the annotations
- **Dataset statistics**: i.e., Duration, size(s) in appropriate format (MB, packets, etc), number of packets broke down per IP address, protocols… (be exhaustive as possible)

4.2.5 Data restrictions

- **Publicly available data:** Whether data are public
- **(If not) Plan to make data open**: Whether there is a plan to make data public.
- **Data accessibility**: Whether data can be accessible to the consortium, or to specific partner(s) in case they cannot be public.
- **(If yes) Accessibility period**: What is the time period the data can be accessible to the consortium, or to specific partner(s)
- **Possibility of data dissemination**: Specify if the data can be used for public dissemination (without revealing the full content of the data, aggregated view)
- **Data ownership**: identification of the data owner
- **Legal issues**: Specify the confidentiality level of the dataset and the license under which the dataset could be opened and offered publicly.

4.3 ARCADIAN-IoT datasets

This is the first version of the DMP, and given that the project is in a very early stage, most of the information characterizing the datasets has not been determined yet.

Currently available information has been collected as a result of the work being developed on WP2, namely the use cases and requirements definition and initial architecture discussions.

4.3.1 Emergency and vigilance using drones and IoT

4.3.1.1 General

Table 1. General information for Emergency and vigilance domain

Field	Description
Reference Number	1
Title	Specific data sets for Emergency and vigilance using drones and IoT
Version	1
Description	TBD
Data origin	Drone
Dataset availability	TBD
Future revisions anticipated	TBD
Owner	LOAD
Contact person	Davide Ricardo
Related use cases	TBD (under specification)
Utility / potential use	Emergency Face recognition Threats signs / public security

4.3.1.2 Context

Table 2. Context information for emergency and vigilance domain

Field	Description
Observable human data	Biometric data
Directly observable devices	Drone sensors (light, temperature, camera, microphone)
Directly observable SW	IoT application
Indirectly observable devices	TBD
Indirectly observable SW	TBD

4.3.1.3 Data access

Table 3. Data access information for emergency and vigilance domain

Field	Description	
Dataset provided as data file(s)	*Yes*	
Remote accessibility	Yes / No	Yes
	Protocol	TBD
	Message format	TBD
	Pull/push	TBD
	Provided interface	TBD
(In case data not accessible) Target retrieval mode	TBD	

4.3.1.4 Data description

Table 4. Data description for emergency and vigilance domain

Field	Description	
Data format	JSON / TBD	
Encryption	TBD	
Data format description	TBD	
(Non-standard format) Tool for reading	TBD	
Dataset generation	Data monitored in a real system with real users	*Yes*
	If not, data generation approach	TBD
Attack	Datasets containing attacks	TBD
	(If yes) Annotated attacks	TBD
	(If yes) Granularity of the labels / annotations	TBD
Dataset statistics	NA	

4.3.1.5 Data restrictions

Table 5. Data restrictions for emergency and vigilance use case

Field	Description

Publicly available data	No
Plan to make data open	TBD
Data accessibility	Yes
Accessibility period	TBD
Possibility of data dissemination	TBD
Data ownership	TBD
Legal issues	Personal Data Protection issues

4.3.2 Secured early monitoring of grid infrastructures

4.3.2.1 General

Table 6. General information for grid infrastructures domain

Field	Description
Reference Number	2
Title	Specific data sets for grid infrastructures domain
Version	1
Description	To be updated
Data origin	Customer live infrastructure & BOX2M lab
Dataset availability	After TRL4 achieved
Future revisions anticipated	TBD
Owner	BOX2M ENGINEERING
Contact person	Alexandru Gliga / Ovidiu Diaconescu
Related use cases	Industrial manufacturing infrastructures (production lines with electrical engines); industrial exploitation infrastructures (rigs); industrial energy distribution infrastructures (substations) with multiple purpose commercial application (building, factory, warehouse, lightning network). Concrete use cases TBD (under specification)
Utility / potential use	Grid & utilities companies Oil & gas & mining companies Any manufacturing / maintenance entity Any building service manager Any public authority (for its operated portfolio, e.g. buildings and lightning) Any strategic or critical entity (defence, telecom) Machineries & gear vendors (willing to embed the technology into their products)

4.3.2.2 Context

Table 7. Context information for grid infrastructures domain

Field	Description
Observable human data	N/A
Directly observable devices	IoT devices (acting as gateways)
Directly observable SW	Device firmware; Encryption middleware platform; End IoT platform
Indirectly observable devices	Field sensors (through IoT devices); communication modules (through IoT devices)
Indirectly observable SW	3rd party platforms connected to both middleware platform or End IoT platform

4.3.2.3 Data access

Table 8. Data access information for grid infrastructures domain

Field	Description	
Dataset provided as data file(s)	Yes (CSV format)	
Remote accessibility	Yes / No	Yes (option 1 - from software platform front end; option 2 – by API, using a 3rd party platform configured properly and credentials provided by software platform owner)
	Protocol	HTTPS (TLS)
	Message format	To be updated
	Pull/push	By front end = pull (http request); by API = both possible, up to type of 3rd party platforms
	Provided interface	By front end = GUI; by API = back-end coding, customised (as it is today); GUI (TRL6 stage of software platform)
(In case data not accessible) Target retrieval mode	API & front end, for any nonexposed yet data; requires specific development (coding)	

4.3.2.4 Data description

Table 9. Data description for grid infrastructures domain

Field	Description	
Data format	TBD	
Encryption	TBD: 2 formats are under evaluation, both using hybrid concept proposed (hardware & software)	
Data format description	TBD	
(Non-standard format) Tool for reading	TBD: it will be deployed according to type of encryption implemented; it is also requested by ENISA	
Dataset generation	Data monitored in a real system with real users	Yes (even in lab environment, data is real, being generated by specific sensors network integrated with IoT devices)
	If not, data generation approach	We will consider simulator data generation only after validating the technology with live data; reason for simulator generation is just for stress tests / big data tests
Attack	Datasets containing attacks	TBD
	(If yes) Annotated attacks	TBD
	(If yes) Granularity of the labels / annotations	TBD
Dataset statistics	TBD; conditioned firstly by TRL4 achieving and multiple type of trial customers implemented	

4.3.2.5 Data restrictions

Table 10. Data restrictions for grid infrastructures domain

Field	Description
Publicly available data	Not in the moment
Plan to make data open	Once TRL4 achieved, for the deployed technology
Data accessibility	TBD: it may be shared with any interested benefiter, after a dedicated assessment done by BOX2M and an approval by the consortium
Accessibility period	TBD
Possibility of data dissemination	High level concept = YES Type of crypto chip = NO (just shared with the consortium)

	Open source used in firmware and middleware platform = YES Specific code and open source adaptions = NO Type of sensors = YES IoT device (type, vendor) = YES IoT device & monitored energy circuits IDes = NO Communication module: technology=YES, type=NO, vendor=YES Effective customer energy data = YES (anonymized vs. customer name, location, type of industrial asset monitored) Penetration tests methodology = YES Penetration tests results = YES Commercial application domains & industries = YES
Data ownership	Customer (if it is used a live infrastructure) or BOX2M (if it is used the lab)
Legal issues	Customer NDA and / or compliancy stipulated into contract

4.3.3 Medical IoT

4.3.3.1 General

Table 11. General information for medical domain

Field	Description
Reference Number	3
Title	Specific dataset for Medical Domain
Version	1
Description	Vital signs from patient
Data origin	Patient. Recollected by RGB modules and sent by the smartphone.
Dataset availability	After TRL4 achieved
Future revisions anticipated	TBD
Owner	RGB/UNAV
Contact person	Ricardo Ruiz
Related use cases	TBD (under specification)
Utility / potential use	Hospitals. Patients with a special treatment.

4.3.3.2 Context

Table 12. Context information for medical domain

Field	Description
Observable human data	N/A
Directly observable devices	Phone devices (acting as gateways) WEB browser
Directly observable SW	Device firmware
Indirectly observable devices	N/A
Indirectly observable SW	Firebase encryption services

4.3.3.3 Data access

Table 13. Data access information for medical domain

Field	Description	
Dataset provided as data file(s)	*Yes (JSON)*	
Remote accessibility	Yes / No	Yes
	Protocol	HTTPS (TLS)
	Message format	TBD
	Pull/push	By front end = pull (http request);
	Provided interface	By front end = GUI;
(In case data not accessible) Target retrieval mode	API & front end, for any nonexposed yet data; requires specific development (coding)	

4.3.3.4 Data description

Table 14. Data description for medical domain

Field	Description	
Data format	TBD	
Encryption	Firebase Encryption	
Data format description	HTTPS	
(Non-standard format) Tool for reading	TBD	
Dataset generation	Data monitored in a real system with real users	*Yes (even in lab environment, data is real,*

		being generated by users connected to RGB modules)
	If not, data generation approach	When we don't have a user to generate the data or if we need to do a specific test, we use the saved data from another day repeated in a loop.
Attack	Datasets containing attacks	No
	(If yes) Annotated attacks	
	(If yes) Granularity of the labels / annotations	
Dataset statistics	Conditioned in the first place by the achievement of TRL4 and multiple types of test clients implemented	

4.3.3.5 Data restrictions

Table 15. Data restrictions for medical domain

Field	Description
Publicly available data	No (just into consortium)
Plan to make data open	No (just into consortium)
Data accessibility	No (just into consortium)
Accessibility period	TBD
Possibility of data dissemination	No (just into consortium)
Data ownership	Customer (if it is used a live infrastructure) or RGB/CUN (if it is used the lab)
Legal issues	Customer NDA and / or compliancy stipulated into contract

5 ALLOCATION OF RESOURCES

Most information regarding resources allocated for data management is still to be discussed by the consortium partners, e.g. estimated costs, main responsible for data management, and particularly the potential costs for long term preservation. At the moment, based on the limited information, it is expectable that domain' dataset owners will be responsible for all "owned" dataset decisions.

6 DATA SECURITY

6.1 Storage of digital data

Securing stored data involves preventing unauthorized people from accessing it, and preventing accidental or intentional destruction, infection or corruption of information. Steps to secure data involve understanding applicable threats, aligning appropriate layers of defence and continual monitoring of activity logs taking action as needed. This means that a multi-tier approach needs to be adopted from all the partners.

The proper method of storage and the appropriate community along with levels of access for privileged users are important considerations for comprehensive protection. Improperly stored information along with overly permissive accounts are a centralized topic in many high-profile breaches. All partners within the ARCADIAN-IoT project will follow a specific set of guidelines to comply with the project's requirement for storage of digital data:

- Data availability must be guaranteed;
- Confidential data must be stored using access protection;
- Strictly confidential information must only be stored in an encrypted mode;
- Confidential data must not be stored in online services which were not previously approved by the ARCADIAN-IoT consortium;
- Any exception from this measure must explicitly be approved by the project's Steering Committee;
- Modifications to data with high integrity requirements must be documented and approved by the partners.

6.2 Sharing of data

Data sharing in the context of ARCADIAN-IoT refers to the process of making confidential data available to authorized partners. To prevent impact on the confidentiality and integrity of data while it is being shared (and enabling auditability trail in case of compromise), a set of processes need to be adopted between all partners. Shared confidential data is often copy-protected to prevent the creation of unauthorized copies from malicious actors. The following practices will be applied by all partners:

- For the exchange of confidential data, only services (particularly online ones) approved by ARCADIAN-IoT must be used;
- Strictly confidential data sent by email must be encrypted;
- Encryption/decryption keys and other access mechanisms need to be communicated between the partners in a secure manner
- A process to rotate keys and access controls in case of compromise will be implemented.

6.3 Data disposal, deletion and destruction

Protecting confidential and sensitive data from accidental disclosure is of paramount importance. A key area in data security is the disposal of confidential data, in both electronic and paper formats. Confidential information discarded in the trash or recycling bin is legally and effectively open to anyone. Additionally, so is any data stored on discarded or donated computer technology, like hard drives and thumb drives. Electronic data kept beyond its usefulness invites mischief or accidental breach. The secure disposal, deletion, and destruction of data aims to make data unrecoverable from other parties. The following best practices concerning disposal of no longer useful data will be adopted:

- Confidential paper documents which are no longer needed must be disposed of using data protection boxes or shredded
- Confidential data which is no longer needed must be securely erased
- Data storages of mobile end devices and data carriers which are no longer needed must

be securely erased.

- If it is not possible to erase data storages of mobile end devices or data carriers then the end device or the carrier must be destroyed.
- Tamper-resistant hardware platforms such as secure elements, secure enclaves, SIMs, etc., which are used to store confidential data must be destroyed.

7 ETHICAL ASPECTS

The identification and management of potential ethics issues within ARCADIAN-IoT project is addressed in Task 1.5 (Ethics and security management). Thus, an Ethics guide (D1.6[11]) including recommendations regarding ethics such as the involvement of humans or management of personal data is released in M3.

Additionally, several ethical requirements (e.g. the need to establish an Ethics Advisory Board, clarification on involvement of children or vulnerable humans, etc) were identified following the Ethical screening procedure at the Grant Agreement Preparation phase. These will be addressed in WP7 – Ethics Requirements.

[11] ARCADIAN-IoT D1.6 – "Ethics Guide", to be released July 2021

8 CONCLUSIONS

This document provided the preliminary data management plan for ARCADIAN-IoT project. The current DMP provides an initial positioning of the consortium with respect to the research data Findability, Accessibility, Interoperability and Reusability; given the 3 addressed domains, it is possible to observe the diversity in data – spanning both types, formats, restrictions and sharing possibility. Moreover, while the project did opt out of the ORD pilot, the public availability of some of the datasets to be obtained experimentally within the project is expectable in some cases and as long as some requirements are fulfilled (e.g. data anonymization is enforced).

This first version leaves some important aspects unaddressed, such as providing incomplete information regarding data formats, future dataset sharing, or allocated resources; nevertheless, it is a living document which will be updated whenever relevant advances – e.g. resulting from discussions between consortium partners, progresses in use cases specifications - take place.

ANNEX – DMP KEY ISSUES

The DMP structure was influenced by H2020's DMP template[12]. The table below provides a summary of the addressed DMP issues – serving as guideline for future updates of the document:

DMP component	Issues to be addressed
Data summary	• State the purpose of the data collection/generation • 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) • 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
FAIR Data Making data findable, including provisions for metadata	• Outline the discoverability of data (metadata provision) • 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
Making data openly accessible	• Specify which data will be made openly available? If some data is kept closed provide rationale for doing so • 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
Making data interoperable	• 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?
Increase data re-use (through clarifying licences)	• Specify how the data will be licenced to permit the widest reuse possible • 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 useable 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

[12] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#A1-template

Allocation of resources	• 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

Data security	• Address data recovery as well as secure storage and transfer of sensitive data
Ethical aspects	• 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
Other	• Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)