AI4EUROPE

D1.5

Data Management Plan and Ethics Responsibilities (v1)

Project Title	AI4Europe
Contract Nº.	101070000
Type of Action	Horizon CSA
Topic	HORIZON-CL4-2021-HUMAN-01-02
Project start date	1 July 2022
Duration	42 months

Funded by the European Union

Deliverable title	Data Management Plan and Ethics Responsibilities
Deliverable number	D1.5
Deliverable version	1.0
Contractual date of delivery	30 September 2022
Actual date of delivery	11 November 2022
Nature of deliverable	Report
Dissemination level	Public
Work Package	WP1
Task(s)	T1.4
Partner responsible	ULEI
Author(s)	Alan M. Sears (ULEI), Jan de Bruyne (ULEI), Alejandro Martinez (ITI)

Abstract	This Data Management Plan (DMP) document defines all the procedures to handle the data collected or generated and how they are processed and preserved in the AI4Europe project. In addition to explaining how the project will incorporate FAIR principles, this DMP addresses the GDPR, data security, intellectual property rights, ethical aspects, allocation of responsibility, and other issues.
Keywords	Data Management Plan, DMP, Data life cycle, FAIR data principles, Artificial Intelligence, AI, GDPR, Data security, Ethics, Intellectual property, Platform

Copyright

Funded by
the European Union

Contributors

NAME	ORGANISATION
Alan M. Sears	Leiden University
Alejandro Martinez	Instituto Tecnológico de Informática
Jan de Bruyne	Leiden University

Peer Reviewers

NAME	ORGANISATION
Floor Luub	Eindhoven University of Technology
Tanvir Singh Badwal	University College Cork
Daniel Alonso Román	Instituto Tecnológico de Informática

Revision History

VERSION	DATE	COMMENTS
1.0	26/10/2022	First full version
1.1	11/11/2022	Changes throughout to address comments of reviewers

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Acronyms

Acronym	Open form
AIoD	AI-on-Demand
API	Application Programming Interface
DCAT	Data Catalogue Vocabulary
DMP	Data Management Plan
DPO	Data Protection Officer
DSSC	Data Spaces Support Centre
EC	European Commission
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, and Reusable
GDPR	General Data Protection Regulation
ORDP	Open Research Data Pilot
REST	Representational state transfer

Index of Contents

Index of Figures

Index of Tables

1. Executive Summary

This deliverable D1.5 describes the Data Management Plan (DMP) for AI4Europe and is based on the Horizon Europe Data Management Plan Template document version 1.0 of 5 May 2021. The AI4Europe DMP describes the strategic decisions that have been taken to properly manage the data that will be generated and/or handled by AI4Europe consortium during the project life cycle; this includes data processed within the project consortium and the data processed as part of the digital platform for which AI4Europe is the steward. The DMP also describes all the decisions for making this data Findable, Accessible and Reusable (FAIR). Due to its nature, the DMP is a living document that will be updated periodically according to the progress of project activities.

The main objective of AI4Europe is to support and facilitate a sustainable digital platform and experimentation environment (AI-on-Demand) through the creation of open research channels and mechanisms that foster the European AI research ecosystem, academic and industrial, and that maximises the academic, social, and industrial impact while it seamlessly integrates other projects, platforms, and solutions.

This deliverable is the first version of DMP. The intended audience for this report is internal: the 24 organisations participating in AI4Europe from 15 EU countries. The DMP will establish consistent practices between partners to increase the efficiency and robustness of data handling during the delivery of the project. New versions of the DMP in the form of reports will be submitted as deliverables at M18 and M42.

Starting with a brief illustration of the project, this report describes all the agreed procedures for securely handling and managing the data during the data management lifecycle while leveraging data ‘FAIRification’. This report also includes a summary of ethical, legal, and other possible issues that might emerge during the project development and of which the AI4Europe consortium is aware of and actively addressing.

2. Introduction

This deliverable presents the first version of the Data Management Plan (DMP) of the project AI4Europe. This DMP is part of the task “*T1.4 Project data management, legal and ethics support*”, included in the Work Package “*WP1 Project Management*”.

This first version of the DMP is due to the EC as a deliverable in M03, and the DMP will be refined during the rest of the project lifecycle on a continual basis. New versions of the DMP will be also submitted as deliverables at M18 and M42.

The following document was written using both the HORIZON EUROPE Data Management Plan Template Version 1.0 and Regulation (EU) 2016/679 (GDPR) as main references.

2.1 Objective of the document

The DMP defines all the procedures to handle the data collected or generated and how they are processed and preserved. It describes the approach to making AI4Europe data Findable, Accessible, Interoperable and Reusable (FAIR) by indicating what data will be generated, collected and processed, what standards will be applied, how research data will be preserved and what parts of the datasets will be shared for evaluation purposes and to comply with Open Research Data Pilot (ORDP) requirements. The document will also address ethical issues and some data security principles.

Globally, the DMP has a twofold objective: on the one hand, it offers an overview of the datasets that partners are planning to collect and generate during the project, and the main policies that will be put in place from the partners regarding the management and processing of data. On the other hand, it provides a methodology, guidelines and tools for the partners regarding personal data protection, privacy, and other ethical aspects. This deliverable is a living document and will be updated as the project evolves.

2.2 Project overview

AI scientists and researchers are required to invest a lot of effort to identify trustworthy, high-quality datasets, physical resources, algorithms, and find efficient mechanisms to communicate, cooperate, and engage in an open and transparent manner. A predecessor project, AI4EU, delivered the AI-on-Demand (AIoD) platform with two main tools, namely a content management system enabling access to AI research assets and select pilots and a tool for experimentation.

The main objective of AI4Europe is to, building from AI4EU’s work, support and facilitate a sustainable digital platform and experimentation environment through the creation of open research channels and mechanisms that foster the European AI research ecosystem, academic and industrial, and that maximises the academic, social, and industrial impact while it seamlessly integrates other projects, platforms, and solutions.

AI4Europe will inherit and attract a set of stakeholders and communities from the EC's investment in several projects.

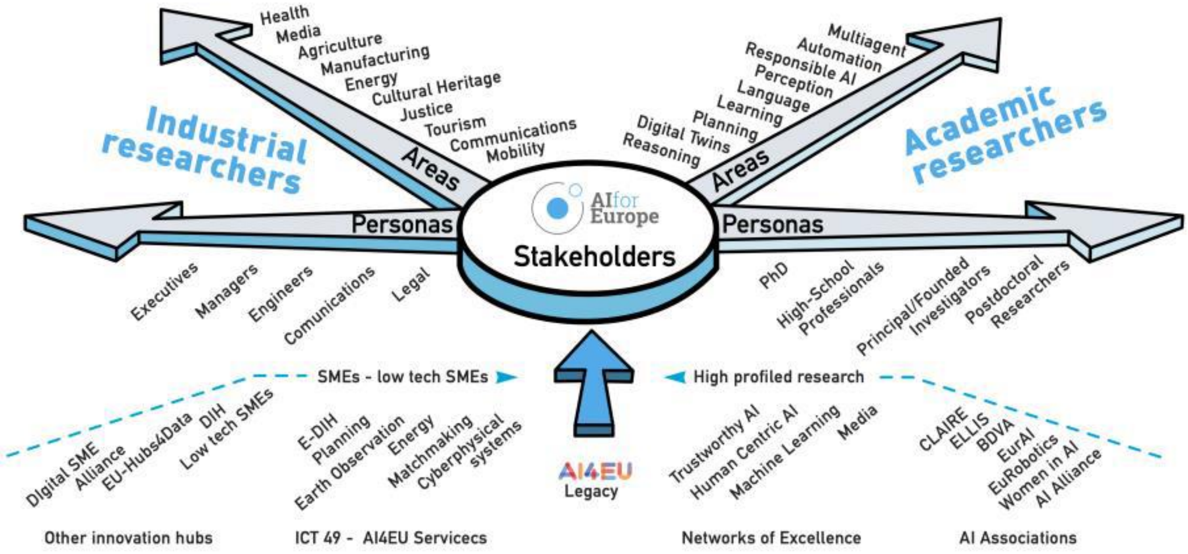


Figure 1. AI Research Personas and Areas

The AI4Europe project consists of:

1. Leveraging existing assets (services, communities) to maximize uptake.
2. Supporting and integrating, where appropriate, ongoing and future outputs from national and European projects.
3. Providing a set of tools that can be easily extended by the community and that ensure both research Excellence and Responsible AI.
4. Interconnecting activities with the AI Data and Robotics PPP (ADR PPP) and the DE AIoD platform to ensure coverage on strategic communities and industrial focus.

2.3 Structure of the document

This deliverable will first explain the methodology of the DMP and project in Section 3. Afterwards, an overview of the data handled by the project is given in Section 4. Standards and principles are discussed in Section 5, including those pertaining to FAIR data, the GDPR, data security, intellectual property rights, and ethics, as well as other developments in regulatory requirements relating to AI and data. The allocation of responsibility in regard to data management in the project is discussed in Section 6. This document provides conclusions in Section 7.

3. Methodology

Understanding dataflows, interactions, and potential issues among the different tasks of the project is key for a successful Data Management Plan. AI4Europe, as the cornerstone project for the success of the AI-on-Demand platform, presents a series of peculiarities which will require this DMP to be unique:

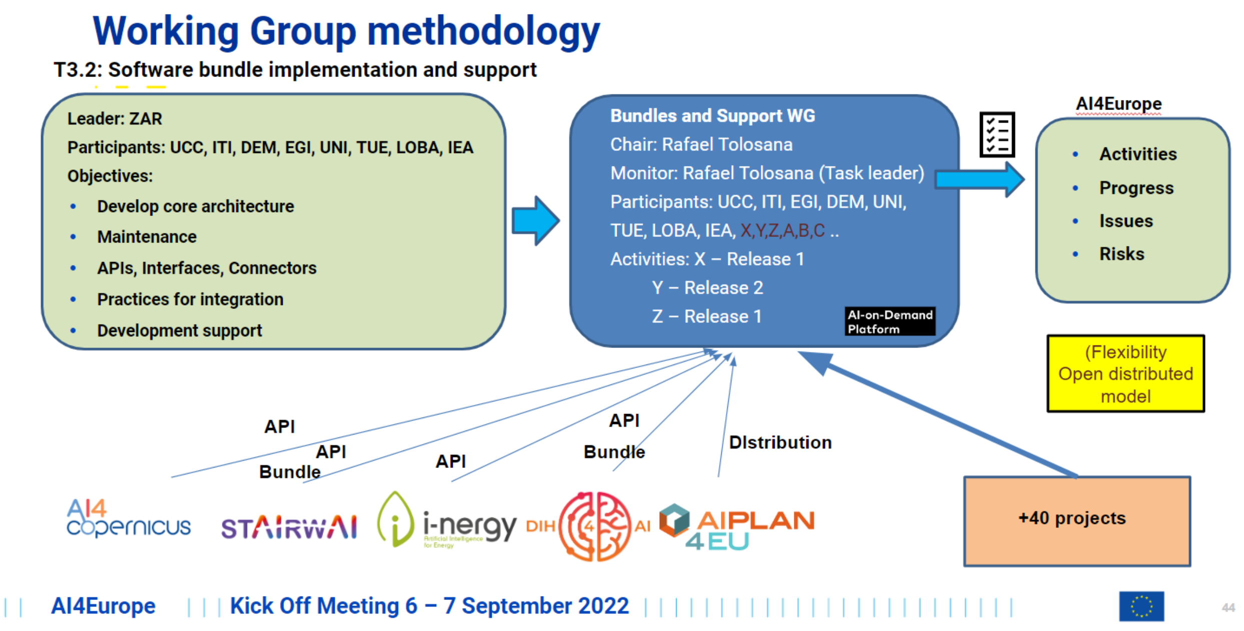
1. The close relation between AI4Europe, the project itself, and the development of the AIoD platform. As the cornerstone project, most data/outputs are expected to be exploited on the platform; the DMP strategy should encourage the release of open access data as much as possible.

2. The open nature of AI4Europe developments. Understanding the Working Groups (WGs) methodology implemented for the project and how partners and external participants are going to interact is pivotal for the success of the DMP.

3. As a complex project where a myriad of stakeholders will interact with project partners and among each other, the DMP should aim to be as flexible as possible to enable collaboration without compromising security (see Figure 2).

It is worthwhile to briefly overview the working groups project methodology to understand the DMP methodology proposed for this first iteration. Originated by the need of involving the community on AI4Europe developments and aligning the project's ethos with the open distributed nature of the Platform, multiple WGs will be spun off and allow external actors to collaborate and influence AI4Europe developments.

An obvious question arises with the formation of working groups, and it's how AI4Europe is going to ensure that developments do not drift from the consortium's vision of the Platform. To ensure that, each WG will be led by a Chair and Monitor board (the latter being involved in AI4Europe's project tasks), ensuring that working groups are completely aligned with task objectives. An example of the multiple links and stakeholders involved in a working group can be found at Figure 2.



Working Group methodology

T3.2: Software bundle implementation and support

Leader: ZAR
Participants: UCC, ITI, DEM, EGI, UNI, TUE, LOBA, IEA
Objectives:

- Develop core architecture
- Maintenance
- APIs, Interfaces, Connectors
- Practices for integration
- Development support

Bundles and Support WG
Chair: Rafael Tolosana
Monitor: Rafael Tolosana (Task leader)
Participants: UCC, ITI, EGI, DEM, UNI, TUE, LOBA, IEA, X,Y,Z,A,B,C ..
Activities: X – Release 1
Y – Release 2
Z – Release 1

AI4Europe

- Activities
- Progress
- Issues
- Risks

(Flexibility Open distributed model)

+40 projects

AI4Europe | Kick Off Meeting 6 – 7 September 2022

44

Figure 2. Working Group methodology example presented at AI4Europe's Kickoff

The DMP's methodology is aligned with the working groups, since we anticipate that they will be the hubs where different participants share data for common purposes. In each WG, both the Chair and the Monitor will be the main focal points between the data managers and the WG. The main actors involved in the Data Management workflow are:

- **Data Managers**: Formed by ULEI (task leaders) and ITI (task co-leaders) personnel. They are responsibilities for the fulfilment of Data Management activities by keeping a track record of data and other valuable outputs being generated, manipulated and/or reused from other sources; ensuring that enough legal policies are implemented to allow a safe exchange of data while respecting the open philosophy of AI4Europe; serving as the consortium's focal point for any data related legal/ethical questions; and promoting the release of Open Access data and outputs to nourish the EU's Open Science ecosystem.
- **WGs Chair and Monitor**: Main leaders of each of the Working Groups, they will be the focal points for Data Management notifications; they ensure that all WG participants are aligned to Data Management policies implemented; and they will periodically check that all WG datasets/outputs documentation is being updated.
- **WG Participants**: All actors (internal and external from AI4Europe) involved in a working group. They will keep track of all datasets and outputs being generated, manipulated and/or reused from other sources to nourish Data Management documentation.

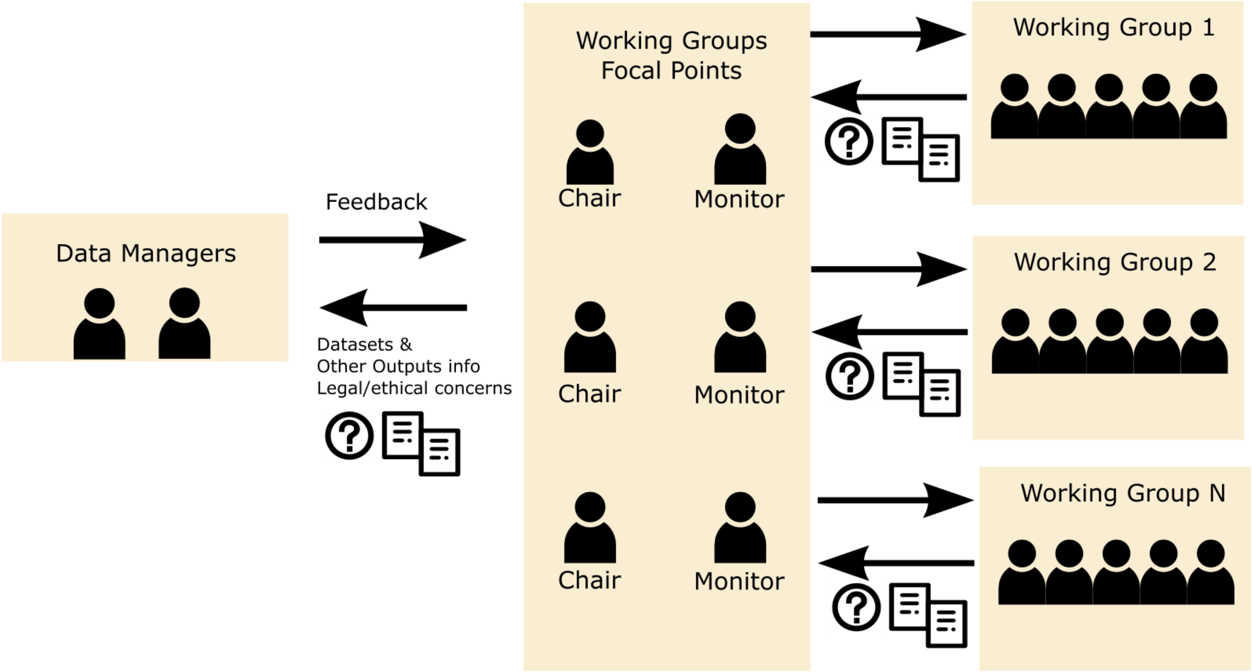


Figure 3. Overview of Workflow

4. Data Summary for AI4Europe

This Data Management Plan outlines the data which will be collected, processed, and/or generated by AI4Europe, a broker for AI resources within the European AI community and interested company and research stakeholders. The goal of the DMP is to ensure that the data is findable, accessible, interoperable, and re-usable (FAIR).

The project will handle different types of data such as technical, open research, and personal data. To get a better understanding of the data that will be used in the project as well as to ensure FAIR data, we sent a survey to the different project partners. This will allow us to gain a more complete overview of the data collection and processing.

The purpose of processing data within the project is to enable the functionalities of the AIoD platform, in order to allow for the efficient dissemination and sharing of AI assets and resources.

Given the fact that AI4Europe is a continuation of the AI4EU project, certain aspects regarding data collection and processing are envisaged. The AIoD platform will collect data regarding AI resources such as datasets, apps, software components, libraries, docker containers, AI models, and services. The collected data is further processed (e.g., review process, meta-data enrichment) and made accessible to the community under defined terms. This collection of AI resources will be useful for the entire European AI community.

Two categories of datasets may be distinguished: I) public datasets or datasets provided by the members of the consortium, II) external data sources, which are integrated into the AI4Europe platform.

In terms of data sources, we anticipate a number of types of data, as outlined:

- User profiles (e.g., anonymised demographic data, click data, etc.)
- User-generated content (e.g., comments, discussion entries, tags, etc.)
- AI resources (e.g., algorithms, source codes, models, containers, etc.)
- Data from the AI4Europe platform
- Data from other platforms

More detailed information about the expected AI4Europe data will be provided in the upcoming versions of the DMP, once the information is available.

5. Standards and Principles

5.1 FAIR Principles

The FAIR principles refer to a concise, domain-independent, high-level and measurable set of guiding principles and practices to apply on a wide range of scientific data or metadata. They are the result of the work in 2014 of a community of stakeholders representing academia, industry, funding agencies, and scholarly publishers, which were then adopted the same year by the European Commission as the data guidelines for the Horizon 2020 (H2020) framework programme. They put "specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals".[1] The term "FAIR" refers to the characteristics of data or metadata of being Findable, Accessible, Interoperable and Reusable. In practice, the elements of the FAIR principles are related, but independent and separable. Any combination of the principles can be applied incrementally. Thus, this modularity of the principles, as well as their distinction between data and metadata, facilitate their support on a wide range of special circumstances. The FAIR principles can also be applied to non-data assets which need to be identified, described, discovered, and reused in the same manner as data. These principles constitute then a general guide to FAIRness of data. But they are not themselves a standard or a specification. Precisely, they precede implementation choices and do not necessarily suggest any specific implementation solution. Instead, they act as a guide for data implementers, publishers and managers to evaluate whether their particular implementation choices are rendering their digital research artifacts FAIR. They form the basis for a long-term care of valuable digital assets composed by the data produced by the research project, while keeping the goal of being discovered and re-used by further research.

5.1.1 Findable data

The first step in (re)using data is to locate them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.[2] To be *Findable*:

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data it describes

F4. (Meta)data are registered or indexed in a searchable resource

[1] Wilkinson, D. et al (2016), The FAIR Guiding Principles for scientific data management and stewardship [Online] Sci Data 3.

[2] See https://www.go-fair.org/fair-principles/.

The project plans to provide an overview of the project website/platform (AI-on-Demand) that outlines available AI resources and their main characteristics. Each resource will be described with a minimum of provided metadata, including information on the data collection methods and processes, the sensibility of the data, usage and access rights, information on each field of the content, and related keywords.

Furthermore, resources will use a naming convention and be annotated with version numbers and timestamps, where the first number of a version changes with a major change in the data collection method (e.g., new survey, new UI features) and the timestamp indicates the date of publishing. The project will be aggregating AI assets from a variety of sources, which may include their own naming conventions and versioning, yet the project will ensure that all indexed AI assets will get clear names, IDs, and versioning.

To support the completeness of metadata, the project will provide a metadata template to stakeholders upon the integration of resources into the AIoD infrastructure.

The template will be a living document that might be expanded to fit project-specific requirements, such as data-related API's needs. The current template, which was inherited from AI4EU, is structured as follows:

Entity		Description
NAME		A name given to the resource.
DESCRIPTION		Text describing the resource e.g., abstract.
HOMEPAGE		URL
VERSION		A version number of the entity. The format is arbitrary chosen by the provider.
VERSION_RELEASE_DATE		The release date of the entity.
DOCUMENT	FILENAME	The name of a file that is added to an AI Resource as complementary information. (E.g., readme, details to data collection, questionnaires, etc.)
	CONTENT_TYPE	Refers to the internet mime-type like (text/csv, audio/mp3, video/h.264)
	STORAGE_LINK_BINARY	Link to the document.
DISTRIBUTION	FILENAME	The name of a distribution of the resource.
	CONTENT_TYPE	Refers to the internet mime-type like (text/csv, audio/mp3, video/h.264)
	EXECUTION_ENVIRONMENT	e.g., Windows, Linux, Mac.
	STORAGE_LINK_BINARY	Link to the distribution.

Entity		Description
AI_TECHNICAL_CATEGORIES		A selection of different technical categories to be assigned to the resource, such as EXPLAINABLE_AI, VERIFIABLE_IA, PHYSICAL_AI, INTEGRATIVE_AI, COLLABORATIVE_AI, ALGORITHM_SELECTION, COMPUTATIONAL_LOGIC, COMPUTER_VISION, CONSTRAINTS_AND_SAT, DECISION_SUPPORT_SYSTEMS, HEURISTIC_SEARCH, KNOWLEDGE_REPRESENTATION, MACHINE_LEARNING, MULTI-AGENT_SYSTEMS, DEEP_LEARNING, PLANNING, NATURAL_LANGUAGE_PROCESSING, SPEECH/AUDIO_PROCESSING, DIALOGUE_PROCESSING, PROBABILISTIC_MODELS, SEMANTIC_WEB, ROBOTICS, REASONING. Any other category can be given as free text.
AI_BUSINESS_CATEGORIES		A selection of different business categories to be assigned to the resource, such as AI_FOR_AGRICULTURE, AI_IN_HEALTH, AI_FOR_CITIZEN_SERVICES_&_EDUCATION, AI_FOR_ROBOTICS, AI_FOR_INDUSTRY_AND_MANUFACTURING, AI_IN_AUTONOMOUS_DRIVING_AND_MOBILITY, AI_FOR_ART_AND_MUSIC, AI_FOR_ENVIRONMENT_AND_SUSTAINABILITY, AI_FOR_IOT, AI_FOR_CYBERSECURITY, AI_FOR_MEDIA, AI_FOR_TELECOMMUNICATION, AI_FOR_FINANCE_&_INSURANCE, AI_FOR_LAW, AI_IN_RETAIL_AND_ECOMMERCE, AI_IN_SOFTWARE_ENGINEERING, AI_IN_HUMAN_RESOURCES, AI_FOR_TRUST_AND_PRIVACY, AI_FOR_AMBIENT_INTELLIGENCE, AI_FOR_SPACE, AI_FOR_AIR_TRAFFIC_MANAGEMENT, AI_FOR_FASHION. Any other category can be given as free text.
LICENSE	CATEGORY	A selection of commonly used licenses, i.e. GPL, APACHE, LOGPL, MIT, COMMERCIAL, or any other license that is not depicted in the categories above can be given as free text.

Entity		Description
LICENSE_URL		A URL explaining the details of a selected license model.
SUPPORT_TYPE		A selection of different support types to be assigned to the resource, such as FREE_SUPPORT, COMMERCIAL_SUPPORT, DISCUSSION_FORUM. Any other category can be given as free text.
SUPPORT_URL		Link to the manner of support. If applicable.
GDPR_COMPLIANT		YES/NO
PROVIDER		Organisation that publishes or provides a resource.
AUTHORS		Creators of the resource.
LINKED_AI_RESOURCE		URI to the resource itself.
KEYWORDS		Freely chosen keywords describing the resource.
LABELS		Will be assigned as part of the review process in the AI4Europe platform. It indicates the specification of the docker container e.g., ACUMOS_READY, BEAT_READY, ELG_READY, MUNDI_READY, BONSEYES_READY, VERYFIED_PROVIDER, OFFICIAL_RESOURCES. RATING Ratings will be collected as
RATING		Ratings will be collected as feedback within the AI4Europe platform, assigned to and stored with the resource.

Table 1. Metadata template for AI4Europe entities

In addition to the metadata described in the table, dataset providers are compelled to attach additional documents such as:

- A description of the study
    - Method of research
    - Applied tools (e.g., questionnaires)
    - A date indicating the termination of the data collection period
    - Language
- Data documentation / usage manual
- Any other information that might be of interest to a data user

While this reflects the current state of the platform, AI4Europe may make changes to metadata system, whether for efficiency reasons or to better accommodate new types of assets or resources that are made available on the platform. Metadata will be provided for datasets, code, models, notebooks, and other AI assets indexed by the platform. We will use data models for these descriptions using the leading standards (e.g. DCAT for datasets). Developing metadata ontologies or vocabularies for other AI assets will be performed as part of WP3 (see the task on data models), and will be aligned with the project ontology and vocabulary. Further, keywords will be used in the metadata so that these textual descriptions can be searched by search engines, in addition to categories and data type information which can be filtered through searches. Finally, this metadata will be held in databases that can be accessed through APIs and indexed by search engines.

5.1.2 Accessible Data

Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation. To be Accessible:

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Project partners and other project stakeholders will collect and provide a range of data. Research outputs by project partners will be stored and made available through an open access repository (e.g. Zenodo) where possible. For the platform, data will also be aggregated from a variety of trusted repositories (e.g. Zenodo, OpenML, HuggingFace, GitHub, etc.); such data will be enriched with metadata that we store ourselves.

Depending on the characteristics of each AI asset or dataset, access and usage rights will be assigned, and access modalities will be documented. Access rights will be assigned per AI asset or dataset, which will be managed through an authentication service based on an open standard (e.g. OAuth). This will be developed within WP4. Along these lines, AI4Europe will provide documentation on and access to all collected assets, either directly or through a link to the host repository, through a standardised access protocol; a REST API will be used to make all indexed AI assets searchable. Future versions of the DMP will further elaborate upon the exact mechanisms to gain access to different categories of AI assets or datasets.

A data access committee is not currently foreseen as needed to evaluate or approve access requests to personal or sensitive data. In instances where the project has indexed datasets for a source in which data access must be requested, it will be described as such in our metadata; however, persons seeking access to such data will still have to obtain access from the source of the data directly.

5.1.3 Interoperable Data

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing. To be interoperable:

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

AI4Europe endeavours to enhance the interoperability of AI assets from various sources. Leading cross-discipline data models (e.g. DCAT for datasets) will be adopted to this end. Metadata will also be enriched to maximise interoperability, and a clearly defined metadata template will be used (see, e.g., Table 1). The exact details (e.g. which data models and which additional metadata will be used) will be fine-tuned in WP3.

Even though the main AIoD repository is expected to be initiated as a central one, the goal is to be extended as a Data Space, leveraging Gaia-X and EOSC interfaces, among others. As such, publicly available data sources for AI4Europe will be identified and made available through the AIoD platform to increase their findability and accessibility.

A common project ontology and vocabulary will be needed to ensure semantic interoperability of all AI assets shared and exchanged among platform users, and the project will leverage the expertise of the Data Spaces Support Centre (DSSC) in realising this goal. This will be developed within WP3. These ontologies may be extended when necessary or new ones may be created where none yet exist. Qualified references to other data will also be utilised; for instance, machine learning experiments will reference the datasets, models, and code involved.

5.1.4 Reusable Data

The ultimate goal of FAIR data is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings. To be reusable:

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

In terms of licensing, the resources can be either categorised under the following five different licenses (GPL, APACHE, LGPL, MIT, COMMERCIAL) or completed with any other officially approved license model provided as free text. As complementary information, the platform expects a link to the description of the license model.

The AIoD platform will not store all AI resources but act as a gateway to access relevant resources and data. Arrangements to meet the access criteria of resources are the responsibility of the resource provider and the resource consumer.

The metadata that is collected when uploading a resource assures its re-usability due to a specified categorisation in terms of Business Category, Technical Category and Distribution Type. Using these categories, specific datasets can be found and retrieved easily. This will be implemented within the AIoD data catalogue, a metadata database which is synced among all participating nodes. In addition, Support type, Rating, and Distribution type can be used to filter available resources. Data catalogue management will be further developed within WP5.

With respect to datasets coming from external platforms, e.g., from the open machine learning platform OpenML.org, the data quality will be assured by the services integrating these data sets into AIoD. In the case of OpenML.org, for example, social indicators such as the number of likes, downloads and runs will be used as a proxy for data quality.

All data will remain reusable for the whole duration of the project. Its reuse after the end of the project will be determined when the sustainability model of the platform is established and the project assets, including data, are transferred. This model will be investigated and developed as part of WP2.

5.2 GDPR

To the extent that AI4Europe will involve personal data collection and processing, it will be subject to the provisions of the GDPR which enshrines the following key principles (without considering exemptions):

- data must be processed fairly, lawfully and only for the purpose for which it was collected and further processed;
- Data cannot be disclosed without authorisation unless there is an overriding act of law or legitimate grounds to do so;
- Subject to certain exemptions, individuals have a right to access the information relating to them and to ask for correction of inaccurate data;
- Information cannot be transferred beyond the European Economic Area boundaries without consent or adoption of other adequate protection measures;
- Organisations are usually required to register or notify the processing of personal data unless the data processing is simplistic, or a data protection officer has been appointed;
- Organisations must have adequate security measures in place

Some activities that the project will engage in will likely involve the processing of personal data. This may include, *inter alia*, personal data collected through the website or through questionnaires, and personal data collected for newsletters or for the organisation of events.

While our position as scientific researchers permit us derogation from the prohibition on processing (sensitive categories of) personal data, we are nevertheless aware that it remains incumbent upon us to provide specific and suitable safeguards so as to protect the fundamental rights and privacy of data subjects. Some of these safeguards are already detailed above. The project further undertakes to ensure that any personal data collected will also be treated in accordance with Article 49 of the GDPR. In particular, personal data collected will be processed fairly and lawfully, and further, personal data collected will be used only for research purposes as specified in our original proposal. The data will be adequate, relevant, and not excessive in relation to the purposes for which they are collected. We will endeavour not to collect, and we will expunge all data that is not directly project related.

5.3 Data Security

During their storage, data, and particularly personal data, should be protected against any type of modification through the implementation of data security principles. Such security principles relate to authentication, accounting, confidentiality, communication security, data integrity and availability. Securing stored digital data involves preventing unauthorised people from accessing it as well as preventing accidental or intentional destruction, infection, or corruption of information. While data encryption is a popular mechanism, it is just one of many techniques and (privacy-preserving) technologies that can be used in implementing a tiered data security strategy. Moreover, Article 25 of the GDPR requires that a data protection by design and by default approach should be taken, which entails that appropriate technical and organisational measures are taken that correspond to the risks to the rights and freedoms of natural persons (e.g., sensitive data requires a higher level of security).

Steps to secure data involve understanding applicable threats, aligning appropriate layers of defence and continual monitoring of activity logs, taking action as needed. The proper method of storage along with levels of access control for privileged users are important considerations for comprehensive protection. Improperly stored information along with overly permissive accounts are a centralised theme in many high-profile breaches. Partners within AI4Europe will follow a specific set of guidelines to comply with the project’s main requirement for the storage of digital data (e.g. data availability must be guaranteed; confidential data must be stored using access protection; strictly confidential information and personal data must only be stored in an encrypted mode).

This section applies to the AIoD platform, and for any assets or resources hosted by the platform. The AIoD platform is the central node in a distributed system, which grants sovereignty over the data that other participating nodes choose to host themselves. In this situation, those participating nodes control the storage and backup of their data, as well as the governance mechanisms that control access to their data.

5.3.1 Storage and Backup

Data corresponding to the current AIoD platform is stored within the IMT TeraLab[3] infrastructure. TeraLab ensures that backups are stored on dissociated physical servers. Backups are incremental and their perimeter is specified in accordance with the project representative.

Protocol	Technology & usage
SQL	**The Content Management Component (CM) (Drupal) and the Identification & Authentication Management (IAM) WSO2 are both using MariaDB to store information.**
NoSQL	**The SEARCH component indices and the metadata of the AI Resources Catalogue are stored using an ElasticSearch Index**
Distributed File System	**Backups, Logs and AI Resources attachment from the AI Resources Catalogue are stored using Block and object CephFS. Distributed file system allows storage extension without unplanned service disruptions**

Table 2. Backup protocols and their usage in relation to technologies used

The backup strategy was built according to a two-level process. First, each component runs on-premises data backup to package and copy data to a specific location on a different drive at specified intervals. Then within the infrastructure, a remote backup is performed to copy to a different Backup Virtual Machine.

As AI4Europe takes stewardship of the AIoD platform, this backup strategy might change according to the developing needs of the platform.

5.3.2 Access and Security

Each project partner gets individual access to the main data repository. The research partners will get unrestricted read-access to the data. Commercial partners will get access based on project needs. Data access will be given to named users. Relevant transactions will be logged within the system. The main data repository is to be located in AI4Europe’s infrastructure, utilizing the security measurements in place.

The personal data that is collected through the website or through questionnaires, as well as the personal data collected for newsletters or for the organisation of events, will

[3] Teralab is a secure and sovereign platform aiming at facilitating the realization of projects of innovation, research and education around the Data field (Big Data, and Artificial Intelligence in particular) (see https://www.teralab-datascience.fr). The ambition is to allow major institutions and companies, SMEs and start-ups, researchers, teachers, and students to work together in this area.

not be stored on the main data repository. Only the consortium partners which need access for their activities will have access to such data.

5.4 Intellectual Property Rights

All digital research data generated by the project and made available through the repository will be made available with the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a licence with equivalent rights, unless it would be against the beneficiary's legitimate interest (including commercial exploitation), or be contrary to other constraints (e.g. EU competitive interests). Where such data is not open access, it will be noted and justified within the present document, the Data Management Plan. The metadata of deposited data will be open under a Creative Common Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded), in line with the FAIR principles, and include reference to Horizon Europe funding.

In terms of licensing data or resources provided by external parties through the AIoD platform, the resources can be either categorised under the following five different licenses (GPL, APACHE, LGPL, MIT, COMMERCIAL), or completed with any other officially approved license model provided as free text. As complementary information, the platform expects a link to the description of the license model.

5.5 Ethical Aspects

The guiding principles at the heart of the AI4Europe project are the highest ethical standards, the protection of privacy and the validity of data and its accurate representation. In adhering to these principles and remaining cognisant of concerns that arise in the Work Plan, the project will take the following steps, in addition to those detailed above, towards addressing these:

- Compliance with the Assessment List for Trustworthy Artificial Intelligence (ALTAI), which has been developed by the High Level Expert Group on AI (AI HLEG);
- The availability of partners with Ethics, Privacy, and Legal expertise to all project staff members at the outset of the project and throughout the duration of the project;
- Assurance that Privacy by Design, Ethics, Legal and Societal Impact requirements are included as research and development mandates integrated into the project plan in compliance with GDPR Article 25 (data protection by design and default).

While we recognise the numerous benefits of AI research that can positively affect efficiency, productivity, and more broadly the flourishing of humanity, we are also aware of the detriments that can occur when, for example, biased datasets are used (e.g. in predictive policing). When we become aware of such issues hosted on the platform, we will take remedial measures for the affected asset or resource, which may ultimately result in its removal. A more detailed process for handling ethical issues in the project will be further elaborated upon in the next version of the DMP.

5.6 AI and Data Related Evolutions

The AI4Europe project will of course follow-up and implement where necessary ongoing and relevant evolutions with regard to the AI Act and Data (Governance) Act. Although these initiatives are still ongoing, the project will endeavour to incorporate their mandates at an early stage of the project.

Of particular relevance, the AI Act will categorise different forms of AI according to risk, which will entail certain obligations. For example, those in the unacceptable risk category are prohibited, whereas those with less risk have only transparency obligations (see Figure 4).

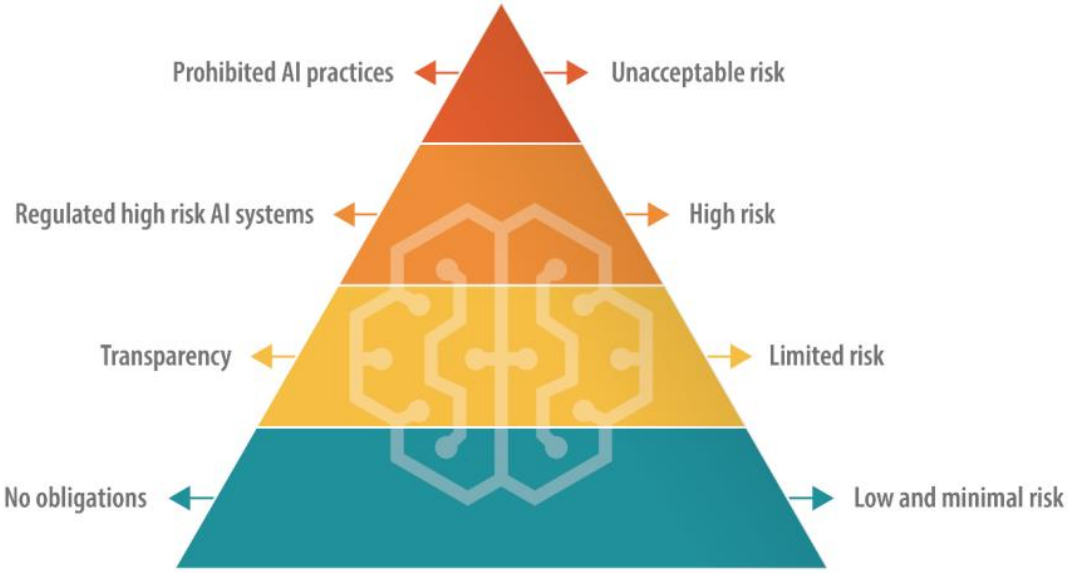


Figure 4. Risk assessment and related obligations under the AI Act

AI4Europe will incorporate a risk assessment procedure into the AIoD platform. This will be further detailed in the next version of the DMP.

6. Allocation of resources

6.1 Costs for making AI4Europe data FAIR

The costs for making data FAIR are estimated to be relatively small. The costs for storage of data on the AIoD platform are included in the project costs. External parties also provide data and assets which are made available on the platform but they host the underlying data; in this case those parties bear the costs of storage. The distributed metadata database is lightweight and the cost for hosting the file is born by each participating node.

6.2 Responsibilities for data management in the project

Each project member is responsible for ensuring that the data generated or collected within complies with the principles and standards set out in this DMP, including that the data is FAIR. ULEI leads and ITI co-leads the data management task and will aid the other partners in complying with the standards set out within the DMP and FAIR principles. Alan M. Sears will serve as the primary point of contact for any questions or concerns regarding data management in the project, as well as serve as the Data Protection Officer, until it is deemed necessary to appoint a separate person. The leaders of the data management task will also track the data that each partner is generating or collecting—updating the DMP as necessary—and categorise any research data generated as open access or closed access, identifying where open access materials may be located.

The implementation of the DMP will be monitored by UCC, as the Project Coordinator, and reported on a regular basis to the Executive Board, as well as to the EC through periodic reporting procedures. Furthermore, consortium partners have the responsibility to make sure their activities are in line with all applicable local, government and international laws, regulations, and guidelines, some of which are detailed in this document.

For the main data repository hosted as part of the AIoD platform, the host of the platform manages the access rights to the data—including the monitoring and controlling of access to the data repository—as well as other relevant data security measures. The details of the relevant mechanisms will be developed as part of WP3 and WP5, and this will be further elaborated upon in future versions of the DMP.

7. Conclusions

This report describes the Data Management Plan for the data which is collected, generated and/or processed within the AI4Europe project. It is important to note that the Data Management Plan is a living document and will be constantly updated until the end of the AI4Europe project.

7.1 Summary

AI4Europe will support a sustainable digital platform, AI-on-Demand, that facilitates the sharing of AI assets and resources that fosters the European AI research ecosystem. The platform will integrate other projects, platforms, and solutions.

The first section is the Executive Summary for the Data Management Plan. The second section gives an introduction to both the DMP and the project itself, as well as provides the structure for this document.

The third section outlines the methodology of both the DMP and the project, while the fourth section provides an overview of the data managed within the project and the general approach taken for data management within the project.

The fifth section discusses the different principles and standards that are applicable to the project. It explains the goal of making data FAIR (findable, accessible, interoperable, and reusable), and how this is achieved within the project. It also addresses applicable legal standards such as the GDPR and those pertaining to intellectual property rights, as well as standards for data security. Ethical aspects and issues are also discussed, and developing standards pertaining to AI are envisioned and taken into account.

Finally, the sixth section explains the allocation of resources in the project, including the costs for making data FAIR and the responsibilities for data management, before concluding in the present section.

7.2 Tentative roadmap for future versions of the DMP

The present Data Management Plan v1 gives a first overview of the data processed in AI4Europe and describes the data categories of the project, providing information on their management and the implementation of the FAIR principles in the project, as well as outlining data security, ethical aspects, and intellectual property rights issues.

In the next version of the Data Management Plan (D1.6, v2), more details will be provided. This is particularly true regarding the exact nature and types of data that are envisioned to be collected and used, their format, size, and utility, This also holds for the metadata framework (to be further developed and refined in WP3), and the AI-on-Demand data catalogue (to be developed in WP5), as well as the corresponding standards and methodologies used for both of these. Consideration will also be given to further ethical issues that have arisen and a mechanism for handling them, as well as the continued hosting of the AIoD platform for which we do not yet have subsequent information at this stage of the project.

The final version of the DMP (D1.7, v3) will provide a comprehensive overview of the data generated, collected, and/or processed by the project, its lifecycle, and how it complies with FAIR and ethics principles, among other standards.

8. Appendix

This is the questionnaire that was sent to multiple partners in order to gain a better understanding of data usage in the project.

Questions

FAIR data

1. Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

What naming conventions do you follow? Do you provide clear version numbers?

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Will metadata be offered in such a way that it can be harvested and indexed?

2. Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Data:

Will all data be made openly available? 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.

Will patents or other protection of intellectual property be sought?

Will the data be accessible through a free and standardized access protocol?

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project? How will the identity of the person accessing the data be ascertained?

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

3. Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

4. Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Will the provenance of the data be thoroughly documented using the appropriate standards?

Describe all relevant data quality assurance processes.

------End of Document------

Consortium

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

ÖREBRO UNIVERSITY

ITI INVESTIGATE TO INNOVATE

KINIT

telenor

Universidad Zaragoza
1542

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Fraunhofer
IAIS

eit Digital

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

ARISTOTLE UNIVERSITY OF THESSALONIKI

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Universiteit Leiden
eLaw

egi

THE ARCTIC UNIVERSITY OF NORWAY • UIT •

European DIGITAL SME Alliance

Université Paris Cité

IeAT

LOBA®

FGS

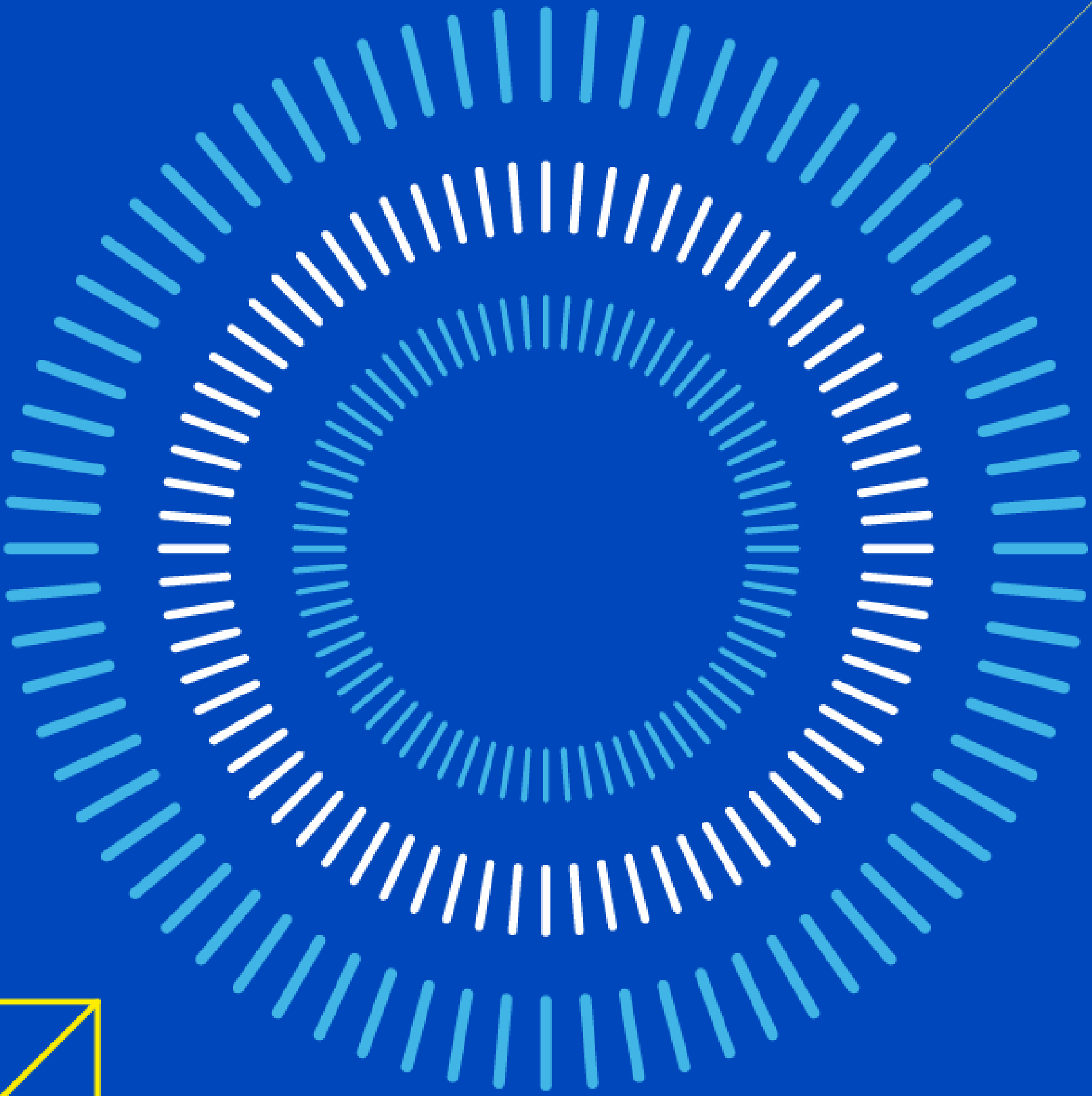
ΔΗΜΟΚΡΙΤΟΣ
DEMOKRITOS

A!
Aalto University

ZRC SAZU

AI4EUROPE

www.ai4europe.eu
info@ai4europe.eu



Funded by
the European Union