



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101135437



AI services for COntinuous trust
in emerging Digital Environments

D1.2 – Data Management Plan

Grant Agreement no.:	101135437
Project Title:	AI-CODE - AI services for COntinuous trust in emerging Digital Environments
Contractual Due date:	31/05/2024
Actual submission date:	10/06/2024
Lead beneficiary:	DS Tech

Disclaimer

“Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.”

Copyright

Copyright © 2024 the AI-CODE Consortium. All rights reserved.

Consortium composition

N	Role	Short name	Legal Name	Country
1	COO	DS TECH SRL	DS TECH SRL	IT
2	BEN	CERTH	ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS	EL
3	BEN	FBK	FONDAZIONE BRUNO KESSLER	IT
4	BEN	UPM	UNIVERSIDAD POLITECNICA DE MADRID	ES
5	BEN	ATC A.E.	ATHENS TECHNOLOGY CENTER ANONYMI VIOMICHANIKI EMPORIKI KAI TECHNIKI ETAIREIA EFARMOGON YPSILIS TECHNOLOGIAS	EL
6	BEN	DW	DEUTSCHE WELLE	DE
7	BEN	KInIT	KEMPELENOV INSTITUT INTELIGENTNYCH TECHNOLOGII	SK
8	BEN	Debunk.org	DEBUNK EU	LT
9	BEN	EIPCM	EUROPEAN INSTITUTE FOR PARTICIPATORY MEDIA EV	DE
10	BEN	SRU	STICHTING RADBOUD UNIVERSITEIT	NL
11	BEN	EURACTIV	EURACTIV MEDIA NETWORK B.V	NL
12	BEN	CEPS	CENTRE FOR EUROPEAN POLICY STUDIES	BE

Deliverable number	D1.2
Deliverable Title:	Data Management Plan
Description of the deliverable:	This deliverable will describe how the AI-CODE project manages data throughout its life cycle, in order to be compliant with the regulatory framework.
Type/Nature	☒ R - Document, report ☐ DEM - <i>Demonstrator, pilot, prototype</i> ☐ DEC - <i>Websites, patent filings, videos, etc.</i> ☐ OTHER ☐ ETHICS – <i>Ethics requirement</i> ☐ ORDP – <i>Open Research Data Pilot</i>
Dissemination Level	☒ Public – fully open ☐ Sensitive — limited under the conditions of the Grant Agreement ☐ EU classified — RESTREINT-UE/EU-RESTRICTED, CONFIDENTIEL-UE/EU-CONFIDENTIAL, SECRET-UE/EU-SECRET under Decision 2015/444. For items classified under other rules (e.g. national or international organization), please select the equivalent EU classification level.
Work Package number	WP1
Work Package leader	DS Tech
Author(s) in alphabetical order	Michela Mattaloni (DS Tech) Paula Gurtler (CEPS)
Internal Reviewers in alphabetical order	Appointed Internal Reviewers Andreas Griefer (DW) Ivan Srba (KInIT) Roberto Moro (KInIT) Other reviewers Giulio Spinozzi (DS Tech) Manos Schinas (CERTH) Mathias Becker (EIPCM) Symeon Papadopoulos (CERTH)

Authors

Name	Partner	e-mail
Michela Mattaloni	DS Tech	m.mattaloni@dstech.it
Paula Gurtler	CEPS	paula.gurtler@ceps.eu

History of changes

Version	Date	Changes
0.1	20th May 2024	First version
0.2	24 th May 2024	Revision by internal reviewers
0.3	5 th June 2024	Version integrating reviews from internal reviewers
1.0	10 th June 2024	Final version

Table of contents

1	Executive Summary.....	6
2	Introduction	6
3	Definitions	7
4	Data Summary	8
4.1	FAIR Data Management	11
4.1.1	Making data findable, including provisions for metadata	11
4.2	Making data openly accessible.....	11
4.3	Making data interoperable.....	12
4.4	Re-use of data.....	13
5	Methodology.....	13
5.1	Data acquisition.....	13
5.2	Data storage and backups.....	13
5.3	Data categories and taxonomies.....	13
6	Data security	14
6.1	Provision for Raw Data.....	14
6.2	Provisions for Internal Research Data.....	14
7	Applicable Legislation	15
7.1	General Data Protection Regulation (GDPR)	15
7.1.1	Data protection impact assessments (DPIA).....	15
7.1.2	Informed Consent	16
7.1.3	Privacy by design.....	16
7.1.4	Due Diligence Documentation.....	16
7.2	Fundamental Rights	16
7.3	Ethical Frameworks	16
7.3.1	Due Diligence Documentation.....	16
8	Allocation of resources.....	17
9	Conclusions	17
10	Annex 1: Data Summary form.....	18

List of Tables

Table 1: Preliminary list of data categories	9
--	---

Table of acronyms

Acronym	Explanation
AI-CODE	AI services for COntinuous trust in emerging Digital Environments
ALTAI	Assessment List for Trustworthy AI
DMP	Data Management Plan
DPIA	Data Protection Impact Assessment

FAIR	Findable, accessible, interoperable and reusable
GA	Grant Agreement
GDPR	General Data Protection Regulation
CA	Consortium Agreement
WP	Work Package

1 Executive Summary

This document is the Deliverable 1.2 “Data Management Plan” (DMP) of the AI-CODE project. This deliverable relates to activities in WP1 “Management” under task 1.3 “Data Management Plan and IPR issues” which overall objective deals with data management as an overall issue. The aim of this task is to deal with the definition and sharing of policies and the identification of actions for data management and associated software produced as a result of the project.

The intended audience for the presented document is the AI-CODE consortium, the European Commission and the project’s main stakeholders and services’ users.

The DMP will describe the data management cycle for all the datasets to be processed under the project and will cover:

- the purpose of data processed and its relation to the objectives of the project;
- the types and formats of data that will be collected, processed and/or generated by the project;
- the re-use of existing data and the procedures for the re-use;
- the methodology and standards to be applied to the data management;
- the handling of data during and after the project.

The document describes and presents the following content:

- definitions,
- summary of the data categories processed under the project,
- the FAIR data principles applied to the project,
- the methodology used to create the DMP plans regarding the data processed within the project, including data acquisition, data storage and backup, and data categories and taxonomies,
- the provisions adopted for data security standards,
- the applicable legislation,
- the provisions and adopted measures for the processing of personal information in the project implementation,
- a discussion on the resources responsible for data management and for the compliance to the FAIR principles and
- draws conclusions and sets future goals.

Following the EU’s guidelines regarding the DMP, this document may be updated, if appropriate, during the project lifetime.

2 Introduction

A Data Management Plan is important for the AI-CODE project since the project involves:

- 1) **collecting data** describing users and their communication patterns in next-generation social media platforms (**WP2**) using public APIs (Application Programming Interfaces) as well as collecting data describing the content of messages, such as text, images and videos (**WP4-WP7**). A DMP helps in outlining the sources of data and how it will be collected, stored, and managed throughout the project.

This ensures transparency and accountability in data collection practices.

- 2) Activities such as reconstructing social networks, characterizing risks associated with disinformation spreaders, and analyzing **user requirements (WP3)**, **user profiles and user input, text, videos and images (WP4-WP7)** necessitate detailing methodologies, algorithms, and tools that will be used for data analysis, ensuring reproducibility and reliability of results.
- 3) **User and stakeholder engagement** is solicited throughout the duration of the project to collect feedback through co-creation sessions (WP3) and **validation activities (WP8)** to iteratively design, improve and develop the prototypes of the services. The DMP provides a clear overview of the project's data management practices, making it easier for team members, collaborators, and stakeholders to understand and contribute effectively.
- 4) **Collecting and re-using data** from existing datasets, besides collecting new datasets, to train methods, seed content into services and other related activities (**WP4-WP7**).
- 5) Given the sensitive nature of the data being collected, particularly in analysing social media behaviour, ensuring **user privacy** and **adhering to ethical guidelines** is paramount. The DMP outlines the measures taken to protect user privacy, obtain necessary permissions or consent, and mitigate any potential ethical concerns associated with data usage.
- 6) The project may generate valuable datasets and insights that could benefit other researchers or organizations. The DMP includes provisions for **data sharing**, outlining the conditions under which data will be made accessible to other researchers, ensuring **proper attribution**, and **promoting the re-use of data** beyond the project's scope.

3 Definitions

Data Management Plan: A formal working document which outlines how datasets will be handled both during the active project phase and its conclusion.

Data Owners: the individuals or groups of individuals who are held accountable for a dataset and who have legal ownership rights to a dataset even though that dataset may have been collected, collated or disseminated by another party.

Dataset: a collection of data created under the AI-CODE activities, events, services or secondary data which is published with due permission of the user. This includes also datasets already existing outside the AI-CODE project that will be used to train models, seed content into services and other related activities. Datasets are a fundamental unit of data management.

Legal Account: a location on an AI-CODE server used to store information from a legal entity.

Personal Account: a location on an AI-CODE server used to store information from a personal user.

Metadata: Information about datasets stored in a repository. Metadata is the term used to describe the summary information or characteristics of a set of data. For example, with respect to geographic information this normally means the What, Who, Where, When and How of the data.

Newly collected data: means original data collected by project partners and/or uploaded by users especially for the purpose of the project.

Re-used data: means data that has been captured for another purpose than the purpose in mind. Secondary data that is being reused, usually in a different context.

Publicly available data means information publicly available, which can be freely accessed, used and reused.

User: a person who uses AI-CODE services and is responsible for the creation of a Personal Account which may differ from the legal account.

4 Data Summary

The main purpose of the data captured and generated by the AI-CODE project is to understand rapid developments in next-generation social media, gaining insights into the workings of generative AI and its potential impact on information dissemination. Another purpose is to develop AI-based tools and services tailored to support media professionals to navigate emerging digital environments effectively and credibly. Data capture further facilitates the identification of new forms of content manipulation enabled by generative AI, as well as the development of algorithms and methodologies for detecting and mitigating disinformation and misinformation spread through these platforms.

In summary, reusing existing datasets, as well as capturing and generating new data in the AI-CODE project serves the overarching goal of empowering media professionals with innovative AI-based solutions to combat disinformation, ensuring media freedom and pluralism, and delivering credible and truthful information in rapidly evolving digital environments. **AI-CODE services** are designed to actively **assist media professionals** in detecting forged content, disinformation and multiple deepfake and in determining reliability of information. Services will **support media professionals** in producing reliable information thanks to a better understanding of how generative AI works, its capabilities and risks and how to best use it to analyse and interpret data to combat disinformation.

The AI-CODE project will produce several datasets during the lifetime of the project. The datasets are categorized by the type of origin of their data. There are two types of origin of data identified in the datasets: **Newly collected data** means original data collected by the project partners for a specific purpose. **Re-used data** means data that has been collected within other activities, within the project or not. It may also contain publicly available data, which can be freely accessed and is public domain. Also, the data processed under the AI-CODE project can generally be divided into two categories: “**non-personal data**” and “**personal data**”.

The second category, “Personal Data” comprehends the results from the stakeholders’ engagement activities, more specifically contact details from the participation in workshops, seminars, educational activities, the results of questionnaires. Other Personal Data may be collected in the process of analysing and collecting profiles and requirements from users of social networks. These types of data will follow the provisions explained in the par. 6 Data security.

Regarding the formats of datasets, the following categorisation will be used:

- Text: plain text (TXT), HTML, XML, PDF/A.
- Databases: XML, CSV, sqlite3.
- Image: JPEG, JPG-2000, PNG, TIFF.
- Audio: AIFF, WAVE, .mp4, .mp3.
- Containers: TAR, GZIP, ZIP.
- Video: mp4, AVI., WMV., AVCHD., WebM., FLV.
- Other.

A preliminary overview of data categories anticipated in the AI-CODE project is listed in Table 1 below. The list is indicative and allows estimating the data that AI-CODE will process. This list may be adapted (by adding or removing data categories) as well as supplemented with a list of specific datasets (with more detailed information per each dataset) during the project.

Table 1: Preliminary list of data categories

#	Data category	Description and purpose	Type of origin of data	WP	Format	Data controller
1	AI-CODE Stakeholder contact database	Description: Contact data from the stakeholders of the AI-CODE Project, including first and last name and e-mail address. Purpose: the purpose of processing data under this operation is the subscription to the project's newsletter to communicate publications, events and other activities from the AI-CODE project, which may include opportunities, services, innovation and technology transfer, participation in surveys and questionnaires and other communication and dissemination activities.	Newly collected data	WP9	.xls	AI-CODE Stakeholders
2	User data from next generation social media (Fediverses and metaverses)	Description: Data from social media platforms that will help to describe content, user behaviour and communication patterns. Data will consist mostly of 1) content of social media posts (in three modalities: text, images and videos), 2) users' interaction above such content (implicit and explicit feedback and reactions), 3) metadata about the instances of fediversed platforms. Purpose: reconstruct the social networks and characterise the risks associated with the action of super-spreaders of disinformation, social bots, and how the new social context provided by next-generation social media may contribute to disseminating and giving credit to such content, including echo chambers and polarised communities.	Newly collected and re-used data	WP2	.txt, .tsv, .csv, doc, sqlite3.	FBK, KInIT, ATC, DW, DEBUNK.org, UPM, DS Tech, EIPCM, RU
3	User requirements and specifications data	Description: Data collected in surveys/questionnaires, interviews and workshops from target groups (i.e. journalists, media professionals etc.) regarding user needs and requirements for the development of services. Purpose: the main purpose of the data collected in the surveys/questionnaires, interviews and workshops is to develop the User Requirements (functional, non-functional, and technical) to ensure that AI-CODE services are intuitive and easy to use, embody user-centred ethical requirements, include all needed and requested functionalities. A comprehensive Functional Specification that describes all necessary functionalities required for the AI-CODE	Newly collected data	WP3	.xls, .csv, .txt, .doc, .pdf, .mp4, .mp3, .jpg, .png,	EIPCM, DS Tech, RU, CERTH, ATC, UPM, FBK

		Services will collect this data.				
4	Design and technical development data	Description: Data collected from target groups (i.e. journalists, media professionals etc.) to ensure the AI-CODE service design and technical development using tools and methodologies for co-creation such as storyboarding, design thinking, mock-up development or prototyping techniques. Purpose: the main purpose of the data collected via co-creation sessions with target groups is to lead the development of more user-centred and effective services that meet the needs and preferences of end-users and stakeholders.	Newly collected data	WP3 – WP7	.xls, .csv, .txt, .doc, .pdf, .mp4, .mp3, .jpg, .png, video	AI-CODE partners
5	Content data	Description: Data such as texts, videos, images, and audios that will be collected from next-generation social media to test the services developed by AI-CODE with the aim of detecting fakes and deepfakes for text, images, and videos. Hyperlinks to webpages for media professionals to check content. Purpose: to train disinformation detection models and validate the efficiency and effectiveness of AI-CODE services using real case scenarios.	Newly collected and re-used data	WP4-7	.mp4, .mp3, .jpg, .png, .webp, .html, URL	AI-CODE partners
6	Feedback data	Description: Data report on end-user experiences in using AI-CODE services in real case scenarios. Purpose: to report and improve the experience of AI-CODE services.	Newly collected data	WP8	.pdf, .html	AI-CODE partners
7	Data from previous projects	Description: Research data and considerations from previous funded projects such as vera.ai and AI4TRUST. Purpose: re-use of data and results already investigated and achieved in previous projects to serve as baseline for further evolving collected experiences into next-generation social media and advance prototypes of possible disinformation detection systems for metaverses and fediverses.	Re-used data	WP4-5	.mp4, .mp3, .jpg, .png, .html, URLs	CERTH, FBK, KInIT, UPM

4.1 FAIR Data Management

The Data Management Plan is a key element for a good data management in a project. This DMP is based on the guidelines from the European Commission on FAIR Data Management in Horizon Europe. FAIR stands for findable, accessible, interoperable and re-usable.

The AI-CODE Project will generate and process a significant amount of data. The generation and processing of these datasets will be collaborative among all the partners in the consortium. This DMP will identify and categorize the type of data that will be processed under the scope of the AI-CODE Project. Processing of data means all the operations performed on data, including by manual or automated means. It includes collecting, recording, structuring, storage, alteration, consultation, use, disclosure, alignment, restriction, erasure or destruction of data.

The project has an obligation under the Grant Agreement, article 17, to promote the action by disseminating its results “as soon as feasible, in a publicly available format, subject to any restrictions due to the protection of intellectual property, security rules or legitimate interests.”

4.1.1 Making data findable, including provisions for metadata

Making data easily findable is useful for nurturing further research in similar domains and/or ongoing or future initiatives such as other European projects to ease the understanding of the project and the access to re-usable data.

Digital Object Identifiers (DOI), keywords and metadata standards will be implemented according to the needs of the AI-CODE’s progress and introduced as practices to optimize the potential of finding and re-using data.

Creation of AI-CODE’s communities on data repositories and search keywords to optimise data re-use:

Furthermore, to facilitate the search of AI-CODE related data, a project community will be created in Zenodo. The community will be named including the following information that will help to make related data easily findable:

- Acronym of the project: AI-CODE
- The term “European Union (EU)” and “Horizon Europe”
- The grant number.

Other keywords related to the project may be identified and used to simplify the research of the data in the designed platform. A directory with all the available keywords per category will be built and updated during AI-CODE’s implementation, optimizing data re-use procedures.

Metadata means “data about data”. It is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. The adoption of a good metadata management will ensure the project keeps the quality of data, so users know it can be trusted.

The datasets processed by AI-CODE will have appropriated metadata compiled, which will provide a summary of the characteristics of a dataset. A good metadata record enables the user of a dataset or other information resource to understand the content of what they are reviewing, its potential value and its limitations.

4.2 Making data openly accessible

For each dataset, the partner that owns the data will indicate whether the dataset will be made openly available. If certain datasets cannot be shared (or need to be shared under restrictions), the partner will explain why, clearly separating legal and contractual reasons from voluntary restrictions. The openly accessible data (including the associated metadata and documentation) will be deposited in an open-access repository. Project participants will deposit their data in a research data repository and take measures to make the data available

to third parties. The third parties should be able to access, mine, exploit, reproduce, and disseminate the data. This should also help to validate the results presented in scientific publications.

The preferred repository is Zenodo (<https://zenodo.org/>), an OpenAIRE compliant repository, without precluding the further use of other repositories (e.g., institutional, subject-specific, etc.). All open data deposited in Zenodo will be assigned a persistent and unique DOI. The open data will be accessible without restrictions.

As for the datasets including “personal data”, these will be kept confidential and with all the security measures in place. Details on the procedures and methodology adopted for the processing of the datasets containing personal information, is further explored in Section 7 of this DMP.

The AI-CODE project takes a holistic **open science approach** in all three main phases and across all project’s outcomes. As to open science practices, the partners will ensure:

- **Open access to peer-reviewed scientific publications.** All project-related publications (research papers, technical reports, etc.) will be published in an open-access manner at widely accepted or organisational repositories (green open access), such as arXiv.org or Zenodo, or under suitable open licences (gold open access). Metadata of deposited publications will be made available under a Creative Common Public Domain Dedication (CC0) or equivalent, in line with the FAIR principles. The metadata will provide information at least about the following: publication (author(s), title, date of publication, publication venue); Horizon Europe funding; grant project name, acronym and number; licensing terms; persistent identifiers for the publication, the authors involved in the project and, if possible, for their organizations and the grant. Where applicable, the metadata must include persistent identifiers for any research output, or any other tools and instruments needed to validate the conclusions of the publication.
- **Open access to source code, AI-models and reproducibility.** Source code and AI-models developed during the project will be made available as open source under suitable licences at standard portals (e.g., Github, Hugging Face). Furthermore, the AI-based outcomes will be shared by means of the AI-on-Demand Platform. Reproducibility will also be ensured by making the best performing methods freely available to researchers as web services, with easy-to-use REST APIs, accompanying documentation, and AI-CODE samples.
- **Open access to data.** Datasets created by the project will also be made openly available to researchers (e.g., under a suitable CC-BY licence), provided they do not raise IPR, ethical, and legal issues (cf. Data Management section).
- **Open access to participation.** The whole methodology of the AI-CODE project is fundamentally based on close involvement of stakeholders thanks to the adoption of co-creation methodologies. In addition, the project will provide opportunities for participation and cooperation with other researchers (especially those who are involved in EU funded projects related to the fight against disinformation). Dissemination activities organised during the project will provide another opportunity for participation of the wider European and global R&D community.

4.3 Making data interoperable

Making data interoperable allows systems and services to exchange and consume data amongst each other. AI-CODE will put special attention on the use of standardised data relying in established **vocabularies, standards, formats or methodologies** to facilitate data exchange and interoperability whenever possible.

The adoption of controlled vocabularies in descriptive metadata fields to support consistent, accurate, and quick indexing and retrieval of relevant data is foreseen in the data management. Keywords will be adopted for indexing data and metadata. As controlled vocabularies change within different disciplines, these keywords may be updated during the project to increase the interoperability of the project’s data and metadata. To ensure the interoperability, all datasets shall use the same standards for data and metadata capture and creation.

This section will be further developed once vocabularies, standards or methodologies will be identified to make data interoperable between partners and other institutions, whenever applicable for the dataset.

4.4 Re-use of data

During the project, each data controller shall make data accessible to the AI-CODE partners under the terms of the Grant Agreement (GA) and the Consortium Agreement (CA).

Whenever the processing includes personal data, a Joint Controller Agreement will be in place.

After the end of the project, the partners will define the possible licensing terms and restrictions for each dataset, particularly for third-party use.

The re-use of publicly available data is also part of the methodology adopted for the AI-CODE data management plan.

5 Methodology

5.1 Data acquisition

Data managed under the AI-CODE Project will be acquired through different actions, according to its origin, as defined in table 1. Newly collected data will be collected through analysis of user data on next-generation social media, through surveys/interviews and workshops, through feedback collection.

Other data to be processed under the AI-CODE project might come from a different activity of the project, such as re-used data from previously funded projects or publicly available data.

5.2 Data storage and backups

Datasets will be managed closely with clear responsibility and the AI-CODE consortium will be made accountable for ensuring that data management procedures are followed. Data will be held securely with adequate provision for their long-term care as defined in Section 6.

Datasets will be stored on the:

- Shared repository of the project for use of all partners;
- Internal repositories from the partner who is the data controller of the datasets providing access to the partners when needed at the scope of the project according to provisions defined in Section 6;
- AI-CODE services platform/repository to train AI-models to develop, maintain and improve the AI-CODE offer;
- Open-access repositories such as Zenodo.

Procedures for the backup and recovery of data, defining the frequency and reliability will be adopted by the controllers of each processing operation. Each controller must ensure the adoption of appropriate organizational and technical measures to guarantee data storage and backups.

5.3 Data categories and taxonomies

The definition of the data categories, vocabulary, and taxonomies to be adopted by AI-CODE is a key aspect to achieve the interoperability between databases, as well as to adopt a common language for the project implementation, allowing a common terminology across the Work Packages (WPs).

The data categories to be used by AI-CODE comprise the taxonomies used in EU organizations, EU statistic frameworks and furthermore they will be defined based on the results of the requirements analysis (WP3).

At this early stage of the Project, not all the data categories and taxonomies can be defined, but it is important to emphasize the value of defining an effective and comprehensive data categorization to organize the datasets and the metadata to be processed in the project.

6 Data security

This section defines the general data security provisions for AI-CODE newly collected data, especially two more restrictive AI-CODE data categories which are raw data and internal research data.

At the time of the DMP publication, the concrete definition of final technical and operational details of the implementation of requirements have not yet been decided. This section will be updated with those details in future versions as the AI-CODE infrastructure is developed and the use cases progress.

6.1 Provision for Raw Data

Where applicable, the respective use case leaders (see Table 1) will serve as the data controller for raw collected data and will be responsible to implementing the following security measures:

- Storage that is secure, encrypted and password protected by individual access credentials in the infrastructure from its institution.
- Transfers from and to this storage should be similarly encrypted and protected by individual access credentials.
- Participants will be identified using pseudonymized study IDs or completely anonymized.
- The study's keys and access will be restricted only to the study/pilot staff and the developers of the applications.

6.2 Provisions for Internal Research Data

Research data is obtained from the previous level of raw generated data by performing a data preparation operation. This data preparation operation will be carried out, under the responsibility of the original data controller of the raw collected data, within six months of the completion of the study that generated it and will involve:

- The anonymization of the collected data by deleting the participant pseudonyms, other study keys and the sociodemographic/sensitive data.
- The duly justification of any remaining personal and sensitive data to make sure that it is relevant and limited to the purposes of the project activities (in accordance with the 'data minimisation' principle).

The resulting data is called Research Data and has the following security requirements:

- Storage that is secure and password protected by individual access credentials in the infrastructure from its institution.
- Transfers from and to this storage should be similarly protected by individual access credentials.
- This information will be kept for at most 10 years after the end of the pilot / case study that generated it.

The AI-CODE project consortium partners are the primary target audience of research data. To access it, the AI-CODE partner will send a written request to the dataset controller in which the following information will be contained:

- i) name of the persons to be granted access,
- ii) duration of this access and
- iii) activities for which the data will be used.

The data controller will then decide to approve the request and share access information if the request is aligned with the AI-CODE project objectives.

The partner receiving access to research data, commits to:

- Follow the rules in the AI-CODE CA and AI-CODE GA related to data management and secrecy.
- To use the data during the timeframe and for the activities specified in the written access request and to delete it after these purposes have been achieved.
- To keep data secure and comply with the same security requirements as the data controller.

7 Applicable Legislation

This section will detail the legislations, regulations and guidelines that the project will consider for its data management procedures and practices.

At every step of the project, a broad legal framework combined with ethical frameworks will be considered as they are relevant to the project. The machine learning development process and the case studies specifically will be continuously assessed for alignment with existing and forthcoming EU Directives and Regulations. This DMP focusses primarily on the GDPR, while in the AI-CODE project overall, all relevant regulation will be continuously monitored, particularly the AI Act, that is forthcoming at the time of this DMP.

7.1 General Data Protection Regulation (GDPR)

When personal data is concerned, the legislative cornerstone of the data management practices in AI-CODE is the GDPR Regulation (EU) 2016/679.

The GDPR regulates the processing of personal data in the EU and aims to ensure individual fundamental rights.

Personal data according to Article 4(1) is defined as follows:

“‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.”

The AI-CODE project ensures that personal data is:

- processed lawfully, fairly and in a transparent manner in relation to the data subjects
- collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes
- adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed
- accurate and, where necessary, kept up to date
- kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data is processed and
- processed in a manner that ensures appropriate security of the data.

Requirements under the GDPR further include a data protection impact assessment, informed consent, and privacy-by-design.

7.1.1 Data protection impact assessments (DPIA)

All AI-CODE use case leaders will act according to the article 35 of the GDPR performing the Data Protection Impact Assessment (DPIA) when personal data is collected for further processing.

In accordance with Art.35 (7) the DPIA includes at least:

- a systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller;
- an assessment of the necessity and proportionality of the processing operations in relation to the purposes;
- an assessment of the risks to the rights and freedoms of data subjects referred to in paragraph 1; and
- the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Regulation considering the rights and legitimate interests of data subjects and other persons concerned.

7.1.2 Informed Consent

For personal data to be processed, free and fully informed consent from the data subject is required. Where AI-CODE partners collect primary data themselves, the AI-CODE consortium ensures that all participation in these project activities are voluntary, and the subjects are aware of their rights. They do so by obtaining fully informed consent in advance and clearly documenting the data subject's consent. The research subject should further be informed about the privacy practices via a privacy statement.

7.1.3 Privacy by design

All research and development activities conducted within the scope of the AI-CODE project will be mindful of the "Privacy by Design" principle of article 25 GDPR, as well as article 22 GDPR assuring compliance with automated individual decision-making including profiling.

7.1.4 Due Diligence Documentation

In order to keep an auditable document trail of the project data management activities and agreements as per the GDPR, the following signed documents need to be accessible digitally in a secure project shared storage space:

- final version of the DPIA document
- Privacy Statements: the final text of the Privacy Statement given/displayed to study participants.
- Proof of informed consent: either a signed document (in the case of a paper/in person-based study) or a table summarizing the names/pseudonyms (in the case of a digital based study) of the persons that have voluntarily accepted participating in the study.

The AI-CODE consortium will also ensure transparency on the recruitment process, inclusion/exclusion criteria for participation, as well as the procedures of how informed consent is obtained from study participants.

7.2 Fundamental Rights

All partners will adhere to relevant national and international laws, guidelines and policies, that aim to protect the human rights, dignity and integrity of the data subject and user of the AI use cases.

These include:

- Universal Declaration of Human Rights (United Nations)
- Charter of Fundamental Rights of the European Union (2010/C83/02)
- Convention for the Protection of Human Rights and Fundamental Freedoms (Council of Europe)

7.3 Ethical Frameworks

The Ethics Guidelines for Trustworthy AI will be considered at every step of the project and provide an additional AI governance framework to ensure responsible use and creation of AI.

We will ensure that the project's methodology complies with the 'do no significant harm' principle as per Article 17 of Regulation (EU) No 2020/852 on the establishment of a framework to facilitate sustainable investment, also known as the EU Taxonomy Regulation. This means that the project is designed in a way it is not significantly harming any of the six environmental objectives.

Over the course of the AI-CODE project, at least one assessment of the use cases with the accompanying Assessment List for Trustworthy AI (ALTAI) will be carried out.

7.3.1 Due Diligence Documentation

For the sake of keeping an auditable document trail of the project data management activities and agreements as per the ethical frameworks, the following signed documents need to be accessible digitally in a secure project shared storage space:

- The results of the ALTAI assessment,
- If ethical approvals were required and obtained, the final version of the document and all associated material submitted to the institutional Ethical Committee for their approval. The letter or email where

the committee expresses their feedback and approval should also be preserved.

8 Allocation of resources

DS Tech S.r.l. is the partner responsible for the data management within the project. DS Tech S.r.l. will report the updates in the DMP to the WP Leaders, to include them in the Periodic Reports of the project. Nevertheless, each partner of the AI-CODE project will be responsible for the safe storage of their own datasets. Each partner will be responsible for the identification of new datasets that will be handled within their own tasks. Each partner will be responsible for filling in the Data Summary form (Annex 1) and upload it on the project repository platform and notify it to DS Tech S.r.l.

DS Tech S.r.l. will be responsible for updating this new dataset information into the DMP and for uploading the dataset to the AI-CODE repository. If the dataset will be made openly accessible, the dataset owner will be responsible for uploading it to Zenodo.

For the publications, the consortium will publish them in scientific journals that allow open access. The associated costs will be claimed as part of the Horizon Europe grant.

9 Conclusions

This DMP sets the framework of an effective management of the data processed under the AI-CODE project.

It identifies the datasets to be processed, its purposes, format, origins, and proposes procedures for acquisition, storage/backup, security and methodology to ensure the alignment with the European Commission guidelines on FAIR Data Management.

This document will be progressively updated as the AI-CODE project evolves and develops. This might entail updating the identified datasets (addition or removal) to be collected and processed, defining more details related to the methodology for acquisition, storage and back-up of data as well as defining categories and taxonomies to facilitate their findability, interoperability and re-use.

The AI-CODE project is committed to follow the guidelines of the European Commission on FAIR Data Management as well as ensure that the highest standards policies on data management are in line with the project execution.

10 Annex 1: Data Summary form

This Data Summary form is a template for the collection of information related to data that are newly collected/generated or re-used within the scope of the AI-CODE project and that will serve the purpose of keeping up to date the Data Management Plan.

10.1 Name of the beneficiary

[Enter text]

10.2 Data processing

Dataset identifier
[Enter text]
WP and Task
[Enter text]
Dataset description
[Enter text] Describe the dataset in a few lines.
Purpose of the data
[Enter text] Explain the purpose of the data collection/generation and its relation to the objectives of the project.
Data controller
[Enter text]
Format of data
☐ Text: plain text (TXT), HTML, XML, PDF/A ☐ Databases: XML, CSV, sqlite3. ☐ Image: JPEG, JPG-2000, PNG, TIFF. ☐ Audio: AIFF, WAVE, .mp4, .mp3. ☐ Containers: TAR, GZIP, ZIP. ☐ Video: mp4, AVI., WMV., AVCHD., WebM., FLV. ☐ Other.
Type of origin of the data
☐ Newly collected/generated ☐ Re-used (please explain): [Enter text]
Expected size of data
Give an estimation: e.g. a few kB/MB/GB, etc. ☐ Fixed: never change after being collected or generated ☐ Growing: new data may be added but the old data is never changed or deleted ☐ Revisable: new data may be added, and old data may be changed or deleted

Utility of the data
<i>[Enter text]</i> Indicate to whom the data might be useful e.g. partners, stakeholders, citizens, etc.
Software involved in the data processing
<i>[Enter text]</i>

10.3 FAIR Data

10.3.1 Making Data Findable

Standard identification mechanism and metadata
<i>[Enter text]</i> Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers).
Naming convention
<i>[Enter text]</i> Indicate any naming convention that you follow.
Keywords
<i>[Enter text]</i> Provide keywords to optimise possibilities for re-use.

10.3.2 Making Data Openly Accessible

Dataset accessibility
<i>[Enter text]</i> If confidential, explain why, clearly separating legal and contractual reasons from voluntary restrictions.
Software tools
<i>[Enter text]</i> Indicate what methods or software tools are needed to access the data.
Repository to be used to deposit the dataset
<i>[Enter text]</i> Where will the data and metadata be deposited? Preference should be given to certified repositories which support open access where possible.

10.3.3 Making Data Interoperable

Are the data interoperable?

[Enter text]

Brief description of the tools/vocabularies used to make the data usable by other researchers, institutions, etc.

10.3.4 Increasing Data Re-use

Restrictions

[Enter text]

If applicable, define the data licensing approach to permit the widest re-use possible. Indicate the chosen license tools.

Indicate for how long it is intended that the data remains re-usable

[Enter text]

Explain how data quality is assured

[Enter text]

10.4 Allocation of resources

If applicable, indicate any allocation of resources to data management

[Enter text]

10.5 Ethical aspects

Please indicate if you have identified any ethical or legal aspect that could impact the data sharing

[Enter text]

10.6 Data Security

Indicate the provisions in place to ensure data security

[Enter text]