



**The Framework Programme for
Research & Innovation Actions (RIA)**

Project Title:

Secure and Safe Internet of Things



SerIoT

Grant Agreement No: 780139

[H2020-IOT-2017] Secure and Safe Internet of Things

Deliverable

D10.4. Data Management Plan

Deliverable No.		D10.4	
Workpackage No.	WP10	Workpackage Title and Task type	Project Management
Task No.	T10.3	Task Title	Management of Data, Knowledge & IPR issues
Lead beneficiary		IITiS	
Dissemination level		PU – Public	
Nature of Deliverable		ORDP: Open Research Data Pilot	
Delivery date		30 June 2018	
Status		F: final	
File Name:		[SerIoT] D10.4-Data_Management_Plan_final_R1_V1.1.pdf	
Project start date, duration		01 January 2018, 36 Months	



This project has received funding from the European Union's Horizon 2020 Research and innovation programme under Grant Agreement n°780139

Authors List

Leading Author (Editor)				
<i>Surname</i>	<i>Initials</i>	<i>Beneficiary Name</i>	<i>Contact email</i>	
Gelenbe	EG	IITIS	e.gelenbe@imperial.ac.uk	
Co-authors (in alphabetic order)				
<i>#</i>	<i>Surname</i>	<i>Initials</i>	<i>Beneficiary Name</i>	<i>Contact email</i>
1	Nowak	SN	IITIS	snowak@iitis.pl
2	Domańska	JD	IITIS	joanna@iitis.pl
3	Drosou	AD	CERTH	drosou@iti.gr
4	Baldini	GB	JRC	gianmarco.baldini@ec.europa.eu
5	Monschiebl	BM	Atech	Bernhard.Monschiebl@austriatech.at
6	Ververidis	CV	HIT	c.ververidis@hit-innovations.com
7	Txomin	TR	Tecnalia	txomin.rodriquez@tecnalia.com

Reviewers List

List of Reviewers (in alphabetic order)				
<i>#</i>	<i>Surname</i>	<i>Initials</i>	<i>Beneficiary Name</i>	<i>Contact email</i>
1	Thomos	TN	UESSEX	nthomos@essex.ac.uk
2	Ramos	JLHR	JRC	jose-luis.hernandez-ramos@ec.europa.eu

Document history			
Version	Date	Status	Modifications made by
V0.1	2018.05.25	1st draft version	IITiS
V0.2	2018.06.15	2nd draft version, including input from partners	IITiS
V0.3	2018.06.18	3rd draft version, including input form CERTH, that clarify information about central, public repository.	IITiS
V0.4	2018.06.19	4nd draft version, including comments from internal reviewer (UESSEX)	IITiS
V1.0	2018.06.27	Final version, including comments from internal reviewer (JRC) and input from TECNALIA (in 1 Data Summary chapter)	IITiS
V1.1	2018.06.29	Final version	IITiS

List of definitions & abbreviations/terms

Abbreviation/ Term	Definition
CIA triad	Confidentiality, integrity and availability of data are often denoted as CIA triad, as a model designed to guide policies for information security within an organization.
C-ITS	Cooperative Intelligent Transport Systems
D	Deliverable
Dataset	Digital information created in the course of research but which is not a published research output. Research data excludes purely administrative records. The highest priority research data is that which underpins a research output. Research data do not include publications, articles, lectures or presentations.
DMP	Data Management Plan: A formal working document which outlines how datasets will be handled both during the active research phase and after the project is completed. DMPs must be addressed at the earliest phase of the research lifecycle
DOI	Digital Object Identifier - a persistent identifier or handle used to uniquely identify objects, standardized by the International Organization for Standardization (ISO).
DSP	Description Set Profiles

ETSI	European Telecommunications Standards Institute
IoT	Internet of Things
JSON	JavaScript Object Notation
LOD	Linked Open Data
M	Month
Metadata	Information about datasets stored in a repository/database template. A text document's metadata may contain information such as the document's size/length, the author's name, when the document was written, and a short summary of the document.
SDN	Software Defined Network
SerIoT Repository	A central SerIoT digital repository is a mechanism for managing and storing digital content.
SerIoT	"Secure and Safe Internet of things" project
SVN	Apache Subversion (abbreviated as SVN), a software open source under the Apache License versioning control system, distributed as.
T	Task
VCS	Veritas Cluster Server
VOC	Volatile Organic Compounds
WAPI	WLAN Authentication and Privacy Infrastructure
WP	Work Package
XML	eXtensible Markup Language

Executive Summary

This report describes the initial Data Management Plan for the SerIoT project, funded by the EU's Horizon 2020 Programme under Grant Agreement number 780139. The purpose is to set out the main elements of the SerIoT consortium data management policy for the datasets generated by the project.

The DMP presents the procedure for the management of datasets created during the lifetime of the project and describes the key data management principles. Specifically, the DMP describes the data management life cycle for all the datasets to be collected, processed and/or generated by a research project, including the following processes:

- handling of research data during and after the project implementation,
- what data will be collected, processed or generated,
- what methodology and standards will be applied,
- whether data will be shared/made open, and
- how data will be curated and preserved.

The methodology for the DMP is as follows:

1. Create a general data management policy (the strategy that will be used by the consortium to address all the datasets);
2. A DMP template will be created and sent to the partners of the consortium in order to be filled with information for each relative data set;
3. Analyze the completed by the project's partners DMP templates.
4. Creating an updated version of SerIoT project DMP.

The current document formulates the general data management policy. The project's partners provided preliminary information. DMP Template was created (see Appendix 1). As the detailed assumptions regarding Use Cases will be formulated by M12, the DMP template will be sent for complement/ revised and the DMP document will be updated accordingly.

The initial version of the SerIoT DMP was developed according to guidance by EUROPEAN COMMISSION (HORIZON 2020): HORIZON 2020 DMP [1].

The structure of the document is as follows: Section 1 provides the initial assumptions of the datasets generated during the lifetime of the project, including assumed types and formats of data, the expected size of the datasets and the data utility. The specific description of how SerIoT will make this research data findable, accessible, interoperable and reusable (FAIR) is outlined in Section 2. Sections 3 to 6 outline the policy in relation to data resources, security and ethics. Section 6 contains the conclusions.

Table of Contents

List of definitions & abbreviations/terms.....	3
Executive Summary	5
List of figures	8
Project Participants	9
1 Data Summary.....	11
1.1 Purpose of the data collection and generation	11
1.2 Types and formats of datasets.....	12
1.3 Origin of data	14
1.4 Re-use of existing data.....	14
1.5 Expected size of the data	14
1.6 Data utility.....	14
2 FAIR data	14
2.1 Data management policy	14
2.2 Making data findable, including provisions for metadata	15
2.2.1 Data identification	15
2.2.2 Naming convention	15
2.2.3 Version number	16
2.2.4 Metadata	16
2.2.5 Dataset description	17
2.2.6 Keywords	17
2.3 Making data openly accessible	17
2.3.1 Data repository.....	17
2.3.2 Methods or software tools are needed to access the data	18
2.3.3 Data access committee.....	18
2.3.4 Identity of the person accessing the data	18
2.4 Making data interoperable	19
2.4.1 Interoperability of data	19
2.4.2 Metadata vocabularies, standards or methodologies for interoperability	19
2.5 Increase data re-use (through clarifying licenses)	19
2.5.1 Licensing and data sharing policies	19
2.5.2 Data availability for re-use	19
3 Allocation of resources.....	19
3.1 Long term data preservation	20
4 Data security	20
5 Ethical aspects.....	20






6	Other issues.....	20
6.1	National and EU regulations	20
7	Conclusions	20
	References.....	22
	Appendix 1 DMP Template.....	23

List of figures

Fig. 1 SerIoT global acquisition architecture.	11
Fig. 2 SerIoT central repository (screenshot of main page).	18

Project Participants

	<p>Instytut Informatyki Teoretycznej i Stosowanej Polskiej Akademii Nauk (IITIS, Coordinator, Poland)</p>
	<p>Centre for Research and Technology Hellas, Information Technologies Institute (CERTH, Quality Manager, Greece)</p>
	<p>Joint Research Centre – European Commission (JRC, Belgium)</p>
	<p>Technische Universität Berlin (TUB, Germany)</p>
	<p>Deutsche Telekom AG (DT, Germany)</p>
	<p>Hispasec Sistemas S.L. (HIS, Spain)</p>
	<p>HOP UBIQUITOUS SL (HOPU, Spain)</p>
	<p>Organismos Astikon Sygkoinonion Athinon (OASA, Greece)</p>
	<p>ATOS SPAIN S.A. (ATOS, Spain)</p>
	<p>University of Essex (UESSEX, UK)</p>

	<p>Institute of Communication and Computer Systems (ICCS, Greece)</p>
	<p>Fundacion TECHNALIA Research & Innovation (TECNALIA, Spain)</p>
	<p>AUSTRIATECH - GESELLSCHAFT DES BUNDES FÜR TECHNOLOGIEPOLITISCHE MASSNAHMEN GMBH (AUSTRIATECH, Austria)</p>
	<p>Grupo de Ventas Hortofrutícolas (GRUVENTA, Spain)</p>
	<p>HIT Hypertech Innovations LTD (HIT, Cyprus)</p>

1 Data Summary

1.1 Purpose of the data collection and generation

The main purpose of data generation and collection is to introduce the prototype implementation of IoT platform across all IoT domains (e.g., embedded mobile devices, smart homes/cities, security & surveillance, etc.) and to optimize the information security in IoT networks in a holistic, cross-layered manner (i.e., IoT platforms and devices, Honeypots, Fog networking nodes, SDN routers and operator's controller).

SerIoT will produce a number of datasets/databases during the lifetime of the project. The data will be both analyzed using a range of methodological perspectives for project development and scientific purposes, and will be available in a variety of accessible data formats.

The data sources are IoT devices and IoT systems deployed in the use cases locations, created in-cooperation with project's four industrial partners: OASA, AustriaTech, HOPU and Tecnalia (see Fig.1). According to that on general level four separate datasets categories will be created. For example:

- Intelligent Transport Systems in Smart Cities: Partner AustriaTech, which will provide data from the Road Side stations within the development phase of the SerIoT's monitoring, will include several data sets with C-ITS messages.

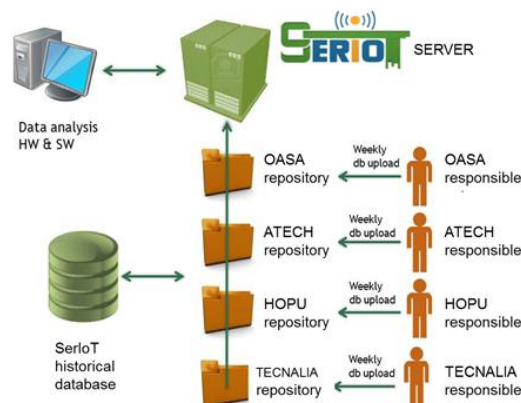


Fig. 1. SerIoT global acquisition architecture.

The main goal of work in the project is designing and implementing the SerIoT system and its components. Specifically, the data generated within WP1 and WP2 will have impact on architectural formal modelling, analysis, and synthesis, verification of the SDN-Controller and Secure Router Design as well as automated penetration testing.

Two types of data will be generated in WP2. The first type will be models, in a selected model checker language. The second type of data generated in this WP will be the security, safety and quantitative properties of the IoT communication architecture. Such security properties are in accordance with rules of confidentiality, integrity, availability, authentication, authorization and non-repudiation. Quantitative properties could be for example, "what is the probability of a failure causing the system to shut down within 4 hours?", "what is the probability of the protocol terminating in error state, over all possible initial configurations?".

Furthermore, for task T2.3, formal verification will work at the property level (a group of output points make up a property). A common misconception, which should be avoided, is

that formal verification ensures a 100% bug-free design. Simulation evaluations are not as effective in detecting all the potential issues within today's complex chips, despite the big progress that has been achieved in stimulus generation. Besides, extracted netlists of modern designs are in most cases too large for (re-)simulation, which creates a gap in the verification flow. On the other hand, given a property, formal verification exhaustively searches all possible input and state conditions for failures. Therefore, if the set of properties formally verified, collectively constitutes the specifications, it can be assumed that a design meets its specifications. Formal verification can be classified further into two categories, equivalence checking which determines whether two implementations are functionally equivalent and property checking that takes in a design and a property which is a partial specification of the design and proves or disproves that the design has the property.

According to results of WP1 and WP2, the SDN network, which is the core of SerIoT network system, will be implemented. WP3 partners will develop and implement software algorithms and methods described in WP1 and WP2, as well as algorithms and methods developed within work of WP4 partners. The outcome of the work will be the source code of prototype modules, extending capabilities of SDN switch and SDN controller, as well as capabilities of existing fog architecture elements.

The source code will be programmed in e.g. C, C++, Java, Python, etc., accompanied by files enabling their compilation and deployment in devices used for testing - project files, makefiles etc., and split into programming projects. Thus, team members will be able to download, compile and deploy the code for testing, bug fixing or further development.

The safe and reliable way of depositing the source code is using the VCS repositories. Repositories of VCS may be stored locally on partners' servers or on external servers. Part of solutions will be shared with the community as Open Source projects. In that case, public repositories like GitHub or GitLab will be used.

The source code of software implementing new methods developed in the SerIoT project will be created during work in mainly WP3, WP4, WP5 and WP6. The code development may concern also WP2 (e.g., extensions to model verification software), WP7 (e.g., software enabling integration of solutions developed by partners) and WP8 (e.g., test scripts).

The largest volume of projects' data will be obtained from test sites. Corresponding WP4, (IoT Monitoring Security and Mitigation) will deal with the research and development of a cross-layer data collection infrastructure, as well as the actual data generated by IoT devices.

More specifically, the data will be delivered by IoT data collection infrastructure and will include key measurement mechanisms in order to deal with the effective management of information related to the IoT threat landscapes. Moreover, a sophisticated multi-layer anomaly detection framework will run on the datasets and will detect in early stages malicious attacks at peripheral devices, honeypots and core network nodes. Real-time data will be processed in order to extract important IoT system features for anomaly detection monitoring and generate design-driven security monitors. The reason is to detect non-consistent IoT behaviors utilizing IoT design artefacts such as requirements, architecture and behavioral models. Effective and resource-efficient cross-layered mitigation techniques deployed on the data, will tackle with emerging vulnerabilities in the IoT landscape. Finally, the data processed through a robust cross-layer Decision Support framework will assist in the identification of malicious activities, attacks and root cause analysis related to the IoT-enabled ecosystem.

1.2 Types and formats of datasets

Data produced by SerIoT includes the following categories: experimental data (related to data from test sites), models, simulations and software. At the current, early stage of the project

implementation, the final list of datasets, formats and access rules cannot be predicted in detail. Most of the project data will be related to testing in test sites - real IoT environments (according to list of assumed use cases and scenarios). The general SerIoT data types are presented in Table 1.

Table 1. Dataset generic description fields.

#	Datasets	Related WP
1	Models, system design, specifications	WP1, WP2
2	Repository of codes, code documentation	WP2, WP3, WP5, WP6
3	UC1: Surveillance	WP7, WP8
4	UC2: Intelligent Transport Systems in Smart Cities	WP7, WP8
5	UC3: Flexible Manufacturing Systems	WP7, WP8
6	UC4: Food Chain	WP7, WP8

According to each pilot use case, types and formats of anonymized data, collected for SerIoT will differ. For example, HOPU will provide data regarding food chain supplies such as temperature or humidity while AustriaTech will provide data regarding vehicle traffic or emergencies on the road. The related datasets will consist of C-ITS messages, captured by a test vehicle with the use of dedicated application. These datasets will be provided as Wireshark PCAPs files with the same capture protocol stack: ETH/GN/BTP/<target>, “target” being <CAM|DENM|IVI|SPAT|MAP> with GeoNetworking/BTP stack transport type Single-Hop Broadcasting (SHB).

All this data will be collected by the data acquisition platform (WAPI server) in order to be processed then by the different modules.

Gruventa will provide data collected from the vehicle (track), and will contain tracks’ information (Vehicle ID, total Km covered by the vehicle, partial Km, GPS status, GPS position, date, time, dashboard alerts, dashboard light status, on board temperature, outdoor temperature, insight temperature, Humidity, VOC level).

In order to perform the evaluation in the Automated Driving Scenario (TECNALIA) different performance indicators will be assessed using a range of measures that will be monitored and logged. For that, both, sensors and questionnaires, will be used depending on the nature of the performance indicator. The required measures to calculate and evaluate the performance indicators will be defined in a validation matrix. Raw data will be acquired through sensors, intended a sensor as any method to obtain relevant data in the tests. This information will be post-processed obtaining the derived measures from raw data, and also synchronizing the data coming from different data loggers in order to have coherent global registers from TECNALIA’s pilot site. Then data will be logged to a local data base following a data format and table structure agreed by project partners. TECNALIA will store also the logging files in their own local server. After storing all logged data in the local server, these files will be sent to the SerIoT Central Data Repository using ftp communications. This repository is nowadays available and there is a directory for each pilot site with enough space to store all data to be logged for the project. This will allow partners (within the evaluations in WP8) to compile early reports and also to provide a backup service for the pilot sites.

Detailed information of the use cases and application scenarios are currently formulated (more detailed assumptions will be made in second version of D1.2, that will be issued by M12) and the detailed analysis will be finished with the second issues of D1.2 deliverable. Thus, the first version of the document presents the basic assumptions.

1.3 Origin of data

In SerIoT, the assumed origins of data are:

- Honeypots (WP5)
- SDN router packet inspection (WP3, WP5)
- SDN router high-level communications (WP3)
- Different IoT traffic (devices, vehicular IoT etc). (WP4, WP7, WP8)

The honeypots will provide data dynamically to the detector algorithms, to detect anomalies and malware installed on the device. Different IoT traffic is the traffic collected from test sites (data collected from IoT devices, sensors, actuators and additional modules installed). SDN router packet inspection data results from analysis of flows of data that are analyzed, collected and sent to controllers by network nodes. SDN router high-level communications are collected by higher layers of the SerIoT framework e.g. related to timely information to/from analytics module, root causes analysis & mitigation engine, multi-level visualization engine.

1.4 Re-use of existing data

In specific cases datasets already exist, e.g., obtained by industrial partners from existing IoT environments. For example, for the Smart City Use Case (with key partner AustriaTech) some of C-ITS data already exist and will be used to develop the monitoring application of SerIoT. This can be used to improve recognition of incorrect information and be able to therefore monitor the incoming as well as outgoing communications of the Road Side Stations. Those data were previously captured during testing and evaluation of C-ITS projects which use the ETSI standardized C-ITS specifications.

1.5 Expected size of the data

Dataset size might vary, depending on the pilot and the amount of information sent to the data collection infrastructure by each IoT sensor. The dataset size corresponds also to the amount of messages collected during the operation and the needs of the monitoring device. The expected size of the produced Use Cases datasets will be between 5MB and 5GB.

The other datasets (related to WP1-3) are code repositories, model descriptions, modeling and simulation results. The expected sizes will be relatively small of about 1GB.

The information about expected and actual sizes of the data will be updated.

1.6 Data utility

Except the internal needs to use the dataset (in order to develop SerIoT component e.g. monitoring application for C-ITS Road Side Stations, and test them), the data may be useful for research purposes in future projects, which have interest in IoT devices and Cyber-security. Moreover, the dataset will include data related to automated transport, and will be useful to researchers in automated transport more focused on secure communications for safety.

2 FAIR data

2.1 Data management policy

In general, being in line with the EU's guidelines regarding the DMP [1], each dataset collected, processed and/or generated in the project comprise of/ includes the following elements:

1. Dataset reference and name
2. Dataset description

3. Standards and metadata
4. Data sharing
5. Archiving and preservation

All datasets in project repositories (public and confidential) will be supplemented with additional metadata, identifiers, keywords as described in the following subsection, where, we provide a generic description of datasets elements in order to ensure their understanding by the partners of the consortium.

2.2 Making data findable, including provisions for metadata

At first, databases are created and used by corresponding WPs and maintained in local repositories of responsible partners. At this stage, datasets will be confidential and only the members participating in the deployment of WPs or the consortium members will have access to them. Then, the selected data will be made accessible through the data repository (See 2.3.1). The more detailed specification of the datasets that will be available to the public will be presented in the updated version of DMP.

A DOI is assumed to be assigned to key datasets (assumed at least for the central repository) for effective and persistent citation. DOI will be assigned when a dataset is uploaded to the repository. This DOI can be used in any relevant publications to direct readers to the underlying dataset.

2.2.1 Data identification

SerIoT will follow the minimum Data Cite metadata standards [2] in order to make data infrastructure easy to cite, as a key element in the process of research and academic discourse. Recommended DataCite format for data citation is relatively simple and follows the format:

Creator (PublicationYear). Title. Publisher. Identifier

It may also be desirable to include information about two optional properties, Version and Resource Type. If so, the recommended form is as follows:

Creator (PublicationYear). Title. Version. Publisher. ResourceType. Identifier

E.g.

Organisation for Economic Co-operation and Development (OECD) (2018-04-06). Main Economic Indicators (MEI): Finance | Country: Argentina | Indicator ID: CCUS, 01/1959 - 12/2017. Data Planet™ Statistical Datasets: A SAGE Publishing Resource [Dataset]. Dataset-ID: 062-003-004. <https://doi.org/10.6068/DP163F9ED671E6>

2.2.2 Naming convention

SerIoT naming convention for project datasets will comprise of the following:

1. A prefix "SerIoT" indicating a SerIoT dataset.
2. A unique chronological number of the dataset
3. The title of the dataset
4. For each new version of a dataset it will be allocated with a version number which will, -for example, start at v1_0.
5. A unique identification number linking e.g. with the dataset work package and/or deliverable/task, e.g., "WP4_D4.3".

E.g.

SerIoT.11234.serSDN_edge_node_traffic.v1_12.WP3_T2

2.2.3 Version number

On general level simple version numbering is assumed. Version number consists of Version/ subversion (e.g. measurements_1.12). For specific cases selected database versioning best practices are recommended and applied (e.g. for integration of source code with external databased in [3]).

For the WP2 purposes (IoT Architectural Analysis & Synthesis) two stages of formal modelling and analysis are assumed. Therefore, there will be two version numbers. The first stage is preliminary, counting from M1 to M12. Within this stage, the infrastructure will be set up and studied. Moreover, formal modelling in architectural and high performance level will be conducted. The second stage, counting from M13 will perform formal modelling and verification in code level. More specifically, scripts will be run in order to observe if particular parameters or constraints are being verified.

For the code versioning compilation number will be included to the version number. The code versioning system will also be adopted (e.g. SVN). Tutorial and examples can be found e.g. in SVN Tutorial [5].

2.2.4 Metadata

The specific metadata content is presented in table below. The content (See) is preliminary, contains generic data and will be further defined in future versions of the DMP.

The assumed file format for metadata is XML. The detailed metadata structures to describe specific content will be developed and presented in an updated version of DMP. Additionally, content specific metadata are linked in the generic description.

Table 2. Dataset generic description fields.

Dataset Name	Name according to naming conventions
Title	The specific title of the dataset
Keywords	The keywords associated with the dataset
Work Package	Related Work Package of the project
Dataset Description	A brief description of the dataset
Dataset Benefit	The benefits of the dataset
Type	Type of dataset (XML, JSON, XLSX, PDF, JPEG, TIFF, PPT)
Expected Size	The approximate size of the dataset
Source	How/why was the dataset generated
Repository	Expected repository to be submitted
DOI (if set)	The DOI assigned (if valid) when dataset has been deposited in the repository
Date of Submission	The date of submission to the repository once it has been submitted
Date of Update	The date of update
Publisher/Responsible partner	Lead partners responsible for the creation of the dataset
Version	Version/ subversion number (to keep track of changes to the datasets)
Link to additional metadata	Link to content specific metadata (will be defined in next versions of DMP)

2.2.5 Dataset description

There are not formal requirements for the dataset description formulated yet. In particular, the description will depend on the type of a dataset. However, it is recommended to publishers of dataset to provide following information:

- The nature of the dataset (scope of data)
- The scale of the dataset (amount of data)
- To whom could the dataset be useful
- Whether the dataset underpins a scientific publication (and which publications)
- Information on the existence (or not) of similar datasets
- Possibilities for integration with other datasets and reuse

It is also possible that the description will have additional internal structure (XML).

2.2.6 Keywords

Search keywords will be provided when the dataset is uploaded to the repository, which will optimize possibilities for reuse of the data. Keywords will be part of a general metadata structure.

2.3 Making data openly accessible

In general, research data is owned by the partner who generates the data. Each partner has to disseminate its results as soon as possible, unless there is a legitimate interest to protect the results. WP leaders will propose the access procedure to their developed datasets, conditions for making datasets public (if applicable) and specify the embargo periods for all the datasets that will be collected, generated, or processed in the project. In case the dataset cannot be shared, the reasons for this will be mentioned (e.g., ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related, etc.).

A partner that intends to disseminate its scientific results has to give an advance notice to the other partners (at least 45 days) together with sufficient information on the results it will disseminate (according to Grant Agreement). Research data that underpins scientific publications should be by default deposited in the SerIoT data repository as soon as possible, unless a decision has been taken to protect results. Specifically, research data needed to validate the results in the scientific publications should be deposited in the data repository at the same time as publication.

2.3.1 Data repository

Data and the associated metadata, documentation and code will be initially deposited in the repositories created by each SerIoT pilot. For example, AustriaTech will provide credentials to WP4 partners that need access to their datasets. HOPU will store data securely on a platform on the FIWARE and MongoDB architecture. Then, the dataset could be transferred to the SerIoT data repository (<https://opendata.iti.gr/seriot>), which is established by Partner CERTH (see Fig. 2).

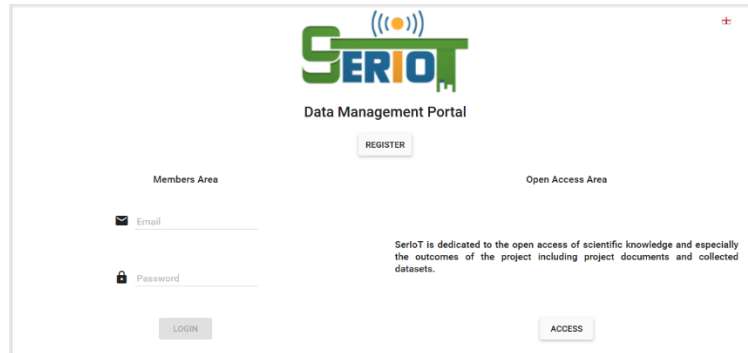


Fig. 2. SerIoT central repository (screenshot of main page).

The datasets will be confidential and only the task / consortium members will have access to them. If a dataset or specific portions of it (e.g., metadata, statistics, etc.) is decided to become of widely open access, it will be uploaded to the SerIoT open data platform. This data will be anonymized, in order to avoid any potential ethical issues with their publication and dissemination.

2.3.2 Methods or software tools are needed to access the data

Two ways of accessing the data repository exist. Firstly, (email and password) credentials are needed in order to have administrator privileges. Such privileges are online editing of the datasets, adding new datasets or downloading public datasets. On the other hand, a single button tagged as “ACCESS”, gives anonymized access to the public datasets only for the purpose of downloading them

To process data stored in a form of XML or JSON files there are available libraries to process the data. For some scientific data MATLAB or OCTAVE have to be used. To use and compile data in code repositories the related software platform has to be used (Linux with JAVA, Python, C++, etc.). The information about platforms will be included to content specific metadata.

2.3.3 Data access committee

The Grant Agreement (GA) does not describe the existence of a data access committee. The access policies will be determined by the owners of the data in agreement with the coordinator and related WP leaders/ partners.

2.3.4 Identity of the person accessing the data

Confidential datasets are stored in each responsible partner’s local repository, accessed by credentials. When the datasets are agreed to become public, only then the data is uploaded to the SerIoT open access repository.

Each pilot use case stores its datasets into local repositories at its premises. The partner provides credentials to other partners that need access to their confidential datasets, and the accessing person is identified. And only when pilots agree, together with the rest consortium, the datasets become anonymized and are uploaded to central open access SerIoT repository (see 2.3.1).

The public part is open accessed and no identification of the person accessing the data is assumed. Valid credentials to identify accessing person are required for editing or uploading new data.

2.4 Making data interoperable

2.4.1 Interoperability of data

The data produced in the SerIoT project are interoperable, allowing data exchange and re-use only inside the SerIoT consortium. Thus, since the SerIoT consortium is composed of fifteen partners from eight different countries, data exchange and re-use will be accomplished.

The use cases data will be first collected by the data acquisition platform, which consists of a WAPI server. Then, the WAPI server will be responsible for the data distribution amongst the different modules, such as the Analytics and the Decision Support System (DSS) module or the Mitigation Engine module. This way the data is made interoperable, allowing re-use between the SerIoT modules (for the use cases data modules developed within WP4).

2.4.2 Metadata vocabularies, standards or methodologies for interoperability

A metadata schema which defines constraints about metadata records is a fundamental resource for metadata interoperability. Existing metadata schemas are assumed to be used, to develop a new schema in order to minimize newly defined metadata vocabularies [6]

Key concepts considered are DSP as a formal basis of metadata schema and LOD, as a framework to connect metadata schema resources. We assume to study and apply two approaches:

- search metadata terms and description set profiles using resources registered at schema registries and the like,
- search metadata terms using metadata instances included in a LOD dataset.

2.5 Increase data re-use (through clarifying licenses)

2.5.1 Licensing and data sharing policies

In general, the coordinator partner (IITIS) along with all work package leaders, will define how data will be shared. WP leaders will propose the access procedure to developed datasets, set conditions for making it public (if applicable), set the embargo periods, the necessary accompanied software and other tools for enabling re-use, for all datasets that will be collected, generated, or processed in the project. In case the dataset cannot be shared, the reasons for this will be mentioned (e.g., ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related, etc.). The plan will be prepared in advance and will be presented in updated version of DMM.

Detailed data sharing policies have not been decided yet but European Union Public License (EUPL) V. 1.1 is considered [4] as a license that has been created and approved by the European Commission.

2.5.2 Data availability for re-use

The time for making the data available for re-use, has not been decided yet. It has also not been decided yet for how long the data will remain re-usable.

3 Allocation of resources

The data repository has been created by the responsible partner (CERTH) to the extent of making data 'FAIR'. In order to access the public datasets stored in the repository for editing or uploading new ones, valid credentials are required. Whereas, only downloading the data does not require any credentials. Furthermore, the repository will use the HTTPS protocol, which helps in the authentication of the accessed repository and protection of the privacy and

integrity of the exchanged data while in transit. The coordinator partner (IITIS) is responsible for the data management.

3.1 Long term data preservation

Resources for long-term data preservation are intended to be discussed in future meetings of the SerIoT project. The details will be presented in the updated version of DMP.

The long-term preservation of open to public datasets is assumed, by archiving them for at least 5 years after the end of the project. The partners will decide and describe the procedures that will be used in order to ensure long-term preservation of the remaining data sets.

4 Data security

Pilot/use cases data in the first period of the project will be stored in use cases partners' repositories. In terms of WP4 data, the CIA triad principles will be followed. The CIA includes the principles confidentiality, integrity, and availability, which are the heart of information security. In other words, confidentiality is the property, that datasets are not made available or disclosed to unauthorized individuals, entities, or processes. Integrity stands for maintaining and assuring the accuracy and completeness of data over its entire lifecycle. Lastly, the meaning of availability is to ensure that the data is available at all times when it is needed.

A central repository (with valid HTTPS certificate) created by CERTH will be maintained for long-term preservation. In this repository, the portion of the dataset that is not restricted by intellectual property rights will be decided to become of open access, whereas the other will remain confidential and will not be uploaded to this repository. The repository will be periodically backed up.

5 Ethical aspects

The SerIoT project has taken into account ethical and legal issues that can have an impact on data sharing, and has dedicated a WP (WP11: Ethics requirements) to ensure compatibility of the activities carried out with ethical standards and regulations. Under this WP, the relevant complex, legal and ethics issues will be tackled. Moreover, deliverable D11.1 (title "*H - Requirement No. 1*") is pointing out ethical issues, including informed consent for data sharing and long-term preservation rules (included in questionnaires concerning personal data).

In order to make the widest re-use possible, the data will be anonymized, to avoid any potential ethical issues with their further distribution. Since the datasets in most cases will not contain personal information (name, surname) data sharing can be spread amongst third parties. In case of confidential datasets containing sensitive information, the re-use of the data will be possible by third parties in order to avoid any potential ethical issues.

6 Other issues

6.1 National and EU regulations

Regulations based on the country of origin of the dataset together with the regulations of the country where the data will be processed, will be followed. More specifically, D11.1 points out all national and EU regulations to be followed.

7 Conclusions

The purpose of this document was to provide the initial plan for managing the data generated and collected during the SerIoT project. Specifically, the DMP described the data management

life cycle for all datasets to be collected, processed and/or generated by a research project. Following the EU's guidelines regarding the DMP, this document will be updated. Current version was created in early state of the project (M6) and details regarding data that will be produced by use cases has not formulated yet.

Dataset from test sites will be supplemented with the ID, metadata and (if applicable) with the related software and documentation. It is assumed to provide at least one dataset of each scenario to the public, available through central SerIoT repository. Finally, datasets will be preserved after the end of the project on the web server.

References

- [1] European Commission (2016), Guidelines on FAIR Data Management in Horizon 2020. Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (accessed 22 July 2016)
- [2] Data Citation Synthesis Group (2014). Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11 <https://www.force11.org/group/joint-declaration-data-citation-principles-final>
- [3] Database versioning best practices, <https://enterprisecraftsman-ship.com/2015/08/10/database-versioning-best-practices/>
- [4] European Union Public License, <https://joinup.ec.europa.eu/collection/eupl/eupl-guidelines-faq-infographics>
- [5] SVN Tutorial, <https://www.tutorialspoint.com/svn/index.htm>
- [6] Find and Combine Vocabularies to Design Metadata Application Profiles using Schema Registries and LOD Resources, <http://dcpapers.dublincore.org/pubs/article/view/3675>

Appendix 1 DMP Template

Data set reference and name	<i>Identifier for the data set to be produced (assumed in DMP naming convention should be used)</i>
Owner of dataset	<i>Responsible partner that owns the data</i>
Data set description	<i>Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale. Information on the existence (or not) of similar data and the possibilities for integration and reuse. Types and expected sizes should be provided.</i>
Standards and metadata	<i>Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created. As a general, general metadata proposed in the SerIoT DMP should be provided.</i>
Dataset storage	<i>Destination storage (central SerIoT repository is suggested, another destination should be justified)</i>
Utilization of dataset	<i>Description to whom the dataset could be useful, and whether it underpins a scientific publication, or technical progress of the SerIoT project.</i>
Archiving and preservation (including storage and backup)	<i>Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.</i>
Data Sharing	<i>Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).</i> <i>In case a dataset cannot be shared, please mention the reasons for this (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).</i>