DATA MANAGEMENT PLAN

DELIVERABLE NUMBER:	D.7.1
DUE DATE:	30.06.2023
DATE OF SUBMISSION:	03.07.2023
NATURE:	R
DISSEMINATION LEVEL:	PU
WORK PACKAGE:	WP7
LEAD BENEFICIARY	KTM

Grant Agreement No 101092850

Funded by the European Union

Document Control Sheet

Deliverable Title:	Data Management Plan
Authors:	Michael Wurzer (KTM)
Contributors:	Dionisios Pnevmatikatos, Konstantinos Nikas (ICCS), Michele Paolino (VOSYS), Foivos ZakkaK (RHAT), Polivios Pratikakis (FORTH), Uwe Dolinsky (CPLAY), Juan Fumero (UNIMAN), Vincent Casillas (SIPEARL), Laurent Eyer (UNIGE), Sergio Saponora (UNIPI), Christos-Alexandros Sarros (UBI)
Reviewers:	Laurent Eyer (UNIGE), Christos-Alexandros Sarros (UBI)
Approved By:	Christos Kotselidis (UNIMAN), Dionisios Pnevmatikatos (ICCS)

Document History

Version	Date	Status	Description/Comments
0.1	16.05.2023	Draft	ToC
0.2	20.06.2023	Draft	First Draft of DMP
0.3	22.06.2023	Draft	Internal Review by KTM
0.4	26.06.2023	Draft	Integrated review comments by UBI
0.5	29.06.2023	Draft	Integrated review comments by UNIGE
1.0	03.07.2023	Final	Final version submitted to EC

DISCLAIMER

AERO has received funding from European Union's Horizon Europe research and innovation programme under Grant Agreement No 101092850. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

TABLE OF CONTENTS

Executive Summary

The document provides the initial data management plan of the AERO project. The deliverable aims to define a framework outlining the AERO policies for data management, sharing, and protection during and after the duration of the project covering topics such as data, metadata content and format, policies for access, sharing and reuse, as well as long-term storage. During the AERO project, the Data Management Plan will be continually assessed and re-evaluated to discover if it has been affected by future results of the work performed in all technical Work Packages. Therefore, the initial framework presented in this deliverable will further evolve during the project lifetime as a living document.

List of Abbreviations & Acronyms

Abbreviation/Acronym	Meaning
CA	Consortium Agreement
DMP	Data Management Plan
DR	Data Release
EC	European Commission
FAIR	Findable, Accessible, Interoperable, and Reusable
GDPR	General Data Protection Regulation
ORDP	Open Research Data Pilot
RIA	Research and Innovation Action
SSL	Secure Socket Layer
UKRI	UK Research and Innovation
WP	Work Package

1 Introduction

The document provides an initial Data Management Plan (DMP) concerning the data processed, generated, and preserved during and after the AERO project. In addition, any topics of discussion regarding data usage, ethics, and security are also discussed in this deliverable. In a nutshell, deliverable D7.1 establishes a framework for the data management policy of the AERO project. Towards this objective, the data, metadata, code, content, and format, sharing policies, storage, and personal data protection measures (if applicable) are entailed. This deliverable will be continually assessed during the project and updated accordingly.

1.1 Background

Deliverable D7.1 - DMP is part of Work Package (WP) 7 "Dissemination, Communication and Exploitation" and reports on the activities concerning Task T7.5 covering the time period from the beginning of the project until milestone M06. It is the first version of the DMP, while updated versions will be included in the Project's annual periodic reports.

1.2 Purpose and Scope

This deliverable defines a data management framework for the AERO project following the "Guidelines on Data Management in Horizon 2020" published by the European Commission (EC) addressing the following questions:

- *What types of data will the Action generate/collect?*
- *What standards will be used?*
- *How will this data be exploited and/or shared/made accessible for verification and reuse?*
- *How will this data be curated and preserved?*

This document will serve as the basis for constant iterations and reassessment throughout the duration of the AERO project. Any changes will be reflected in future versions of this document. These changes regard the whole lifecycle of data that is used within AERO both internally (between the consortium members) and externally (public stakeholders).

1.3 Document Structure

The structure of the document is as follows:

- **Section 1** provides the deliverable's background, purpose and scope, giving its overall structure.
- **Section 2** describes the AERO DMP at a glance, providing insights on the nature of the data that is expected to be used/generated within the AERO project.
- **Section 3** details the processes to be followed for the participation of the project in OperAIRE Open Research Data Pilot (ORDP).
- **Section 4** includes a description of the FAIR principles to be followed for the data used and generated throughout the duration of the project.

- **Section 5** refers to the resources needed for the AERO data collection and management process.
- **Section 6** entails the data security insurance procedures to be followed in AERO.
- **Section 7** comments on the ethical aspects to be considered in AERO during the use of the generated data.
- **Section 8** concludes the document.
- **Appendix I** includes a template to be used in the project as an information notice and consent form.

2 AERO Data Management Plan Overview

Following the FAIR data management guidelines from Horizon2020 - which also apply to the Horizon Europe Framework Program - this deliverable outlines the mechanisms for the vital and important proper data management. This is achieved (and presented) following a methodology that incrementally: 1) identifies the data and artefacts (software, metadata, etc.) used within the AERO project, 2) assesses this data and artefacts in terms of data privacy, sensitivity, and ethical consideration and 3) defines mechanisms for secure data sharing. As a reminder, the collected information included in this tutorial might be augmented during the project (since new data or other artefacts might be added).

2.1 Data Management Plan in Horizon Europe

Following the EC's guidelines during the preparation of the AERO proposal under the "Research and Innovation Action (RIA)", a dedicated section on research data management has been added to the proposal (Section 2 - Impact). For completeness, we also provide the definition of the DMP according to the EC guidelines.

"Data Management Plans (DMPs) are a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. The use of a DMP is required for projects participating in the Open Research Data Pilot (ORDP). Other projects are invited to submit a DMP if relevant for their planned research."

The AERO DMP follows the guidelines on FAIR data management in Horizon 2020 following the templates of both H2020 and Horizon Europe. Following the principles outlined in these guidelines, the management and organisation of data should be based on four fundamental principles which define how research outputs should be processed to be more accessible, interoperable and reusable. This is based on the prerequisite that data must be discoverable (or Findable), Accessible, Interoperable, and Reusable.

Although the EC provides a template for the FAIR principle, no strict technologies, standards, or implementations are imposed, allowing a more flexible implementation of the FAIR principles.

A key characteristic of the AERO project is the reliance and usage of commonly used open-source software systems and frameworks resulting - in most cases - the employment of standardised tools, techniques and data formats.

Regarding the types of data collected in the AERO project, Table 1 provides a summary:

Table 1. Types of Data

Type of Data	Description
Research Data	Research data is the evidence that underpins all research conclusions (except those which are purely theoretical) and includes data that has been collected, observed, generated, created, or obtained from commercial, government or other sources for subsequent analysis and synthesis to produce original research

	results. These results are then used to produce research papers and submitted for publication.
Open Research Data	Openly accessible research data that can typically be accessed, mined, exploited, reproduced, and disseminated, free of charge for the user.
Secondary Data	Secondary data is data that already exists regardless of the research to be conducted.
Metadata	Metadata is data used to describe other data. It summarises basic information about data, which can make finding and working with instances of data easier

The AERO project collected all the information pertaining to this DMP by devising a questionnaire that has been filled in by all project partners. This questionnaire follows the EC's templates regarding FAIR data policies and the received answers summarise (per partner) the following: 1) data summaries, 2) FAIR data, 3) resource allocation, 4) data security, 5) ethical aspects, and 6) other relevant issues regarding data, metadata, and policies.

All the partners of AERO have completed the questionnaire devised by the DMP task leader (KTM) and all answers have been compiled and summarised into this deliverable (D7.1) It is important to know that all partners completed the questionnaires - to the best of their knowledge - up to available data and knowledge to M06. In case new information or data are created after M06 or any information included in this deliverable change, an amendment to this deliverable will be performed. The existing completed questionnaires will serve as the basis for tracking changes per partner. According to the AERO planning, the DMP will be re-evaluated every six months.

2.2 AERO Data Information

In this section, we present the collected information regarding the project's artefact types, formats, access rights and estimated sizes.

2.2.1 Types & Formats of Artefacts Generated/Collected

Table 2 contains the artefact types, descriptions and formats that have been identified within the AERO project.

Table 2 Types of Artefacts

Artefact Type	Description	Indicative Format
Research Item	Vehicle data (KTM Use Case), Gaia DR4/DR5 variable starts catalogue	xls, csv, Gaia Binary Format (GBIN), JSON/BSON, .txt, .docx, .pdf
Software	Source code of all frameworks used and built by partners in the AERO project (mostly open source)	Standard formats used by respective programming languages
Synthetic Dataset	Datasets provided by custom benchmark suites (e.g. Rodinia, FunctionBench)	.csv, .jpg, .avi, .mp4, .json

2.2.2 AERO Artefacts & Access Rights

Based on the conducted survey with the usage of the questionnaires in the AERO project, the concrete artefacts, types and access rights of each partner are listed in Table 3.

Table 3 Table of AERO Artefacts

Partner	Type	Artefact	Publishable/Non-publishable
ICCS	Software	Knative, Kubernetes, containerd, Kata Containers, Firecracker, KVM, the vHive ecosystems, FunctionBench	Publishable
UBI	Software	Kubernetes, CRI-O, Cilium	Publishable
UBI	Software	Maestro	Non-publishable
KTM	Research Item/Data	Synthetic data emulating vehicle information	Non-publishable
KTM	Software	Microservices running the service including Apache Kafka and Databases	Non-publishable
RHAT	Software	OpenJDK, GraalVM, Mandrel, Quarkus	Publishable
FORTH	Software	OpenJDK/TeraHeap, Knot	Publishable
VOSYS	Software	KVM, qemu, libvirt, docker, rust-vmm	Publishable
VOSYS	Software	VOSySMonitor	Non-publishable
UNIGE	Research Item/Data	Based on the Gaia Cycle 4 photometric, spectral, and radial velocity data, we perform a comprehensive analysis of over 2.5 billion sources. Each source has 6-7 time-series associated with an average of 80 observations. With CU7 and DPCG (with Sednai) co-located at UNIGE, the group is responsible for the full processing chain to ingest raw data, construct time-series, curate the data, create ML models, process, including period-search, modelling, specific object studies and analyse and validate the obtained Variables catalogue	Publishable
CPLAY	Software	oneAPI Construction Kit, DPC++	Publishable

2.2.3 Data Sizes

The AERO project is expected to generate research datasets, publications, new service proposals, dissemination material etc., but the expected size of the datasets cannot be currently estimated in high accuracy. A current estimation regarding the use cases, is that data sizes vary from several MBs up to hundreds of TBs (700TB Gaia Data). In any case, the exact data sizes will be reported in the document by the end of the project.

3 Participation to Open Research Data Pilot - OpenAIRE

The open access and reuse of the research data initiated is facilitated by the ORDP – OpenAIRE[1] of the EC. The two main components of the pilot are: a) the development of a DMP and b) the provision of open access to the research data.

A project which participates in the ORDP must follow specific conditions:

- To develop and keep an up-to-date DMP.
- To store the data in a research data repository.
- To make sure that any interested third party can openly access, mine, exploit, reproduce and disseminate the data.
- To include any relevant information and pinpoint - or provide - the tools required to use the raw data to validate the research.

The ORDP is applied:

- To the data/metadata which is required to verify findings in scientific publications.
- To other curated and/or raw data/metadata which have been specified in the DMP

3.1 Publishing Infrastructure for Open Access

The AERO publications infrastructure consists of a process and a number of identified publication platforms that provide long-term open access to all publishable, generated or collected results of the project. Furthermore, the implementation of the project will be done in accordance with the applicable regulations at the national and EU level and, especially, with the General Data Protection Regulation (GDPR) for protection of personal data. This includes also external funding bodies such as the UK Research and Innovation office (UKRI) and the Swiss government which partially fund the AERO project.

The following subsections describe both the publication process and the platforms to be used within the AERO project.

3.1.1 Publishing Process

In AERO we adopted a simple process which decides, in a deterministic way, whether a result of the project must be published. All artefact types created during the AERO project are considered as project “results”, including white papers, scientific publications, and anonymous data. Through this process, each result is classified as public or non-public. A public result must be published as open access, while a non-public result must not be published.

To classify each of the AERO's results, the questions presented below in Table 4 must be answered:

[1] https://www.openaire.eu/

Table 4 Questionnaire for Classification of Artefacts

Artefact Type	Description
Does a result provide significant value to others, or is it necessary to understand a scientific conclusion	If this question is answered with yes, then the result is classified as being public. If this question is answered with no, the result is classified as non-public. Such a result could be code that is very specific to the AERO platform (e.g., driver configuration), which is usually of no scientific interest to anyone, nor does it add any significant contribution.
Does a result include personal information that is not the author's name?	If this question is answered with yes, the result is classified as non-public. Personal information beyond the name must be removed if the result should be published. This also bears witness to the repetitive nature of the publishing process, where results that are deemed in the beginning as non-publishable can become publishable once privacy-related information is removed from them.
Does a result allow the identification of individuals even without the name?	If this question is answered with yes, the result is classified as non-public. Sometimes data inference can be used to superimpose different user data and reveal indirectly a single user's identity. As such, in order to make a result publishable, the included information must be reduced to a level where single individuals cannot be identified. This can be performed using established anonymization techniques to conceal a single user's identity, e.g., abstraction, dummy users, or non-intersecting features.
Does a result include any business or trade secrets of one or more partners of AERO?	If this question is answered with yes, the result is classified as non-public, except if the opposite is explicitly stated by the involved partners. Business or trade secrets need to be removed in accordance with all partners' requirements before the result can be published.
Does a result name technology that is part of an ongoing, project-related patent application?	If this question is answered with yes, then the result is classified as non-public. Of course, results can be published after the patent has been filed.
Can a result be abused for a purpose that is undesired by society in general or contradicts societal norms and AERO's ethics?	If this question is answered with yes, the result is classified as non-public.
Does a result break national security interests for any project partner?	If this question is answered with yes, the result is classified as non-public.

3.1.2 Publishing Platforms

Various platforms are used in AERO to support the open publication of our results. The paragraphs below present the platforms chosen by the project and details their general concepts regarding publishing, storage and backup actions.

Project Website

The AERO project has set up its website, found at: https://aero-project.eu. This website describes the mission and general approach of the project, as well as its development status. Progressively, all developments in the AERO project will be announced on the website via the "Dissemination" webpage which provides access to published deliverables and publications (in pre-camera ready form, or through links to the publisher's websites). All documents will be published using the portable document format (PDF) and downloads are enriched using simple metadata information, such as the title and type of the document (e.g., https://aero-project.eu/publications/). All web page-related data is backed up on a regular basis and the information on the project website can be accessed without creating an account. Finally, the website includes a data protection notice.

Zenodo/Arvix

Zenodo is a research data archive / online repository that enables researchers to share research results in a wide variety of formats for all fields of science. It was created through EC's OpenAIRE+ project and is now hosted at CERN using one of Europe's most reliable hardware infrastructures, with data being backed nightly and replicated to different locations. Zenodo supports the publication of scientific papers or white papers and the publication of any structured research data (e.g., using XML). All uploaded results are structured by the use of metadata, such as the contributors' names, keywords, date, location, kind of document, licence, and others. Considering the language of textual metadata items, English is preferred. All metadata is licensed under CC0 licence (Creative Commons 'No Rights Reserved'). The property rights or ownership of a result does not change by uploading it to Zenodo. The AERO project already has a Zenodo community with all its publications here: https://zenodo.org/communities/aero/?page=1&size=20

arXiv is an open research-sharing platform that currently hosts more than two million scholarly articles in eight subject areas and offers researchers a broad range of services, like article submission, compilation, production, retrieval, search and discovery, web distribution for human readers, and API access for machines, together with content curation and preservation.

It is maintained and operated by Cornell Tech and is widely used in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. AERO will also utilise Arxiv for hosting pre-prints or whitepapers.

Open Access Publications

All publications of the AERO project will follow an open access model. The baseline will be green access and if budget/publisher permitted, a Gold Open Access Model will be used.

Code Repositories

The partners of AERO have agreed to use GitHub as its main source code repository. To that end, an AERO workspace has been created here: https://github.com/AERO-Project-EU

The AERO project at its core builds upon widely available open-source software frameworks (e.g., Quarkus, TornadoVM, KVM, etc.). Therefore, all changes made during the project will be upstreamed to the main repositories. In order to be able to track precisely which code is being generated from the project for quality control and logging purposes, all major repositories have been forked within the AERO Github workspace for staging all changesets.

4 FAIR Data

AERO supports the reuse of research data and follows FAIR principles - a set of guidelines for making data Findable, Accessible, Interoperable, and Reusable; the definitions of which, are provided below:

- **Findable**: The data has a unique, persistent ID, located in a searchable resource, and is supported by meaningful metadata.
- **Accessible**: The data is readily and openly retrievable by utilising common methods and protocols; additionally, the metadata is accessible even when the data is not.
- **Interoperable**: The data is presented in widely accepted standardised formats, vocabularies, and languages.
- **Reusable**: The data has clear licences and accurate, meaningful metadata which conform to relevant community standards and identify its content and provenance.

4.1 Making Data Findable - Provision of Metadata

As agreed by the AERO partners, storage, processing and sharing of data (among project participants) will occur via data exchange platforms, and in particular Google Drive. Data will be accessible even following the end of the project for 2 years. In contrast, interaction with the broader public will be achieved through the official project website.

To make the data findable, a **naming convention** will be used. This convention includes a concise description of contents, the host institution collecting the data and the month of publication.

Version numbering will only be an issue if a participant requests withdrawal of their data, in which case a version number will be added to the filename.

No specific standards or metadata have been identified for the time being regarding the proposed datasets.

For all the AERO use cases, the sensitive data of individuals (if any) to be used by the technical consortium partners to feed their technical software tools will be pseudonymised by the data providers of the consortium. They will be GDPR-compliant, meaning that the data will not identify any individuals, and therefore real names of participants will NOT be distributed.

Data will be shared only in relation to publications (deliverables and papers). Hence, the publication will serve as the main piece of metadata for the shared data. If this is not adequate for the comprehension of the raw data, a report will be shared along with the data explaining their meaning and methods of acquisition.

4.1.1 Discoverability of Data

By considering the FAIR data principles[2], we specify that data/metadata must conform to the following requirements:

- The (meta)data must be assigned a globally unique and persistent identifier.

[2] Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship, 2016.

- Enough metadata must be included, so that the data can be fully interpreted.
- The data must be indexed in a searchable source.

The application of these principles, along with the inclusion of their authentication and authorization details, will result in the data becoming retrievable.

4.1.2 Data Identification Mechanisms

Each project-related document will follow a strict naming convention and will be identified by the project name, along a unique and persistent document type designator and a number provided to the coordinator for submission to the EC. The document version will also be included in the document name and title.

Regarding the project activity and deliverable documents, the relevant task or deliverable numbers will be utilised for document identification, succeeded by a short activity/deliverable title.

An example is given below:

AERO_D7.1_Data-Management-Plan_v.2.0.pdf

4.1.3 Naming Conventions Used

The data generated during the project (datasets, technical reports, deliverables etc.) will be named in a homogenised manner and a version control table will be included. The AERO project document naming recommendations are presented below:

- Easily readable identifier names must be chosen, i.e., brief and meaningful.
- Acronyms with limited acceptance must not be used.
- Abbreviations and/or contractions must not be used.
- Language-specific or non-alphanumeric characters should be avoided.
- A two-digit numeric suffix must be added to identify new document versions.
- Dates follow the YYYYMMDD format. They are stated 'back to front' and use four-digit years.

Naming convention example for deliverables:

AERO_[Deliverable Code from DOA]_[Deliverable name with dashes instead of spaces]_[version]

Naming convention example for reports:

WP[number]_P[Partner Number]_[Description of the activity]

4.1.4 Clear Versioning of Documents

Documents created by the consortium will be versioned. Hence a table will be included in each document outlining the different version numbers by: version, date, status and comments.

Table 5 Revision History Table

Version	Date	Status	Description/Comments
0.1	16.05.2023	Draft	ToC

4.1.5 Standards of Metadata Creation (if applicable)

Any document describing metadata will follow the versioning methodology of deliverables with a number of added information such as: Type, format, identifier, source of collection tools, language, relation to other documents, and access rights.

In addition, a number of mandatory paragraphs will be added: Description of the study or analysis, research methodology, template questionnaires (if used), data documents, user manuals, and other information critical for understanding the metadata.

4.2 Making Data Openly Accessible

When partners declare intent to make data openly accessible following the AERO ethical guidelines, they must notify the project coordinator and then make the data available via the AERO data management repositories - or any other approved means of data sharing.

Regarding scientific publications, as already mentioned in Section 3.1.2, AERO will follow the Green Open Access Model at the minimum conforming to the EC Guidelines on Open Access Scientific Publications and Research Data in Horizon 2020[3] (shown in Figure 1 below).

All papers will be hosted in both the AERO website and the Zenodo AERO community.

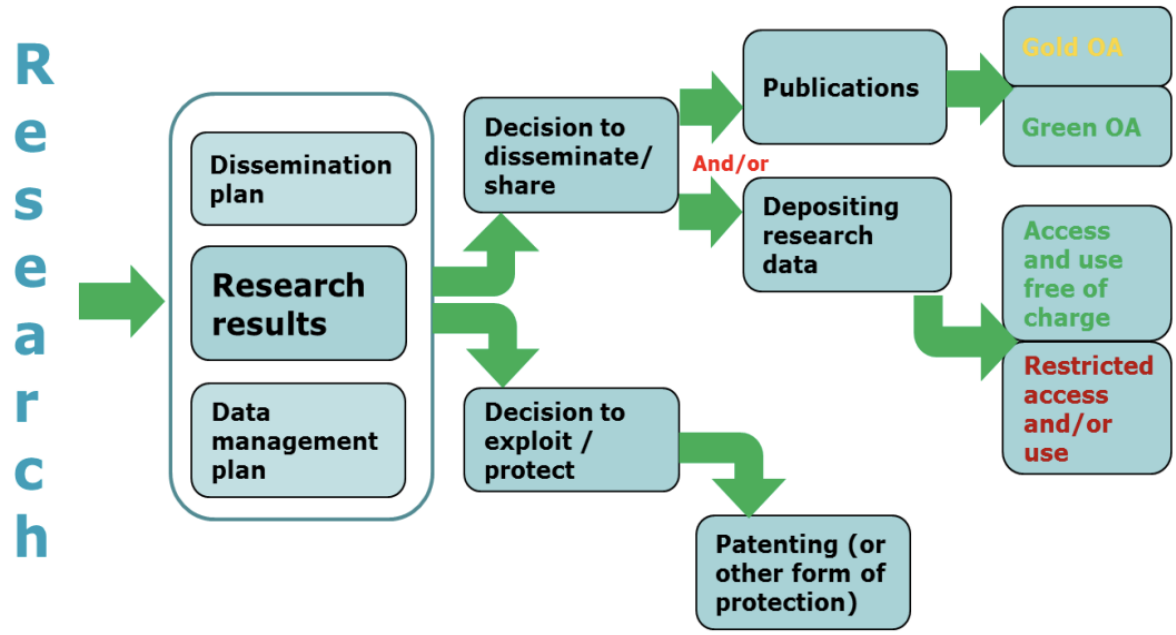

Figure 1 EC proposed strategy for open access publishing[4]

4.3 Making Data Interoperable

The success factors for data interoperability are the standardised naming convention across AERO-generated data and documents and the usage of standard data format. All partners are responsible for satisfying both factors by following the guidelines set in this document. Any deviations shall be

[3] European Commission Directorate-General for Research & Innovation (2017) Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020

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

communicated to the project coordinator, and any decision taken must be reflected to a revised version of the DMP.

4.4 Making Data Reusable

A key success factor for the reusability of data, is the plurality of accurate and relevant attributes accompanying the data. Moreover, it is considered that (meta)data are associated with their provenance and that they meet domain-relevant community standards.

Note that the overall management of knowledge and the provisioning of the Intellectual Property Rights is dictated in detail under AERO Grant Agreement and the Consortium Agreement (CA) stipulating -among others- for the ownership of the background and the foreground knowledge, as well as for the commercial exploitation of the project's results.

4.4.1 Increase Data Reuse through Clarifying Licences

Some of the data will only be available on the website or Google Drive, and their use will be restricted to the research use of the licensee and colleagues on a need-to-know basis. This non-commercial licence shall mention that data may not be copied or distributed and must be referenced if used in publications.

4.4.2 Data Quality Assurance Process

The project coordinator will be in charge of ensuring data quality by guaranteeing that the dataset adheres to the FAIR principles outlined in this plan and that the data is updated.

Personal data will be processed in accordance with EU, and relevant national regulations, as well as the "data quality" principles outlined below:

- Data processing is adequate, relevant, and non-excessive.
- Accurate and kept up to date.
- Processed fairly and lawfully.
- Processed in line with data subjects' rights.
- Processed in a secure manner.

The data quality assurance process will be guided by the REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of April 27, 2016 on the protection of natural persons with respect to the processing of personal data and the free movement of such data.

4.4.3 Length of Time for Data Reusability

The consortium will assist in keeping data reusable after the project is completed, as long as it is feasible. A three-year baseline period has been set; however, this time can be extended with the written agreement of the partners.

4.5 Artefact Template

Appendix I presents the template questionnaire that has been (and will be) used to capture the description of the data produced in the AERO project.

4.6 Data Maturity Models

Data maturity is the measurement of the extent to which an organization is utilizing their data and in general how advanced an organization's data capabilities are. It allows the development and continuous improvement of data governance.

To achieve a high level of data maturity, data must be deeply ingrained in the organization, and be fully incorporated into all decision making and practices.

All the data used by the use case partners of the AERO project (SED, UNIGE, KTM), are used in production by the respective partners following the data maturity models of their organisations. The technologies of the AERO project will utilize this data as part of the executed applications without a direct objective of evolving the data.

Hence, a separate data maturity model in the context of AERO is not considered at this stage since the relevant organisations follow their individual data maturity models.

If this changes during the duration of the project, and a data maturity model specifically for AERO is required, the DMP will be updated accordingly.

5 Resources for Data Collection & Management in AERO

All project's administration data (deliverables, reports, scientific papers, presentations, teleconference minutes, etc.) will be stored at the Google Drive collaboration file repository. That data will be kept for three years after the end of the project and, if requested, for two additional years. It is the coordinator's responsibility to maintain the repository on behalf of the AERO consortium, as well as to handle all project-related data management issues.

The storage and compliance of the research data collected during the project will be part of the Task and WP leader responsibilities. This also includes the responsibility of uploading the data in the AERO Google Drive, when project information needs to be shared.

5.1 Data Management Responsibilities

The responsibility for updating the present document lies at the AERO coordinator, supported by the respective WP leaders and the DMP task leader. This responsibility also includes the development of a relevant strategy to:

- Pinpoint the most appropriate data sharing and data preservation methods.
- Facilitate the efficient use of the data and impose clear rules regarding its accessibility.
- Ensure the high quality of the stored data.
- Guarantee secure data storage.

5.2 Cost of Potential Value of Long-Term Preservation

No extra costs are associated with long-term storage and preservation.

6 Data Security

The AERO project does not involve activities or results that raise security issues, nor "EU-classified information" as background or results.

In addition, for data protection reasons, the AERO data exchange platform (Google Drive) applies the following technical and organisational measures against data acquisition and processing by non-authorized entities, against data processing actions which violate the law and against any change, loss, damage or destruction of the specified data. Those measures include:

- **Information security**: Towards this end, the Secure Socket Layer (SSL) protocol is used in conjunction with the appropriate SSL certificates. The account password is going to exist in the platform solely in an encrypted form, to achieve the desired security level.
- **Options for reading data**: The platform provides the option to render the data available in a read-only or downloadable format, prohibiting unauthorised entities from accessing the information. In addition, invite-only members of the AERO consortium have access to the relevant data.
- **Backup policy**: Auto-backups are performed by Google periodically. Furthermore, each time a document is modified, the previous document version is saved. In addition, local backups are created by the coordinator of the project.
- **Accidental deletion/modification**: Previous dataset versions can be restored if the datasets are deleted, either partially or completely, because of a catastrophic event.
- **Data deletion/modification by users**: Administrators are the only ones allowed to modify or delete any information included in the datasets.
- **Terms and conditions**: All users of the platform have to accept Google's terms of use and conditions.

7 Ethical Aspects

The AERO partners adhere to ethical rules and comply with European legislation on data protection (Regulation (EU) 2016/679 General Data Protection Regulation[5]), the national legislation applicable in countries where the research will be carried out, as well as recommendations and codes of conduct relevant to research activities. The CA has included specific agreements with regard to personal data processing.

The AERO consortium has not identified any additional specific ethical issues related to the activities of the project that are not already addressed in the Grant Agreement. Ethical procedures "Data Protection and Ethics of Research Guidelines" have been specified within the project (and disseminated between consortium members) and these procedures will have to be followed in project activities.

[5] https://ec.europa.eu/info/law/law-topic/data-protection_en

8 Conclusions

The purpose of the DMP is to support the data management life cycle for all data that will be collected, processed, or generated by the AERO project. The DMP is an evolving document during the duration of the project and will be updated at least by the mid-term and final review to fine-tune it to the data generated and the uses identified by the consortium since not all data or potential uses are clear at this stage of the project.

Appendix I

AERO Data Management Plan Questionnaire

*Note: Each data provider should fill in **one questionnaire per dataset**.*

Partner Name	
Corresponding person (email)	
Name of the dataset	
Use Case/Technology Provider	
A – Data Summary	
Are you generating or re-using the dataset?	☐ New ☐ Re-used
	Description of dataset:
What is the type of the described dataset? *The main distinction between data types is between primary data and secondary data. Primary data is data that have been collected for the first time and have not undergone through data processing and/or analysis, yet. Secondary data is data that have been cleaned up, analysed and shared by others (published or unpublished) and they are those that are being typically reused*	*Please Specify:*
	Please provide more details for your answer:
What is the format of the dataset?	*Please Specify:*
Data Schema	*Please Specify:*
What is the expected size of the data?	*Please Specify:*
Why are you collecting/generating or re-using it? What is the relation to the objectives of the project?	*Please Specify:*
	Please Specify:
What is the origin/provenance of the data?	*Please Specify the origin/provenance of the data:*
To whom might the data be useful (data utility)?	*Please Specify:*
Will you apply data aggregation, minimization and anonymization methods to the data?	☐ No ☐ Yes (please specify):
Data Release Roadmap	*Please Specify:*
B- FAIR Data	**Findable, Accessible, Interoperable, Reusable**
B1 - Making data Findable - ***Data are findable when described with metadata and vocabularies in a standardised way, assigned Persistent Identifiers (PIDs) and are registered or indexed in a searchable resource.***	
Will you provide metadata for the described dataset / output?	☐ Yes ☐ No
What type(s) of metadata?	*Please outline what type of metadata will be created and how:*

Do the metadata use standardised vocabularies? *Standardised vocabularies enable greater interoperability across systems. There are generic and discipline-specific metadata standards which have been compiled so as to contain vital information that enable exchange while taking into consideration specificities of work and demands in different disciplines, even in different areas of work within a discipline.*	☐ Yes (please specify): ☐ No (please specify): *Define the metadata standards you use. If there are no standards in your discipline, do you agree to use minimum DataCite metadata standards[6] that Zenodo also uses?*
Are the metadata searchable?	☐ Yes ☐ No *[Searchable metadata are metadata indexed by search engines and are identifiable by web crawlers]*
Are keywords provided in the metadata?	☐ Yes ☐ No *[Outline your approach towards search keywords (e.g. search keywords will be provided when the dataset is uploaded to the repository)]*
B2 - Making data openly accessible - Not all data can be made publicly open, hence data can be FAIR but not open, or open but not FAIR or both FAIR and open. Data is accessible when uploaded in a data repository and retrieved by their PIDs. When data cannot be shared openly, metadata should be provided (even when the data is no longer available). In the case of sensitive or personal data, anonymization or pseudonymization and specific access rights can be applied. Where accessing data requires the use of complementary methods or tools, such procedures should be documented.	
Data	
How is the dataset / output shared?	☐ Open ☐ Shared ☐ Closed *[Indicates access mode for data. If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.]*
What is the reason for limiting access to the dataset / output?	*Please Specify:*
Are there any methods or tools required to access the dataset / output?	*Please Specify:*
Please provide information about the method(s) and tool(s) needed to access the dataset / output.	*Please Specify*
Is documentation about the software needed to access the data included?	☐ Yes ☐ No
Is it possible to include the relevant software (e.g., in open-source code)?	☐ Yes ☐ No (please specify):
Which repository will be used?	*Please Specify:*
Where will the data and associated metadata, documentation and code be deposited?	*Please Specify:*
Is the described dataset / output supported by a data access committee?	☐ Yes ☐ No *[Explain if there is a need for a dedicated data access committee to evaluate/approve access requests to personal/sensitive data, etc.*
Please specify how the dataset / output will be accessed during and after the project ends.	*Please Specify:*
How will the identity of the person accessing the data be ascertained?	*Please Specify*

[6] https://support.datacite.org/docs/datacite-metadata-schema-v44-recommended-and-optional-properties

Please specify how long after the project has ended the dataset / output will be made accessible for?	*Please Specify:*
Metadata	
Will you provide metadata even if the described dataset / output cannot be openly shared?	☐ Yes - Please Specify: ☐ No - Please Specify:
Under which license will metadata be provided?	☐ Creative Commons Public Domain Dedication (CC 0) ☐ Other - Please Specify:
Do metadata provide information about how to access the described dataset / output?	☐ Yes - Please Specify: ☐ No - Please Specify:
Will metadata remain available after the dataset / output is no longer available?	☐ Yes - Please Specify: ☐ No - Please Specify:
B3 - Making data interoperable - Data are interoperable, meaning they can be easily understood and shared with other platforms and systems, when they are created using standard vocabularies and include references to other data and metadata.	
Does your (meta)data use a controlled vocabulary?	☐ Yes - Please Specify: ☐ No *[Controlled vocabularies provide standard terminology as opposed to keywords or tags used to classify information. Examples: taxonomies, ontologies, thesauri.*
If you created the vocabulary, where can it be found?	*Please Specify:*
Have you applied a standard schema for your (meta)data?	☐ Yes - Please Specify: ☐ No *You may browse the Metadata Standards Catalog here: http://rd-alliance.github.io/metadata-directory/*
What is the methodology followed?	*Please Specify:*
B4 - Increase data re-use - Data can be reused when the conditions about how others can make use of the data are well-described following community-standards and are communicated as specified by the owners. Such information can be found in licenses attributed to data and in references about the data provenance.	
What internationally recognised licence will you use for your dataset / output??	*Please Specify:* *[License conditions: Copyright, Creative Commons, Open License, etc. A list of licenses can be found here https://opendefinition.org/licenses/ . Do you agree to use Creative Commons Attribution 4.0 International?]*
What reusability and / or reproducibility methods are followed?	*Please Specify:* *[Examples: Readme files, Codebooks, Data cleaning, Analyses, Variable definitions, Units of measurement, Other]*
Data owner	*Please Specify:* *[List the data owner, the copyright owner, the intellectual property owner]*
Will you provide the described dataset / output in the public domain?	☐ Yes ☐ No
When will the data be made available for re-use?	*Please Specify:* *[E.g., Immediately, after the end of the project (specify the exact time), along with the publication of main results, etc. If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible]*
Do you intend to ensure (re)use by third parties after the project finishes?	☐ Yes (please specify): ☐ No (please specify): *[If the re-use of some data is restricted, explain why]*
How long is it intended that the data remains re-usable?	*Please Specify:* *[Specify the length of time for which the data will remain re-usable, e.g. 5 years after the conclusion of the project]*
Allocation of Resources - Costing of data management includes for example potential use of proprietary services and tools or extra effort needed to perform specific tasks or even to develop tools from scratch.	
What are the costs for making data FAIR in the project?	*Please Specify:* *[Costs might include Storage, Archiving, Re-use, Security, etc.]*

How will these costs be covered?	*Please Specify:* *[Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions). Examples: Use of national infrastructure, Use of institution infrastructure, Infrastructure Grant, Collaboration with other Projects, etc.]*
Identify the people who will be responsible and their role(s) in the management of the described output	*Please Specify:* *[Provide names and responsibilities of researchers' or data managers' data management and stewardship activities that are performed throughout the project for the described output]*
What are the resources for long term preservation?	*Please Specify:* *[Resources: Costs and potential value, who decides and how, what data will be kept and for how long]*
E- Data Security	
Where will the datasets be stored or are already stored?	*Please Specify:*
What security measures are followed? (including data recovery as well as secure storage and transfer of data including personal data)?	*Please Specify:* *[Describe what provisions are in place for data security. Examples: Encryption, IRB protocol, Firewall, Passwords, Hash functions, Physical access control, etc.]*
What conditions do the security measures meet?	*Please Specify:* *[What conditions do the security measures meet? Examples: Data access, Data storage, Data transmission, Data recovery, Data sharing, etc.]*
How will you preserve the described dataset / output in the long term?	*Please Specify:* *[Please describe curation and preservation policies followed for repository content.]*
F- Ethical Aspects	
Are there any ethical or legal issues that can have an impact on sharing the described dataset / output?	☐ Yes - Please Specify: ☐ No - Please Specify: ☐ Unknown - Please Specify:
Does the described dataset / output contain sensitive information?	☐ Yes - Please Specify: ☐ No - Please Specify: ☐ Unknown - Please Specify:
Does the described dataset / output contain personal data?	☐ Yes - Please Specify: ☐ No - Please Specify: ☐ Unknown - Please Specify:
How is the data you process relevant and limited to the purposes of the project? (Please show that what you collect is proportional to the purpose - data minimisation principle)	*Please Specify:*
Will you process "special categories of personal data" (racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, health data, sex life or sexual orientation)? If so - why?	*Please Specify:*
Will you obtain the data subject consent for processing personal data? If not, please provide a valid legal basis for the processing of personal data (e.g. legitimate interest).	*Please Specify:*
Will you process previously collected personal data (secondary use)? If so - please describe the data and justify your right to use data for this project (e.g. consent).	*Please Specify:*
Are you going to transfer/share your data? If so - where and how will you transfer your data?	*Please Specify:*
Will you process any data that cannot be shared? If so - please explain.	*Please Specify:*
What technical and organisational measures will be implemented to safeguard the rights and	*Please Specify:*

freedoms of the data subjects / research participants?	
Please describe your procedures for erasure and deletion of data, if you have them.	*Please Specify:*
Please reference your organisational policy and procedures on personal data management (include national/funder/sectorial/departmental procedures, and ethical standards, for processing of personal data, if any).	*Please Specify:*
Do you make use of other national/funder/sectorial/departmental procedures for data management?	*Please Specify:* *[Indicate if you must adhere to other policies and procedures for data management]*