AIoTwin

Twinning action for spreading excellence in Artificial Intelligence of Things

DELIVERABLE D5.1

DATA MANAGEMENT PLAN

Funded by the European Union

Project number: 101079214
Project name: Twinning action for spreading excellence in Artificial Intelligence of Things
Project acronym: AIoTwin
Call: HORIZON-WIDERA-2021-ACCESS-03
Topic: HORIZON-WIDERA-2021-ACCESS-03-01
Type of action: HORIZON Coordination and Support Actions
Granting authority: European Research Executive Agency

DOCUMENT CONTROL	
Deliverable No.	**D5.1**
Title	Data Management Plan
Lead Editor	
Type	Report
Dissemination Level	PU
Work Package	WP5
Due Date	30.06.2023

AUTHOR(S)

Name	Partner	e-mail
Lodovico Giaretta	**RISE**	**lodovico@kth.se**
Dora Kreković	**UNIZG-FER**	**dora.krekovic@fer.hr**
Ivan Čilić	**UNIZG-FER**	**ivan.cilic@fer.hr**

AMENDMENT HISTORY

Version	Date	Author	Description/Comments
0.01	10-05-2023	L. Giaretta	First draft open for comments
0.02	06-06-2023	D. Kreković I. Čilić	Expanded sections on FAIR Data, Allocation of Resources, Data Security, Ethical Aspects and Intellectual Property Rights
0.02	14-06-2023	I. Podnar Žarko	Updates of sections 2, 3 and 4
0.03	23-06-2023	P. Frangoudis	Internal review
1.00	12-07-2023	I. Podnar Žarko	Final version submitted to EC

Table of Contents

Executive Summary

The Data Management Plan (DMP) presented here is part of the Project Management activities within the AIoTwin project funded under Grant Agreement No. 101079214. This document serves as a comprehensive guide describing the data management procedures in line with the principles of Findable, Accessible, Interoperable, and Reusable (FAIR) data management and the requirements of the Horizon Europe programme.

The DMP sets out the key aspects that will be monitored and the necessary actions to be undertaken for the management of research data within the AIoTwin project. It provides a framework to ensure that the project's data is of high quality, secure and sustainable, while enabling its accessibility and reusability as much as possible, adding value to the research community.

The DMP includes measures to ensure the protection of personal data and confidential information. It outlines strategies for handling and securing sensitive data in accordance with applicable regulations and best practises.

The Data Management Plan is a living document: Its first full version is produced by M6 and its content is revised throughout the life of the project, taking into account the requirements and characteristics of datasets and publications created by project members. A final version of the DMP describing how the data created have been managed and shared throughout the project is submitted at the end of the project.

1 Data Summary

To achieve its research objectives, the AIoTwin project plans to use various data sources related to real-world smart IoT use cases, such as (among others) smart city, smart farming and smart building applications. Data will be obtained in one of the following ways: 1) re-use of publicly available data; 2) collection/generation of data through project activities; 3) generation of synthetic data (for scenarios where data collection is not sufficient). The purpose of data collection/generation is to enable AIoTwin researchers to conduct research in the four project-specific research domains and to complete the joint research project outlined in the AIoTwin project proposal.

The AIoTwin project will employ data of different confidentiality levels taking into account the requirements on open science and research data management in Horizon Europe which states that "digital research data generated in the action must be managed responsibly in line with the FAIR (Findable, Accessible, Interoperable, Reusable) principles" while "open access to research data via a trusted repository should be ensured under the principle *as open as possible, as closed as necessary*"[1].

Publicly available data involves the following:

- Publicly available datasets (public): Datasets that are publicly and readily available in bibliographic databases or public research repositories (e.g., the Stanford Large Network Dataset Collection[2], the PapersWithCode platform[3]) and other publicly accessible data, e.g., in media and statistical reports compiled and published by public authorities or other institutions for public use (e.g., Stockholm City Open Data Portal[4], Statistics Sweden[5], etc.).
- Public data that is open for collection (public): This refers to data that is publicly open for collection but has not yet been compiled into an easily accessible data collection. Examples include public data on social networks that is available for retrieval, structural data from open platforms such as P2P applications, etc. The AIoTwin consortium will not collect data containing sensitive personal information. The collection and processing of public data will be in accordance with the GDPR and copyright legislations.
- Research data generated during project activities (public): Data generated within the AIoTwin project in one of the following ways: 1) collection/generation of data from experiments conducted on hardware (e.g., IoT devices) owned and controlled by the partner institutions during project activities; 2) collection of publicly available data enriched with additional data or metadata (e.g. machine learning generated predictions) generated by methods developed within the project.
- Data related to Dissemination and Communication activities (public): the AIoTwin website, public Project Deliverables, presentations/posters, blogposts, video materials, etc. These data will be useful for other research institutions, other professionals in the field and the general public.

Confidential data includes the following:

[1] Annotated Model Grant Agreement: v1.0 DRAFT – 01.04.2023 (https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/aga_en.pdf, p. 285

[2] https://snap.stanford.edu/data/

[3] https://paperswithcode.com/

[4] https://international.stockholm.se/governance/smart-and-connected-city/open-data/

[5] http://www.statistikdatabasen.scb.se/pxweb/en/ssd/

- Project Management Data (confidential): Grant and Consortium Agreements; recruitment data; confidential Project Deliverables; administrative and financial data; detailed training and visit plans; internal meeting minutes, lecture notes, presentations, video recordings and templates; details of innovation management activities and joint preparation of project proposals.

While the nature and characteristics of the data generated by the AIoTwin project will depend on the individual project tasks and activities, the tools and methods required to re-use the data will be provided where possible. Whenever possible, data will be provided in commonly used data formats (see Section 2.3). All members of the consortium are expected to follow best practises in generating, storing, and sharing data.

Research data generated during project activities will be **made open access and licensed under the latest version of CC BY 4.0**. Research data will typically be linked to a scientific publication or project deliverable, and the author(s) will provide (if applicable) the following information to specify whether the publication re-uses existing or creates new data: **data origin/provenance**, **types and formats of data**, **what is it used for within the project** and **to whom might it be useful outside the project**. Table 1 will be used to track information about such datasets.

Research data will be deposited as soon as possible after its generation, following the requirements of the Horizon Europe programme[6], and at the latest, by the end of the project. In particular, data underpinning a scientific publication will be deposited at the latest at the time of publication.

Table 1. Research datasets reused and created in AIoTwin

	Dataset name	Created or reused	Data origin (with URL)	Data formats	Associated publication
1.					

2 FAIR Data: Findable, Accessible, Interoperable and Re-usable

The FAIR Data Management principles have a goal to describe how research outcomes, in particular datasets, should be structured to improve their discoverability and reusability. The European Commission updated the FAIR data guidelines for projects it funds in 2016[7]. The adoption of these principles is driven by the intention to generate research results that will stand the test of time by linking them to metadata that will enable other researchers to use the knowledge, thus improving its reach and transparency.

2.1 Making Data Findable, Including Provisions for Metadata

The partners of AIoTwin jointly drafted and reached consensus on a Consortium Agreement (CA). This agreement serves as a formal document that outlines the interactions between the partners and contains "legal guidelines" to be followed. The CA contains various clauses, including those dealing with research publication procedures, ownership regulations, intellectual property rights (IPR) background, and other relevant issues. This document was signed by all partners in March 2023 and contains various information

[6] https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-rdm

[7] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

on the discoverability of data. This applies to Section 8 (Results), Section 9 (Access Rights) and Section 10 (Non-disclosure of information).

The data published by the AIoTwin consortium will include metadata as well as tools and instruments needed to access or validate the data. **Persistent and unique identifiers** - such as Digital Object Identifiers (DOIs) - will be used to uniquely identify and reference individual publications.

The **metadata** used by the consortium will follow the known practices of labelling, naming, indexing and version control of data, such as ACM Computing Classification System[8], Metadata Object Description Schema (MODS)[9], ISO/IEC 19506 for describing software systems, etc. In all cases, providing search keywords (IoT and AI) will be required to facilitate the possibilities of re-use. The metadata will include the following fields: **author(s)**, **dataset description or abstract**, **date of dataset deposit or publication date**, **dataset deposit venue**, **dataset licence** (CC 0 or CC BY by default) and **dataset embargo period** (if any).

The association of metadata and unique identifiers will be facilitated using well-established platforms for this purpose, such as Zenodo[10], the OpenAIRE repository hosted by CERN.

The **research publications** will be easily findable (i) via the publisher's web site, (ii) via the project's web site, and (iii) via the well-established platforms. For all publications, the attached metadata will include, among others, (i) the terms "European Union (EU)" and "Horizon Europe", (ii) the name of the action, acronym and grant number, (iii) the publication date, and (iv) persistent identifier.

2.2 Making Data Openly Accessible

The publishable data of the project will be made openly accessible (i) through the AIoTwin website, (ii) on well-established repositories which support open access such as Zenodo, Github[11], arXiv[12] and Youtube[13] as well as (iii) repositories of partner institutions.

2.2.1 Documents

Project deliverables categorized as "Public" will be openly accessible on the AIoTwin website under the Results/Deliverables section: https://www.aiotwin.eu/aiotwin/results/deliverables. These publicly available deliverables will be made downloadable from the website once they have been approved by the consortium and submitted to the European Commission (EC) not to delay the publication process, while documents which have received approval from both the EC and external reviewers will be specifically marked as final. Public deliverables will be provided in the widely adopted PDF format.

Confidential deliverables will not be accessible through the website and will remain internal for the Consortium's internal interactions.

Scientific publications will be made readily accessible (open access) in line with the mandatory open science practice in Horizon Europe. We will **pursue Gold Open Access** when this option is optimal and possible for a particular publication in terms of scientific impact and visibility, as well as availability of

[8] https://dl.acm.org/ccs
[9] https://www.loc.gov/standards/mods/
[10] https://zenodo.org/
[11] https://github.com/
[12] https://arxiv.org/
[13] https://www.youtube.com/

funds for publication fees. The Consortium is aware that **immediate open access** to peer-reviewed scientific publications in a **trusted repository** is the main requirement in Horizon Europe. We are currently considering two such repositories, Zenodo and arXiv, to be in line with the copyright policy of publishers most relevant to our field of study (IEEE, ACM, Elsevier, Springer).

We plan to retain the copyright by depositing an accepted version of the publication (**Author Accepted Manuscript** – AAM) in a trusted repository at the latest at publication time if either 1) the publisher allows for the AAM or the edited version to be uploaded on a repository **without embargo period** (needs to be checked for each particular publisher; a useful tool for checking the publishing policies is the Sherpa Romeo web site, https://www.sherpa.ac.uk/romeo/) or 2) by applying the Rights Retention Strategy[14] outlined by cOAlition S, as advised by Open AIRE[15].

The following attribution shall be used in the submitted paper version, AAM and published paper version: "*This work was funded by European Union's Horizon Europe research and innovation programme under grant agreement No 101079214 (AIoTwin project). The contents of this paper reflect the views only of their author(s). The European Commission/Research Executive Agency cannot be held responsible for the information contained. For the purpose of Open Access the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.*"

The peer-reviewed scientific publications will be published under the latest available version of the Creative Commons Attribution International Public Licence (CC BY 4.0). Details on the scientific publication process, target journals, conferences and other data of the scientific publication will be properly registered and reported in WP3 deliverables.

2.2.2 Software repository and documentation

The repository for software developed within the project, mainly in WP1, will be established on GitHub. Documentation of the developed middleware (description, results, tests, usages, validation) will be included in WP1 deliverables of the project.
Supporting tools:

- MS Teams: The tool for managing internal communication in the project, such as chats, groups, conference calls, common project calendar. More information can be found in deliverable D3.7.

2.3 Making Data Interoperable

When publishing data, the AIoTwin consortium will strive to employ the most common file standards, ensuring machine readability in a variety of environments. Depending on the type of data, the following formats will be preferred:

- Text content: plain text ASCII/UTF-8 (.txt), Acrobat PDF/A (.pdf), Microsoft Office and Open Office formats (.docx, .pptx, .odt, .odp)
- Tabular data: comma- or tab-separated values (.csv, .tsv), Microsoft Office and Open Office formats (.xlsx, .ods)
- Structured and semi-structured data: JSON (.json), YAML (.yml) and XML (.xml).
- Graphic content: JPEG (.jpg), PNG (.png), SVG (.svg) and Adobe TIFF (.tif)
- Audio content: AIFF (.aif) and WAVE (.wav)

[14] https://www.coalition-s.org/rights-retention-strategy/
[15] https://www.openaire.eu/how-to-comply-with-horizon-europe-mandate-for-publications

- Video content: MP4 (.mp4) and AVI (.avi)

The project will seek to use standard vocabularies and ontologies for all published data types to enable interdisciplinarity and interoperability. If this is not possible, explanations and mappings to more common ontologies will be provided alongside the data. In doing so, we draw on the expertise of the project partner TUB, the pioneer in the application of Semantic Web technologies in the IoT domain.

2.4 Increasing Data Re-use

Public project outputs, such as public deliverables, presentations and project results, are made available on the project website and can be re-used under the conditions specified in the document licence.

Whenever possible, data will be published under a permissive Creative Commons licence to encourage reuse of the data by third parties. When used, appropriate attribution of the AIoTwin project is required. Any modifications to the original data or results must be clearly stated. The project results will remain accessible as long as the project website is accessible. The published datasets remain usable indefinitely.

3 Allocation of Resources

The implementation of the Data Management Plan in AIoTwin is carried out in Task T5.3 - Data management and protection of personal data. This task started in M1 and will continue until the end of the project. The leader of task T5.3 is RISE, who will also take on the role of AIoTwin's Data Manager (DM) (see next subsection). In addition, all other project partners will also focus on this aspect throughout the project.

The estimated total resources for this task are 6 person-months (PMs) from all Consortium partners (2 from UNIZG-FER and RISE, 1 from TUW and TUB) to ensure adequate allocation of personnel for the following activities: data management, protection planning, data security, documentation management, licencing and information reuse. The long-term preservation of individual datasets will be discussed as part of the implementation of Task T5.3. These discussions will take place later in the project and are particularly relevant to tasks that may generate new datasets, namely T1.4 - Use case definition and experimental evaluation and T3.2 - Planning, preparation and publication of joint publications.

The Consortium affirms that this division seems appropriate to ensure compliance with the EC guidelines on FAIR data management principles.

3.1 Data Manager

The Data Manager (DM) is responsible for overseeing the implementation of the procedures and steps of the Data Management Plan throughout the project. DM oversees the identification of reused datasets and new datasets created as part of the project, the maintenance of the repository referred to in the DMP, and ensures compliance with other clauses of the document.

RISE will assume the role of Data Manager in the project. The personal details of the Data Manager are:

Name: Lodovico Giaretta
Position: Researcher

Institution: RISE Research Institutes of Sweden
Address: Isafjordsgatan 22, 164 40 Kista, Sweden
Telephone: +46 10 228 41 46
Email: lodovico.giaretta@ri.se

4 Data Security

Each partner institution is responsible for ensuring that all data it collects or produces is stored securely and handled in accordance with all European Union data protection laws (e.g. the GDPR) and local legislation. This applies in particular to:

- The **confidential data** (as described in Section 1) will be handled only by the Partner(s) involved using private data management and storage systems that provide granular and tiered access governed by the Partner acting as Controller of the data. The partner acting as Controller will also be responsible for data management, secure storage and deletion of confidential data beyond the lifetime of the project.
- The **publicly available data** (as described in Section 1) will be stored on trusted repositories, the General Assembly of the consortium will decide which repositories will be used for this purpose. At the completion of the project, all the responsibilities concerning long-term data management and secure storage of the publicly available data will fall on the select service for storing this data, based on the decision and agreement established by the consortium's General Assembly.

Sharing of the confidential data within the project occurs through a team collaboration platform (MS Teams) provided by UNIZG-FER. To access the platform, individuals must create a personal username and password. Passwords are encrypted and known only to the individual, ensuring neither RISE nor the platform provider can access them. The UNIZG-FER administrators associate each individual with the project space. Once this association is established, the user gains access to the project space.

When selecting data formats for private and shared documents (datasets managed by the DMP), it is advisable to prioritize open and non-proprietary formats whenever possible. Depending on the purpose, such as analysis, storage, and sharing, both long-term and short-term formats, as well as dissemination and preservation formats, should be utilized.

5 Ethical Aspects

AIoTwin partners must comply with the ethical principles as set out in the Grant Agreement. All activities must be carried out in compliance with:

- all ethical principles, including the highest standards of research integrity – as set out, for instance, in the European Code of Conduct for Research Integrity[16] – including avoiding fabrication, falsification, plagiarism or other research misconduct;
- applicable international, EU and national law (in particular, privacy and IP legislation).

The AIoTwin consortium has considered the ethical implications of its research on data-driven orchestration middleware for edge environments, specifically in the context of smart city or smart building use cases. The consortium intends to utilize existing datasets for experimentation, ensuring that all data is anonymized and not linked to personally identifiable information. However, as the project progresses,

[16] https://allea.org/code-of-conduct/

additional studies may be conducted to gather personal preferences in smart environments, requiring individuals to participate in further experiments. We anticipate that such studies will pose a low risk to personal data as they will not raise particular ethical concerns about human rights, values and freedoms (e.g. human autonomy, privacy and data protection). Data collected in such studies will be anonymised. Participants will be informed about the nature of the data collected and will be required to give consent to data collection before participating in such studies. The choice of modality and location of storage will be according to the rules set by the data owner.

According to the risk-based approach of the AI Act (Proposal for a Regulation laying down harmonised rules on Artificial Intelligence, COM /2021/206 final), AIoTwin's AI-based techniques are not a high-risk solution and can be classified as limited-risk. The consortium is committed to following the human-centred approach promoted by EC and to fulfilling the specific transparency obligations to ensure that users know they are interacting with a machine so that they can make an informed decision about further participation in the studies.

AIoTwin will adhere to ethical standards and relevant international, EU and national laws, including the principles and values enshrined in the EU Charter of Fundamental Rights, the European Convention on Human Rights and the EU Treaties, as well as the Ethics Guidelines for Trustworthy AI and its Assessment List (ALTAI). It will also take into account regulatory reforms currently under development, both those under the European Strategy for Data (such as the Data Governance Act and the Digital Services Act) and those related to AI and intelligent and autonomous systems (the Artificial Intelligence Act). In addition, the consortium will make extensive use of the Assessment List on Trustworthy Artificial Intelligence (ALTAI) to develop procedures for identifying, assessing and managing potential risks and, more generally, to ensure an ethically sound approach to the development, deployment and/or use of AI-based solutions in AIoTwin.

5.1 Responsibilities

Each partner retains exclusive ownership of all data, information, copyrights and other Intellectual Property Rights (IPR) that existed prior to the commencement of this project. The knowledge and pre-existing know-how will be used exclusively for the purposes granted by the access rights.

All sensitive data collected in the framework of the project will be collected, stored and processed in accordance with the General Data Protection Regulation (GDPR) and relevant data protection laws.

Each partner is responsible for managing the data it collects and produces, for maintaining its confidentiality (if applicable) and for depositing publishable data in relevant repositories (e.g. Zenodo) in accordance with the guidelines set out in Section 2.

6 Other Issues

6.1 Intellectual Property Rights

The intellectual property rights relating to the results achieved within the project are set out in the Grant Agreement and the Consortium Agreement. The copyright and intellectual property rights (IPR) in the data generated, collected or used as part of the individual activities within the AIoTwin project belong to the party generating them. The provisions for access by other AIoTwin partners and for cases of shared ownership are set out in the Consortium Agreement.

Intellectual property rights include patents, copyrights, design rights and similar forms of protection, with the exception of confidential information and trade secrets. In line with the aim to release the project components as open source, open source licences such as Apache, MIT or GNU GPL will be used, respecting the partners' foreground and IPR owners' policies. These arrangements were made during the development of the proposal to avoid conflicts.

Confidential data and IPR will be stored securely in a centralised server repository shared between partners for research and testing purposes. Task 5.3 will define and execute strategies to manage IPR and actively monitor the generation of innovation elements (IEs) and potential IPRs. The Data Manager ensures consistency between the datasets and IEs described in the DMP.

7 Acronyms

AAM	Author Accepted Manuscript
CC	Creative Commons
DM	Data Manager
DMP	Data Management Plan
DOI	Digital Object Identifier
FAIR	Findable, Accessible, Interoperable & Reusable
GDPR	General Data Protection Regulation
IE	Innovation Element
IPR	Intellectual Property Rights

8 List of Tables