SEcure Decentralised Intelligent Data MARKetplace

D1.3 Data Management Action Plan. Second version

Document Identification	
Contractual delivery date:	31/03/2024
Actual delivery date:	01/04/2024 – 02/08/2024
Responsible beneficiary:	ATOS IT
Contributing beneficiaries:	SIE, MYT, EGM, WINGS, UC, INRIA, NUID UCD, SDR, FV, SURREY, LINKS
Dissemination level:	PU
Version:	2.0 (Resubmitted)
Status:	Final

Keywords:
Data, Management, Plan, DMP, FAIR

This document is issued within the frame and for the purpose of the SEDIMARK project. This project has received funding from the European Union's Horizon Europe Framework Programme under Grant Agreement No.101070074. and is also partly funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission or UKRI.

Document Information

Document Identification			
Related WP	WP1	**Related Deliverables(s):**	D1.1, D1.3, D1.4, D2.1, D5.1
Document reference:	SEDIMARK_D1.3	**Total number of pages:**	37

List of Contributors	
Name	**Partner**
Arturo Medela, Miguel Ángel Esbrí	ATOS IT
Elias Tragos Aonghus Lawlor Neil Hurley	NUID UCD
Tarek Elsaleh, Adrian Hilton	SURREY
Alberto Carelli	LINKS
Luis Sánchez	UC
Verónica Gutiérrez	SDR
Maroua Bahri	INRIA

Document History			
Version	**Date**	**Change editors**	**Change**
0.1	31/01/2024	ATOS IT	Table of Contents proposed.
0.2	29/02/2024	ATOS IT	Initial input in all sections
0.3	04/03/2024	NUID UCD	Contributions to chapters 2, 3, 4.
0.4	05/03/2024	SURREY	Input for 4.1.2, 4.1.3, 4.1.4, 4.2.4.2, 4.3.1, 4.3.3, 4.3.4.
0.5	05/03/2024	LINKS	Input for 2.2
0.6	05/03/2024	UC, SDR, SIE, ATOS, INRIA	Input for 3.1, 3.4, 4, 4.1.3, 4.1.4, 4.2.3.5, 4.3.1, 10
0.9	01/04/2024	ATOS	Quality Review Form

Document History			
Version	**Date**	**Change editors**	**Change**
1.0	01/04/2024	ATOS	FINAL VERSION TO BE SUBMITTED
2.0	22/07/2024	ATOS	Addressed revision requests from the reviewers.

Quality Control		
Role	**Who (Partner short name)**	**Approval date**
Reviewer 1	Grigorios Koutantos (WINGS)	27/03/2024
Reviewer 2	Tarek Elsaleh (SURREY)	25/03/2024
Quality manager	María Guadalupe Rodríguez (ATOS)	01/04/2024
Project Coordinator	Miguel Ángel Esbrí (ATOS)	01/04/2024

Table of Contents

List of Acronyms

Abbreviation / acronym	Description
AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
CA	Consortium Agreement
CC	Creative Common
CC-BY	Creative Commons Attribution
CSV	Comma Separated Values
DCAT	Data Catalog Vocabulary
DLT	Distributed Ledger Technology
DMAP	Data Management Action Plan
DoA	Description of Action
.doc or .docx	Microsoft Word document
DOI	Digital Object Identifier
Dx.y	Deliverable number y belonging to WP x
EB	Ethics Board
EC	European Commission
EU	European Union
FAIR	Findable, Accessible, Interoperable and Re-usable
FLAC	Free Lossless Audio Codec
GA	Grant Agreement
GDPR	General Data Protection Regulation
GeoJson	Georeferenced JSON file
GeoTiff	Georeferenced TIFF image
GIS	Geographical Information System
GIT	Global Information Tracker
GML	Geographic Markup Language
HE	Horizon Europe

Abbreviation / acronym	Description
HTTP	Hypertext Transfer Protocol
IoT	Internet of Things
IPR	Intellectual Property Rights
ISO	International Organization for Standardization
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
KPI	Key Performance Indicators
ML	Machine Learning
MPEG	Moving Picture Experts Group
Mx	Month number
OAI-PMH	Open Archives Initiative – Protocol for Metadata Harvesting
ODRL	Open Digital Rights Language
ONNX	Open Neural Network Exchange
OWL	W3C Web Ontology Language
PB	Project Board
PC	Project Coordinator
PDF	Portable Document Format
PNG	Portable Network Graphics
.ppt or pptx	PowerPoint Presentation
.psd	Photoshop Document
RDF	Rich Text Format
RDF(s)	Resource Description Framework Schema
SPARQL	SPARQL Protocol and RDF Query Language
SDMs	Smart Data Models
SSN	Semantic Sensor Network
.svg	Scalable Vector Graphics
TIFF	Tag Image File Format
.txt	Text file document
UK	United Kingdom

Abbreviation / acronym	Description
.vsd	Visio Drawing File
WAV	Waveform audio file format
WP	Work Package
.xls or xlsx	Microsoft Excel document
.zip	“Move at high speed” (meaning)

Executive Summary

SEDIMARK knows the importance of regulating data management issues within a context such as the one posed by the project. A solution will be considered where consortium partners will deposit all underlying information on data-related business processes (data storage, data provisioning, processing etc.) of the SEDIMARK solution clearly and transparently.

The purpose of the Data Management Action Plan (DMAP) is to identify the main data management elements that apply to the SEDIMARK project and the consortium. This document is the first version of the DMAP and will be reviewed as soon as there is a clearer understanding of the types of data that will be collected.

Given the wide range of sources from which data will be collected or become available within the project, this document outlines that the consortium partners will consider embracing and applying the Guidelines on FAIR Data Management in Horizon 2020 and Horizon Europe (HE); "*In general terms your data should be 'FAIR', that is Findable, Accessible, Interoperable and Re-usable"* [1]*, as information about data to be collected becomes clearer*".

Open access is defined as the practice of providing on-line access to scientific information that is free of charge to the reader and that is reusable. In the context of research and innovation, scientific information can refer to peer-reviewed scientific research articles or research data.

The SEDIMARK consortium strongly believes in the concept of open science, and in the benefits that the European innovation ecosystem and economy can draw from allowing the reuse of data at a larger scale.

Hence, as described in this report, the consortium will deposit the public data produced within or collected for the purposes of the project in an open data repository once the exploitation rights are safeguarded. It will permit the user to access, mine, exploit, reproduce and disseminate free of charge. Furthermore, it will provide information as well about tools and instruments necessary for the project results whenever possible.

As Data Security and Data Privacy are of particular concern, SEDIMARK will consider the GDPR (General Data Protection Regulation) privacy principles from conception to design to build a novel privacy-preserving architecture.

Moreover, this report addresses the ethics issues detected in SEDIMARK with the aim of:

- Providing the procedures and criteria that will be used to identify/recruit research participants when involving people (external to the project consortium) in the research activities (including dissemination) needed for the implementation of the project tasks and / or elaboration of the project deliverables.
- Creating the templates of the informed consent/assent forms and information sheets (in language and terms intelligible to the participants) that will be needed to use the information provided in the research activities or for using the personal data for dissemination activities.
- Ensuring that all AI-related research conforms to the Ethical AI guidelines for Trustworthy Artificial Intelligence.

This deliverable takes as a reference the updated guidelines for Horizon Europe Programme regarding open research and innovation, data management and ethics support.

All in all, this report represents a second snapshot taken mid-way through the execution of the project (already in its eighteenth month of activity). The final update on the data management plan will be reported after at M36, September 2025, as part of a dedicated section for data management plan updates.

Due to the nature of this report, iterative and offering annual updates, the new content provided in this iteration will be written in a dark blue colour, sticking to the project's stylistic guidelines.

1 Introduction

1.1 Purpose of the document

This document is the second version of the Data Management Action Plan (DMAP) of SEDIMARK project (GA contract no 101070074). This deliverable (D1.3) is under the activities of task T1.3 “Data Management” within the Description of the Action of SEDIMARK project.

SEDIMARK aims to provide an enriched secure decentralised data and services marketplace, where scattered data from various domains and geographical locations within the EU can be easily generated, cleaned, protected, discovered, enriched with metadata, Artificial Intelligence (AI) and analytics and exploited for diverse business and research scenarios. SEDIMARK aims to develop tools that support and promote the concept of FAIR data, ensuring that highly interoperable data are securely stored, discovered, accessed, and easily/securely reused in various contexts.

The European Commission (EC) provided a document with guidelines for project participants. Those guidelines address aspects like research data quality, sharing and security. According to them, project participants will need to develop a DMAP. This DMAP describes the types of data that will be generated or gathered during the project, the standards that will be used, the ways the data will be exploited and shared for verification or reuse, and how the data will be preserved.

This report has been produced following these guidelines and aims to provide a consolidated plan for SEDIMARK partners in the data management plan policy that the project will follow. This document is the second version delivered in March 2024, M18 of the project. Accordingly, the DMAP will be updated during the lifecycle of the project, which translates in a previous iteration coming in M6, March 2023, and an ulterior and final one in M36, September 2025, respectively.

1.2 Relation to other project work

SEDIMARK DMAP will be written in reference to the concept called “Open access to research data” [2]. The DMAP will be important for tracking all data produced during the SEDIMARK project execution.

Regarding the nature of the data, to fulfil the required security and privacy requirements in this project, which are set by the Data Protection Directive (Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data) and the General Data Protection Regulation (GDPR) [3], the project assumes the differentiation set in this Directive between Personal and No Personal data. Data are considered personal data “*when someone is able to link the information to a person, even if the person holding the data cannot make this link*.” Any data susceptible to being considered as Personal Data will be managed according to this Directive.

1.3 Structure of the document

This document is structured in 10 major chapters.

Chapter 1 is the introductory part of the document.

Chapter 2 presents the methodology followed in relation to data collection and processing, data protection and open science practices for research data.

Chapter 3 presents a summary of data to be handled within the project execution, depicting aspects such as formats, sizes, or purpose.

Chapter 4 recaps how the project aligns with FAIR principles.

Chapter 5 summarizes other research outputs that may be relevant.

Chapter 6 addresses the allocation of resources to proceed with data management-related activities.

Chapter 7 focuses on data security aspects in relation to provisions and storage.

Chapter 8 deals with ethical aspects.

Chapter 9 handles other potential issues related to this data management duties.

Chapter 10 aims to present the template to list project datasets in the upcoming project execution stages.

Chapter 11 wraps up the report with some conclusions.

2 Methodology

Being in line with the EU's guidelines regarding the DMAP, this report aims to address, for each data set collected, processed and/or generated within the SEDIMARK project, the following characteristics:

- Dataset description, reference, and name.
- Methodologies for collection and processing.
- Standards and metadata.
- Whether or not the data will remain private or in open access.
- Archiving and preservation, during the project lifetime and beyond.

2.1 Data collection and processing

Data collected across the SEDIMARK project can take different forms: sensor data, surveys, software data sets, deliverables, etc. They can be grouped into three different categories:

1. Data collected from measurements/surveys and the project's use case pilot sites.
2. Software components implemented by the project's partners.
3. Report/analyses produced by the partners/consortium.

It is part of SEDIMARK's core commitment to produce open data and open-source software components. The partners of the project follow a conservative approach to evaluate the risks before sharing any data sets, by keeping data in private repositories managed by the consortium. Responsibilities will be assigned to data collectors and processors in order to ensure proper management of the access rights. Any dataset containing personal data, or a group of data allowing to trace back the person they refer to, will be categorized as sensitive and will be held confidential, i.e., with restricted access only to the relevant project partner or the consortium. Moreover, it will be discarded once irrelevant to the project, or at the end of the project. Only non-sensitive data will be made public and live beyond the project.

2.2 Data protection

The SEDIMARK Partners commit to implementing data protection procedures during the handling of data, following data minimisation principles and good security practices.

Data Storage and Collection procedures will follow the collection, storage, and management flows of data decided between the involved partners in the SEDIMARK Consortium.

Data can be exchanged according to the needs presented by the owner of the data. Depending on the nature, type and size of data, multiple suitable means could be identified for protecting its exchange.

In general, the SEDIMARK Consortium advises and enforces the exchange of encrypted data to maintain its confidentiality. More important, such exchange should comply with existing and well-known secure transfer protocols that are able to guarantee also the integrity of the data exchanged.

Among the simplest mechanisms for exchanging pieces of data, there are the general repository at [1] and the code hosting at [2], which are the official repositories employed for the SEDIMARK project.

Both solutions provide more granular Access Control mechanisms granting or denying a singular user belonging to a partner of the SEDIMARK Consortium access to the resource of interest.

Such an access control mechanism leads to improved security for the data itself. As drawbacks of this solution, it has to be pointed out that none of them is actually suitable for exchanging large amounts of data, and thus they shall not be used for such purpose. Nevertheless, the interested partners can setup and provide their own custom channels for data exchange.

As an additional method -where it applies, the owner of the data can provide small pieces of data, replicating the feature of interest. The data collected, stored, and managed adheres in such cases to the Data Minimisation principle.

In particular, the sample of data is such that they are able to ensure that the data itself is adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed. Nevertheless, it has to be emphasised that it is not the target of the SEDIMARK project to collect, manage and store personal data along the whole project lifetime.

Moreover, any data collected during the Use Cases operations will be anonymised or pseudonymised, to prevent the identification of individuals through (possibly) personal data. Following the best practices for anonymization techniques through ad-hoc anonymisation tools, the anonymisation procedure will be duly reported.

Finally, in any case, the retention periods for different types of (possibly) sensitive data collected, stored and managed are limited to the strictly required time it is necessary. After that, the data will be securely disposed and deleted.

2.3 Open science practices for research data

SEDIMARK acknowledges the Open Science Policy of the EC and thus will establish open science practices as defined in the HE guidelines. All partners will ensure the early and open sharing of the project outcomes with sharing of reports and pre-prints in open-access repositories (i.e., arXiv [4]). SEDIMARK will ensure the reproducibility of the outputs by covering the three main research processes that reproducibility is based on: reproduction, replication, and re-use.

SEDIMARK will target Open Access Journals and Open Access Repositories for project publications, increasing visibility, usage, and impact of research, utilizing both Green Open Access (self-archiving in open repositories) and Gold Open Access (peer reviewed publications in open access journals or/and repositories) strategies. The active involvement of end users will be used to ensure that the SEDIMARK marketplace will be widely accepted and successful.

3 Data Summary

This and the following sections are based on the EC Data Management Plan template for Horizon Europe programme [2].

This section shall provide a detailed description of these elements in order to ensure their understanding by the partners of the consortium. For each element, there is also a description of the strategy that will be used to address it.

At this point, an estimation of the size of the data should be provided. To this end, the consortium develops several strategies that will be followed in order to address the above elements.

3.1 Data re-use

SEDIMARK will re-use already available data to fulfil its objectives, since the project will ingest datasets available in the four different demonstration sites where use cases will take place (e.g. coming from Open Data portals, smart city platforms, proprietary data lakes, custom IoT deployments, etc.). The activities in which such datasets will be involved in, will result in the production of processed data that may be of interest to diverse stakeholders. For developing the project tools and modules, public well-known datasets are being used by the project partners. The goal is to first test the developed tools on existing datasets so that the tools can also be benchmarked against well-known datasets and then use the tested tools on the project datasets that are produced in the use cases.

3.2 Types and formats of data

SEDIMARK will do its best to employ commonly used or open file types, for instance:

- For text documents: Microsoft Word (.doc or .docx), Rich Text Format (.rtf), PDF (.pdf) or text files(.txt)
- For tabular data: Microsoft Excel (.xls or .xlsx) or Comma Separated Values (.csv) or JSON files.
- For images: TIFF (.tif), JPEG(.jpg) or PNG(.png).
- For audio: Free Lossless Audio Codec (FLAC) (.flac) or WAV (.wav).
- For video: MPEG-4 (.mp4).
- For other commonly used data formats domain/sector-specific we can find a large variety of them such as:
  - geospatial datasets encoded using ESRI Shapefiles, GeoJson, GeoTiff, GeoPackage, Geographic Markup Language (GML), etc.
  - Ontologies and semantic datasets encoded using OWL, RDF(S), SPARQL, etc.
  - AI/ML models encoded using TensorFlow models (saved as protocol buffer files, with .pb extension), Keras models (natively saved as .h5 files), Scikit-Learn models (as pickled Python objects, with a .pkl file extension) or ONNX ML framework independent file format.

With regard to the data minimisation principle which SEDIMARK aims to follow, compression methods are encouraged, but lossless compression should be applied where possible.

Data to be generated and used within SEDIMARK can be split in four main categories of types:

- Internal project data: this relates to data that are internal to the project consortium and are necessary for the administrative processes of the project and the normal day to day activities. This type of data might include partner info, contact names/addresses, etc. similarly external people contact/names/etc. that participate in project activities (i.e., External Advisory Board, EAB, members) or in any user trials.
- Dissemination data: manually collected data used mostly for dissemination and can be i.e., data from surveys, or photos/videos from dissemination activities.
- Use case data: automatically collected data, i.e., from sensors in the use cases or data that are automatically computed, i.e., AI models, analytics, averages, statistics from the raw data.
- External public open data: these are external datasets used within the technical work of the project in order to help build and test the modules and tools.
- Project outputs: this category might include the project documents, reports, deliverables, software, etc.

3.3 Purpose

The EU data economy has grown tremendously, with forecasts predicting to reach 800 billion euros in 2025. Data are becoming the new currency, being exchanged as products or services in marketplaces. Data markets are predicted to reach a value of 100 billion Euros in 2025. Existing data marketplaces are centralised, store the data on the cloud, provide limited to no guarantees about data quality and are governed by single entities that make the rules. SEDIMARK gathers a large team of experts to build a secure, trusted, and intelligent decentralised data and services marketplace, based on Distributed Ledger Technology (DLT) and Artificial Intelligence. With the aim of validating the proposed solutions and increase the replication potential of the project, 4 use cases were identified to address significant scenarios. These use cases are represented by 4 complementary demo-sites ready to upgrade SEDIMARK solutions.

Considering such activities, main project datasets will be used or generated by the activities framed in the definition of project requirements and platform architecture (WP2), the techniques to assure distributed data quality and management (WP3), the plan to assure a secure data sharing within a decentralized marketplace (WP4) and their subsequent merging in one integrated solution and the corresponding demonstration campaigns (WP5). In addition, datasets will be involved in dissemination, exploitation, and standardisation (WP7) and management (WP1) activities.

All these datasets are structured into information and uploaded to SEDIMARK's main storage system on OwnCloud [6]. Additionally, other storage platforms are used, depending on the nature or purpose of the data. Compilation and analysis of this information results in knowledge that is disseminated as reports, press releases or scientific articles, for instance, accessible to different stakeholders. Furthermore, and according to what is explained in project Grant Agreement (Articles 16 and 18), third parties are able to access, mine, exploit, reproduce and disseminate data and information generated or derived throughout the project, according to SEDIMARK's ethics, privacy and data protection requirements.

3.4 Expected size

At the time being, it is not possible to measure the expected size of data that will be handled by SEDIMARK, since it will heavily rely on the datasets that will be finally shared and employed

in the 4 use cases and in the results offered by the corresponding processing activities performed using AI techniques.

For example, in the Santander case, 200 e-bikes belonging to the public bike sharing service will be monitored. The amount of information coming from the sharing service will depend on the actual usage of the e-bikes.

In further iterations of the DMAP, the consortium will be in the position to provide a much more precise view on data size.

Regarding external public open datasets, their size varies significantly depending on the task at hand, but on average these are in the order of tens or hundreds of MBs.

For the Helsinki mobility digital twin related data sets, the smallest expected data sets are few KBs, updated on an hourly basis. The larger datasets that describe the urban mobility environment (e.g. the street areas) are larger (tens of MBs), but their updating frequency is much lower.

3.5 Data origin

Data originally employed in the project activities will primarily relate to its 4 use cases. Hence, in the particular case of the demonstration scenario that takes place in Santander (Spain), data relates to bike sharing service (in terms of availability, location, etc.), represented using Smart Data models for open access.

In the case of accurate energy forecasting, a combination of heterogeneous data will be processed to produce the desired results. Interoperability is of vital importance and the use case will utilize the SEDIMARK information models. To do so, owners of this use case collect and analyse energy-related and customer data in specific locations in Greece.

On the other hand, the use case for valuation and commercialization of water data will heavily rely on open datasets related to GIS, weather observations or water management and distribution infrastructure, as well as anonymized water consumption information in specific parts of France.

Finally, the Mobility Digital Twin use case in Helsinki will employ data which describes the infrastructure, depicts mobility events, and describes environmental and other conditions. The commitment is to use open standards, but it is worth noting that some of the data may not yet have established standards for mobility domain.

Moreover, the existence of an open data enabler in SEDIMARK's architecture suggests certain data from external data providers will be ingested as well into the platform.

3.6 Data utility

Data resulting from SEDIMARK activities may be of interest to diverse types of data consumers, such as:

- Public authorities, who might benefit from new insights to conduct management activities (e.g., water consumption related) and risk mitigation.
- Start-ups or businesses willing to provide new services based on the processed data offered by SEDIMARK.
- Individuals or entities which need tools to conduct data quality management actions.
- Utility providers, customers, research and university audience, internal & external sales network, energy advisors in the energy ecosystem.

- Decision-makers in municipalities, diverse biker users (e.g., individuals, cycling associations), and entrepreneurs on bike-rental field.
- Researchers who are working in the areas of interest of SEDIMARK. Researchers can use a variety of the data produced within SEDIMARK, (i) using the datasets for building their own AI/ML models and applications, (ii) using the AI Models for testing and comparing their own models, (iii) using the SEDIMARK tools for developing their own modules for i.e. cleaning data.

4 FAIR data

The FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles emphasise the capacity of computational systems to find, access, interoperate, and reuse data with no or minimal human intervention because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data [7] [8].

SEDIMARK considers these principles particularly relevant to build the basis for a solid solution. The following sections explain specifically how SEDIMARK will ensure that the diverse datasets handled within the project will comply with those FAIR data principles not only at project level but with other sister projects that extend their work in parallel.

Additionally, it is important to note the different types of data that will be managed within the SEDIMARK project (cf. Section 3.2). In this sense, it is worth highlighting that the main aim of the SEDIMARK project is designing and prototyping a secure data and services marketplace that allows the efficient and privacy-preserving sharing of vast amounts of heterogeneous data (specifically, data catalogued as Use case data within the four main categories of data identified as generated and used in SEDIMARK). Hence, for this kind of data, the data identification, data discovery, metadata creation, publication, and discovery, as well as standards used and mechanisms for accessing that data, are the main subjects of the technical activities carried out in the project. Thus, they are described in the corresponding technical documents and deliverables reporting the progress of those activities. For the specific case of this kind of data, concrete insights will be provided in the following subsections.

4.1 Making data findable, including provisions for metadata

According to the general guidelines [9], to make data findable:

- F1. (Meta)data are assigned a globally unique and eternally persistent identifier.
- F2. Data (including datasets and AI models) are described with rich metadata.
- F3. (Meta)data are registered or indexed in a searchable resource.
- F4. Metadata specify the data identifier.

4.1.1 Data identification

SEDIMARK plans to establish whether the data produced and/or used in the project are identifiable and traceable through a standard identification mechanism (e.g., persistent and unique identifiers such as Digital Object Identifiers, DOI). This will include naming conventions and identifying search keywords.

4.1.2 Data discovery

SEDIMARK plans to establish whether the data produced and/or used in the project are discoverable with metadata. A Catalogue will be the main mechanism for enabling discovery, whereby it will rely on other system entities residing on dedicated or elected baseline infrastructure, or through interactions with Self-listing repositories residing the Asset Providers domain. Self-listings will contain metadata descriptions for representing offerings from Providers, which will be used by the Catalogue's querying mechanism.

4.1.3 Metadata creation

Metadata is 'data about data,' which provide information about other data or information on the content of data, files, documents and other material.

In SEDIMARK's case, data annotation is also a part of the data curation and quality management pipeline that the project will build. So, the data generated and/or exchanged within the platform shall be properly annotated with metadata that will help describe such data and make them more findable. This will be applied not only to data assets but also to other assets generated such as AI models and processing services. For AI models, this includes aspects relating to its general application, methods applied and nature of data it consumes. For services, this will include aspects in relation to what data in processes and operations it can perform.

Moreover, the SEDIMARK Data Marketplace will expose data assets by publishing Offerings, which are metadata in the form of descriptions about the dataset or data stream that a data provider is making available through the marketplace.

What specific metadata will be created will be considered and, in the event that metadata standards do not exist, SEDIMARK will outline the type of metadata created and how it will be created.

4.1.4 Standards

The current version of the SEDIMARK DMAP does not include a compilation of all the metadata about the data being produced in SEDIMARK project, but there are already several domains considered in it which follow different rules and recommendations. What follows is a very early-stage identification of standards that may be employed along the way.

- Microsoft Office 2010 for text-based documents (or any other compatible version). doc, .docx, .xls, .xlsx, .ppt, .pptx. Also, especially where larger datasets need to be dealt with, .csv and .txt file formats will be used. All finished and approved documents will also be made available as .pdf documents.
- Illustrations and graphic design will make use of Microsoft Visio (Format: .vsd), Photoshop (Format: different types possible, mostly .png), and will be made available as .jpg, .psd, .svg, .tiff and .ai files.
- .mp3 or .wav formats for audio files.
- QuickTime Movie or Windows Media Video for video files.

These file formats were chosen because they are widespread and accepted standards. When possible, files will be converted to open file formats for long-term storage purposes.

Metadata will be composed of two formats: a) contextual information about the data in a text-based document, and b) ISO 19115 standard metadata in an .xml file. These two formats for metadata are chosen to provide a full explanation of the data (text format) and to ensure compatibility with international standards (.xml format).

Specifically, for SEDIMARK, pertaining to metadata about the datasets that will be available through the SEDIMARK Data Marketplace in the form of so-called Offerings, will be tailored to the SEDIMARK platform's needs but it will be heavily based on the DCAT-AP recommendation [10] using semantic annotation compliant with RDF [11].

Data from the use cases will be annotated automatically within the SEDIMARK's data curation and quality improvement pipeline. The annotation will follow standards such as OWL [12] and RDF [11] for descriptions that will enable asset discovery, and NGSI-LD for the annotation of

the data asset itself [13]. This is an ongoing research activity, so more details will be provided in the next version of this document.

4.1.5 Metadata discovery and indexation

Under SEDIMARK project's activities, some files may need metadata which will accompany other generated data, such as:

- Tabular data files (.csv, .xlsx) may need an additional '*notes'* or '*readme'* tab containing a brief description of the contents of the file.
- Audio, video and image files should be accompanied by a '*readme'* text file that briefly explains the context of such files (for instance, stating the date of creation, purpose of the file and its authors/creators).
- Scientific publications and other dissemination material may need metadata (e.g., acknowledging the funding received from the HE program, stating the name of the action, acronym and Grant Agreement number and authors if applicable).
- The data from the SEDIMARK use cases will also be annotated automatically. The respective metadata will be indexed and be available to be discovered by the users of the system via the SEDIMARK data discovery mechanisms. This is an ongoing research activity, so more details will be provided in the next version of the deliverable.

4.2 Making data accessible

Data can be considered as accessible when they comply with a couple of rules:

- A.1. (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.

The data generated by the use cases will be made accessible through the SEDIMARK data marketplace, assuming that the users have proper authorization to access them. This data sharing process is the ultimate goal of SEDIMARK, so information will be provided in the respective technical deliverables.

4.2.1 Trusted Repository

SEDIMARK plans to establish a trusted repository where the consortium will store and share the data exchanged within the project execution (excluding the data generated in the use cases, which will be stored locally at the data providers premises, considering that SEDIMARK is a decentralised system). As for which data repository will be used and why this is a topic that will become evident once partners have a better understanding of the data to be collected. At first, SEDIMARK looks at Zenodo [12] as one potentially valid solution.

Zenodo is an EU-backed portal based on the well-established GIT version control system [13] and the Digital Object Identifier (DOI) system [14]. The portal's aims are inspired by the same principles that the EU sets for the pilot; Zenodo represents thus a very suitable and natural choice in this context. The repository services offered by Zenodo are free of charge and enable peers to share and preserve research data and other research outputs in any size and format: datasets, images, presentations, publications, and software. The digital data and the

associated metadata are preserved through well-established practices such as mirroring and periodic backups. Each uploaded dataset is assigned a unique DOI rendering each submission uniquely identifiable and thus traceable and referenceable.

Thus, all collected datasets will be disseminated without an embargo period unless linked to a green open access publication. Data objects will be deposited in ZENODO under:

- Open access to data files and metadata and data files provided over standard protocols such as Hypertext Transfer Protocol (HTTP) and Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH).
- Use and reuse of data permitted.
- Privacy of its users protected.

By default, data access policy will be unrestricted unless otherwise specified. The generic Creative Commons Attribution CC-BY licenses will be used. This license allows:

- Sharing - copy and redistribute the material in any medium or format.
- Adapting - remix, transform, and build upon the material for any purpose, even commercially.

4.2.2 Data Identification

The moment SEDIMARK partners decide on the trusted repository to employ and get to know their specifics and features it will be possible to confirm whether or not such repository assigns an identifier to data, how and the possibility to resolve such identifier to a digital object.

4.2.3 Data

4.2.3.1 Data availability

Most data produced and/or used in SEDIMARK will be made openly available within the project, considering that in certain cases data shared by partners in the consortium will be just partially available. Pilot data which are marked under the open access concept can be found in further deliverables. When external data are considered, a similar analysis must be conducted. For instance, data ingested through the open data enabler will be marked as open access, however, in other cases, this could not be possible.

Each SEDIMARK partner must respect the policies set out in this DMAP. Datasets must be created, managed and stored appropriately and be in line with applicable legislation and the agreements signed.

The Project Coordinator (PC) has a particular responsibility to ensure that all the required datasets for developing the project innovations are made available to the consortium no matter if the datasets are declared confidential (therefore not publicly available). In this event, the project will do its best to either find alternative datasets or ensure that the datasets are appropriately anonymised, and that no confidential information will be shared even between the consortium partners. To promote partner agreements that are not against the spirit of the signed Consortium Agreement (CA) is part of the project coordinator's duties.

In the event that certain datasets cannot be shared or need to be shared under restrictions, appropriate explanations will be provided based on legal and contractual reasons. During subsequent versions of this document, the appropriate intellectual property protection will be described.

The proprietary partner that provides the datasets has the responsibility to manage the different versions and to make sure that the latest version is available to all the partners that require the datasets.

The partner using the collected datasets for development is responsible for safely storing and backing them up during the project and deleting them after the project. All the Consortium members must consult the associated partner(s) before publishing any data that can be associated with a scientific publication or exploitable result, in the open domain.

The project may generate datasets during the project activity that will be evaluated to decide their access level mainly because the publication of those datasets may interfere with the exploitation activities or when datasets include confidential information.

4.2.3.2 Embargo

As per Article 17 in the Grant Agreement partners adhere to the Open Science movement and thus metadata of deposited data will be in line with the FAIR principles and provide information about certain datasets, one of them being related to embargo.

4.2.3.3 Accessibility

Open access will be provided to research outputs like scientific papers and publications, with the publications being uploaded on arXiv.org at the time of submission or being submitted to open access journals. The general principles for handling knowledge and Intellectual Property Rights (IPR) are defined in the GA, being the ones related to accessibility specifically mentioned in its Article 16, in line with HE recommendations.

4.2.3.4 Restrictions

Restrictions on the use of data are established in the Grant Agreement, in its Articles 16, 17 and 18.

4.2.3.5 Data access committee

Due to the nature of SEDIMARK project and its use cases, there is no need to set up a data access committee. Nevertheless, the External Ethics Advisory Board that will be assembled may have a say on topics related to this aspect.

4.2.4 Metadata

4.2.4.1 Public access

As per the Grant Agreement, Article17, metadata of deposited publications and/or data must be open under a Creative Common Public Domain Dedication (CC 0) or equivalent, in line with the FAIR principles (in particular machine actionable) and provide information at least about the following: publication (author(s), title, date of publication, publication venue); Horizon Europe or Euratom funding; grant project name, acronym and number; licensing terms; persistent identifiers for the publication, the authors involved in the action and, if possible, for their organisations 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.

4.2.4.2 Availability

The digital data created by the project will be diversely curated depending on the sharing policies attached to it. For both open and non-open data, the aim is to preserve the data and

make it readily available to the interested parties for the whole duration of the project and beyond. The extension of this period of availability will align with HE guidelines.

4.3 Making data interoperable

Data Interoperability is a key pillar to set the grounds for not only a data marketplace such as the one aimed by SEDIMARK but also for a common European Energy Data Space. In fact, it is necessary to establish a common framework to enable the communication not only at consortium level but with other sister projects and other cross sector data spaces initiatives SEDIMARK may interact with.

To be interoperable, there are certain guidelines to comply with:

- 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.

4.3.1 Interoperability

The aim is to follow relevant international standards, which makes achieving interoperability possible.

For the next update of the DMAP, SEDIMARK will identify data and metadata vocabularies, standards, or methodologies to be followed in order to make its data interoperable. Where possible, the consortium plans to follow standard vocabularies for all data types present in this data set, thus allowing inter-disciplinary interoperability.

Data produced within SEDIMARK will be structured using the NGSI-LD information model. To make and enhance its data interoperable, SEDIMARK will adopt Smart Data Models (SDMs) to provide a standardized representation of data and metadata related to data quality across different domains.

SEDIMARK will assess and select the appropriate data models to enrich and accurately represent the data from its use cases. If needed, existing SDMs will be extended to cover all data and metadata properties.

SEDIMARK's primary focus is on the IoT data value chain. Therefore, in addition to adopting standards and vocabularies prominent in the IoT domain, SEDIMARK will also adopt standards that prominent and emerging data spaces such as DCAT for data catalogues and ODRL for specifying policies for asset exchange.

4.3.2 Best practices

SEDIMARK will follow community-endorsed best practices for enhancing interoperability. This includes best practices for ontology development which includes the reuse of relevant ontology concepts [Ontology Development 101: A Guide to Creating Your First Ontology].

4.3.3 Ontologies and vocabularies publication

SEDIMARK has established an information model for the SEDIMARK ecosystem, which will define an ontology that focuses on the concept of offerings, which includes aspects of participant entities involved in either the provision of assets or its consumption, the asset(s) exchangeable as part of the offering, and policies defining conditions for exchange. It will also employ a taxonomy for describing the types of assets made available for consumption. For

publication, the SEDIMARK information models will have the permanent URI prefix (https://w3id.org/sedimark/{model_name}) as the namespace, to maintain the reachability irrespective of where it will be hosted, to support the provision of the ontology in different formats, and for other ontology engineers to adopt for linking with their information models. The information model will be hosted on a server managed by a consortium partner and will also be made available via the project's GitHub repository.

4.3.4 Mappings to commonly used ontologies

SEDIMARK's approach is to first attempt to extend ontology concepts that are adopted by prominent data space initiatives. It will then attempt to map to other data space information models where concepts match or are similar, since context can differ with terms and vocabularies, depending on the theme of the data space, which in SEDIMARK is of a data marketplace nature.

4.4 Increase data reuse

As aforementioned, the different datasets in SEDIMARK will be annotated according to open-source semantic data models based on existing open standards. This will foster data discovery and re-use both within the project and outside the project.

On the one hand, SEDIMARK aims to leverage the seamless interoperability framework to re-use data from the same pilot into different use cases to develop services that otherwise could not be developed. On the other hand, SEDIMARK will enable the reuse of the datasets through the SEDIMARK marketplace, which has as a central objective the sharing of data through various parties in an easy way. As such, data generated by a data provider will be annotated properly and made available to be consumed by multiple parties and reused further. Additionally, data will be reused to build Machine Learning models for various scenarios and the models will also be annotated and made available to be shared within the SEDIMARK marketplace.

There are certain conditions to consider data reusable:

- R1. Meta (data) has 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 their provenance.
  - R1.3. (meta)data meet domain-relevant community standards.

4.4.1 Data analysis validation

When needed, data produced in SEDIMARK will be provided alongside "readme" files that facilitate data reuse in the long term. These documents will include information on the methodology followed, analyses conducted, definitions of variables employed or units of measurement.

4.4.2 Free of charge data

The strategies sketched out for SEDIMARK and its 4 use cases imply data curated and produced will be made freely available in the public domain and hence permit the widest re-use possible during the project execution and beyond. This policy does not apply to those datasets that may fall under the scope of the monetisation activities that are part of the research activities of the project, where retribution on selling data will be asked for.

4.4.3 License

Data re-use policies will be outlined, discussed and decided as the project moves forward and presented in greater detail at the next revision. Concerning licenses related to the dataset(s), in the event a certain sample is not openly available, the project will consider Attribution-Non-Commercial 4.0 International (CC BY-NC 4.0) [15] license when it comes to increasing the possibility of re-use. Once again, this will be carefully addressed in the event certain datasets are available for purchase.

4.4.4 Third Parties usability

As per Grant Agreement Article 16, data produced in the project will be usable by third parties, not only during the project execution timeframe but also after SEDIMARK comes to an end.

4.4.5 Standards

Data generated in SEDIMARK will have clear license and provenance, being thoroughly documented using the standards considered most appropriate at the time of the project execution.

4.4.6 Data quality assurance

The Data Quality Assurance Process is the responsibility of the whole consortium and will be done according to the guidelines specified in this report.

This process will be supervised by the Project Board (PB) and will consist of the following steps:

- Establishment of the way in which the collection and processing of data is managed.

  This should be done by distinguishing the different types of data managed in the project: data collected or generated in the different use cases, data reused from other EU project initiatives or data from project communication and dissemination activities.
- Definition of the responsibilities of each of the roles involved in the process.
- Determination of the quality parameters according to the six main dimensions that are commonly used to mark the quality of the same: accuracy, completeness, consistency, validity, uniqueness, and timeliness.
- Implementation of the set of technical tools and Key Performance Indicators (KPIs) that will be used to measure and ensure the quality of the data. The type of tools to be employed will be defined throughout the project according to its needs.

The aforementioned data quality process is related to the non-use-case data and includes the consortium and project data i.e. deliverables. SEDIMARK also works on improving the quality of datasets generated in the use cases of the project. In this case, Data quality is tackled via the AI-based data quality improvement tools, removing unnecessary data, reducing the communication exchanges between data provider and consumer (resulting in smaller dataset sizes since only necessary data will be exchanged), leading to less energy consumption. This means ratio of errors to data will be minimized, data completeness will be maximised, and validity of data improved.

5 Other research outputs

The non-open research data will be archived and stored long-term in the ownCloud repository administered by ATOS. This ATOS platform is currently being employed to coordinate the project's activities and to store all the digital material connected to SEDIMARK. If certain datasets cannot be shared (or need restrictions), legal and contractual reasons will be explained.

Moreover, the software modules that will be implemented in the SEDIMARK project will be stored within the GitHub Organization that has been created at https://github.com/Sedimark. During the development phase, the respective software projects will be kept private, but SEDIMARK is following Open Science practices and source code will be opened as soon as stable releases of the modules are ready.

Finally, AI models that will be developed during the project will also be made available through a dedicated model registry such as Neptune.ai [16].

Based on some portion of the data provided by Mobility Digital Twin Helsinki, which refers to images that are captured from multiple webcams that are placed on different roads and highways across Finland (can be found here https://tie.digitraffic.fi/api/weathercam/v1/stations/C01503/history), so from those data, a use case can be made that wants to use Computer Vision and Deep Neural Networks models, for example Yolo Models or models provided by Hugging Face, which are open source, which will be finetuned on the data mentioned before, to detect and count the number of cars, trucks and vehicles overall, that are in the frames offered by the webcams API, and with those numbers of vehicles and the timestamp of the image, some predictive maintenance can be achieved, which refers to the ability to predict in the future what are the most busy roads near Helsinki, and based on this prediction to see if some actions are needed to be taken or not.

Another use case of SEDIMARK is the forecast of water flow levels. This focuses on the Lac de Serre-Ponçon area, where there are a number of water stations. By analyzing historical data collected from several water stations on the lake's course, the goal is to predict future water flow levels at the dam's exit. The data contains recordings from 1960-2023. This use case takes into consideration upstream historical data to forecast daily water flow measurements at the dam's exit.

6 Allocation of resources

6.1 Costs

ATOS, as project coordinator, allocated a part of the overall WP1 budget and person months to activities related to making data or other research outputs FAIR in SEDIMARK. This means, direct and indirect costs in this context will relate to storage, archiving, re-use, security, etc. All partners in the consortium have resources allocated for participating in Task 1.3 “Data management.”

Costs related to open access to research data are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions).

6.2 Data management responsible

Data management in SEDIMARK will be done as part of T1.3 “Data Management” within WP1 “Project Management” and ATOS, as project coordinator, will be responsible for data management in SEDIMARK project.

Hence, and for the time being, the project coordinator is responsible for FAIR data management within SEDIMARK.

6.3 Data preservation

Resources for long-term preservation, associated costs and potential value, as well as how data will be kept beyond the project and for how long, will be discussed by the SEDIMARK consortium during upcoming General Assembly meetings.

7 Data security

The achievement of data governance and privacy in SEDIMARK will come through (1) guaranteeing the right to control the data at any time for data owners, (2) monitoring and controlling the access to the data and (3) preventing unauthorized data access.

7.1 Provisions

The coordinator, acting as responsible for data management, along with all partners in the consortium, will work to ensure Data Security throughout the life of the SEDIMARK project and for a certain period thereafter.

This group will monitor activities or results that may cause security issues; it will manage those issues and ensure that provisions are in place for data security throughout the life of the project. This will include both the internal file repository and pilot data of the project. The requirements for role-based access will be defined and implemented to protect from unauthorized access, use, modification, and destruction of the data. Redundant data storage will be included to secure data recovery. Regarding pilot data, they are not yet defined, and this will be reviewed in the upcoming iterations of the DMAP.

7.2 Storage

For the duration of the project, datasets will be stored on the responsible partner's storage system. Every partner is responsible for ensuring that the data are stored safely and securely and in full compliance with European Union (EU) data protection laws. After the completion of the project, all the responsibilities concerning data recovery and secure storage will go to the repository storing the dataset.

All data files will be transferred via secure connections and in encrypted and password-protected form. As an example, the use of the open-source 7-zip tool [17] providing full AES-256 encryption will be promoted. Moreover, the encryption options implemented in MS Windows or MS Excel may be used as well. Associated passwords will not be exchanged via e-mail but in personal communication between the partners involved.

8 Ethics

This section deals with ethical and legal compliance issues, like the consent for data preservation and sharing, protection of the identity of individuals and companies and how sensitive data will be handled to ensure that they are stored and transferred securely.

Given the fact that SEDIMARK receives funding under the Horizon Europe programme, the project will comply with ethical principles and relevant Union, national and international legislation, including the Charter of Fundamental Rights of the European Union and the European Convention on Human Rights and its Supplementary Protocols.

In addition to these guidelines, interactions with research participants and the subsequent data collection will be compliant with "The EU General Data Protection Regulation (GDPR)" [3]. Since the consortium includes a non-EU member from the UK, as of 1st of March 2024, the UK GDPR still maintains EU GDPR adequacy since the last decision published on 28th June 2021 [20].

8.1 Possible ethical issues

Data protection and good research ethics are major topics for the consortium of this project. Good research ethics meet all actions to take great care and prevent any situation where sensitive information could get misused. This is what the consortium wants to guarantee for this project. Research data which contains personal data will just be disseminated for the purpose for which it was specified by the consortium. Furthermore, all processes of data generation and data sharing must be documented and approved by the SEDIMARK consortium to guarantee the highest standards of data protection.

SEDIMARK partners must comply with the ethical principles as set out in Article 34 of the Grant Agreement, which states that all activities must be carried out in compliance with:

- Ethical principles (including the highest standards of research integrity as set out, for instance, in the European Code of Conduct for Research Integrity and including, in particular, avoiding fabrication, falsification, plagiarism or other research misconduct).
- Applicable international, EU and national law (in particular, EU Directive 95/46/EC).

In case citizens will be involved in research and innovation activities in the project pilot sites, compliance will not only be sought with the above-mentioned documents but also with the relevant local and national regulations and ethical guidelines.

No ethical issues are envisaged so far, nevertheless, this subject will be monitored on a regular basis during the project's life cycle as the data set develops and grows.

Attention to these ethical issues should not fall exclusively on the members of the consortium, who are ultimately interested parties in the process. That is why it is advisable to have opinions from voices external to the project, with proven experience in the project's work areas, capable of evaluating with a critical and well-trained eye the work carried out by SEDIMARK to preserve an ethical approach. These people make up the so-called Ethics Board (EB).

In the context of SEDIMARK, such Ethics Board will proactively support the development of the project, act as a knowledge and guidance forum, and provide advice to the consortium on how to exploit knowledge created by the project.

Hence, SEDIMARK partners will invite the Ethics Board members on a regular basis to discuss ethical aspects considering the project evolution, keeping them well informed of all relevant publications and/or events SEDIMARK-related.

This Ethics Board was appointed by the project coordination team and approved by the project board (PB). The EB will ensure compliance with relevant and ethical roles throughout the different stages of SEDIMARK implementation and will assist in the elaboration of the reports associated with WP7, dealing with ethics topics.

8.2 Consent for data sharing and long-term preservation

Data collection, storage, protection, retention and destruction will be carried out through the intranet system of the SEDIMARK project. Interviewees/beneficiaries/recipients will be informed about data security, anonymity and use of data as well as asked for in accordance. Participation happens on a voluntary basis and via the completion of informed consent prepared accordingly.

8.2.1 Informed consent

Most social science research endeavours are such that human participation requires evidence of the voluntary, free, and informed consent of those who contribute their time, insights, effort and data for the use of researchers. Informed consent, whether in writing (as is most usual) or given orally, is thus the default option. Obtaining informed consent, however, does not in itself guarantee ethical research. In some research settings, this very act and the aim of safeguarding participants' rights and well-being in the research setting may place them at risk of harm in their social context (rather than affording them protection).

When preparing information sheets and consent forms, SEDIMARK consortium will follow a checklist based on common Guidelines for Ethics:

- Give participants a clear explanation of the aims, overall purpose, methods and implications of the research.
- Explain that participation is voluntary.
- Remind participants that they have a right to withdraw their consent at any time without any consequences.
- Explain the degree of benefit, risks, burden or discomfort involved in participation. Give an estimate of the time and effort expected of participants.
- Explain precautions to ensure participants' safety and provide information on insurance if there is any.
- Explain who is funding the research and for what purpose.
- Disclose who will benefit from the research.
- Give a firm commitment to protecting respondents' anonymity and privacy (provided that this can genuinely be guaranteed).
- Make a clear commitment to treating personal and sensitive information confidentially.
- Reassure participants that there are secure procedures for analysing any data gathered.
- Explain clearly who will have access to the data provided by participants. Consider any unintended/unexpected/incidental findings and explain how you intend to deal with such findings.
- Explain briefly where research findings will be published.
- Offer to provide respondents with further information about research if they ask for it.

- Give the name and contact details of the contact person who can answer any queries participants may have.
- Clarify possible uses to which data may be put in future (if this is envisaged) and clarify whether participants will be asked for consent again if this is the case. Cover any issues relating to copyright of data and other materials used in the research.

SEDIMARK will create a consent form to be used across the pilot demonstration sites that safeguard citizens and users participating in SEDIMARK activities.

8.3 Ethics and AI

The European Commission has established a High-Level Expert Group on AI which in 2019 published a set of guidelines for Trustworthy Artificial Intelligence [18]. The main goal for such guidelines was to promote Trustworthy AI and build a framework for the research in EU projects to achieve trustworthiness. The guidelines were built on the concept that AI systems should be human-centric, serving the humanity and improving human welfare and freedom. Thus, AI systems should be built in a way to prevent and minimise risks. According to the guidelines, Trustworthy AI has three main components: (i) to be lawful, (ii) to be ethical and (iii) to be robust.

SEDIMARK aims to use machine learning models to improve the intelligence of the data sharing system. ML components will be used mainly on the data curation and quality improvement pipeline, as well as on providing recommendations for datasets/services to the users of the system. To ensure that SEDIMARK's components abide by the AI guidelines, SEDIMARK aims to use the Assessment List for Trustworthy Artificial Intelligence (ALTAI) for its components. The goal is to perform a self-evaluation on the components to ensure they are ethical and trustworthy. ALTAI identifies 4 ethical principles and 7 requirements that AI research should follow to achieve trustworthiness.

The 7 requirements are:

- Human Agency and Oversight,
- Technical Robustness and Safety,
- Privacy and Data Governance,
- Transparency,
- Diversity, Non-discrimination and Fairness,
- Societal and Environmental Well-being,
- Accountability.

SEDIMARK is preparing a more detailed version of the assessment list requirements and how the project partners should address them and will be included in the next version of this deliverable.

9 Other issues

SEDIMARK consortium will adhere to national legislations and adapt accordingly in discussions with each individual partner. It is the responsibility of each partner to identify and follow national legislation taking action where and when required.

SEDIMARK will be run in both the European Union (EU) and the United Kingdom (UK). The entire storage is in the European continent region; therefore, SEDIMARK will be keeping all its data in countries with "adequate protection." Users within both spaces will be served by servers located in the EU, and the data collected from these users will also be stored in EU-based servers. Moreover, EU-based subjects will not have any means to access any other servers outside their continent, hence providing safe means to capsulate EU data in the continent.

9.1 Procedures and criteria used for Dissemination Activities

Participation of people in SEDIMARK **dissemination activities** is related to all the promotion and dissemination means used in the project. SIEMENS is the leader of the WP6 – "Dissemination, exploitation, standardisation" and therefore they will be responsible for implementing the procedures to comply with the ethical requirements.

9.1.1 Public Website

In correspondence with the project commitments, a dedicated project website has been launched. The project website will be openly available to all internet users worldwide. In line with the GDPR, the website also hosts a **legal notice**, a **cookie policy** and a **credits page.**

9.1.2 Distribution of promotional material: leaflet, newsletter, video, booklet

9.1.2.1 Newsletter

A dedicated SEDIMARK newsletter template was developed in line with the established visual identity.

It is possible to subscribe to the SEDIMARK newsletter directly from the project website.

As an alternative, another way to subscribe to the SEDIMARK newsletter is to provide the email address by signing a dedicated form ("Newsletter Consent" template) that the partners of the project can bring with them when they attend events, conferences and fairs where they have the chance to present the SEDIMARK project. Interested attendees can sign the form and give their consent to be informed and updated about the project receiving the newsletter issues. The subscriber, after completing the form, will receive a confirmation email to fully accept the subscription to the SEDIMARK newsletter.

The newsletter offers the users the possibility to unsubscribe from the service; in this case, the user by clicking the unsubscribe button included in the newsletter template, will have the chance to withdraw his/her consent to receive the subsequent issues.

9.1.2.2 Leaflet, Video and Booklet

Important outcomes from SEDIMARK will include the project`s leaflet(s), video and booklet, as they will be key educational tools for the consortium to actively disseminate. The project leaflet, video and booklet will all be displayed on the project website and advertised through social media, events and conferences etc. All dissemination of the leaflet, video and booklet will be done in accordance with the GDPR and other relevant rules and regulations mentioned throughout this deliverable.

9.1.3 Social Media Campaign

SEDIMARK will utilize social media as an integral part of the project´s communication and dissemination activities, in particular Twitter and LinkedIn accounts. The primary step was the founding of a hashtag (#), which will be used on both social media platforms ("#SEDIMARK"). UC, as Task 6.1 "Dissemination and Impact creation" leader, will advise consortium members on best practices for social media. Such best practices include proper and ethical use of human data always in accordance with the GDPR rules. SEDIMARK will not ask for or record personal information through social media.

9.1.4 Events and Workshops

SEDIMARK communication strategy is also based on the participation of the members of the consortium in events, which can be organised internally (project consortium meetings, workshops), but also in external social events, such as specialised international data spaces conferences and fairs, seminars, and clustering events to network with other EU-funded projects. It may happen that, during these events (organised by consortium members or external stakeholders), photographs or videos might be taken for communication purposes on the project website or realisation of posts on project social media. In this framework, SEDIMARK consortium members apply GDPR strict rules for processing personal data (including images) based on consent. Consent of the data subject means any freely given, specific, informed, and unambiguous indication of the data subject's wishes by which he or she, by a statement or by clear affirmative action, signifies agreement to the processing of personal data relating to him or her (GDPR definitions,11). When someone consents to process his/her personal data, SEDIMARK consortium will process only the data for the purpose for which consent has been given. Moreover, the consortium is aware that the opportunity of consent withdrawal is always given and applicable.

SEDIMARK will prepare a dedicated form to collect consent to take pictures, videos, and audio for communication purposes during dedicated events that the project consortium may organize.

9.2 Procedures and Criteria used for Research Activities

Participation of people in SEDIMARK research activities is mainly related to stakeholders' participation in the surveys and interviews that are planned to be conducted in different tasks of the project. All these activities will consider all the GDPR and Ethics aspects explained in the previous sections.

9.3 Procedures against data losses

In the next years, GDPR laws regarding data usage will become even stricter. This means SEDIMARK partners must be ready to delete all data from customers from all storages the moment they require the consortium to perform a complete deletion of their data.

Moreover, partners in the consortium must make sure their systems (e.g., algorithms, AI models) are robust enough and can deal with data losses produced by such data deletion processes.

- Conclusions

The project recognises the importance of regulating data management issues and this document, the tools, and the procedures here described will evolve and be updated as

SEDIMARK continues and faces new challenges, especially once the pilot site demonstrations start to be active.

In this second version of the deliverable, the document outlines the principles and procedures of the SEDIMARK Data Management Plan (DMP). The SEDIMARK DMP is guided by three core principles:

- **Simplicity (KISS):** Both the management of data within the project and the processes for discovering, accessing, and reusing data during and after the project will be kept as straightforward as possible.
- **FAIRness:** All data utilized in the project must comply with the principles of being Findable, Accessible, Interoperable, and Reusable.
- **Societal Responsibility:** Data generated by the project will be made publicly available unless there are compelling ethical or commercial reasons against such dissemination.

The KISS principle is primarily implemented by assigning the responsibility of tracking the assessment, usage, or production of data within the project to the relevant tasks where such data is involved. Additionally, quality control is integrated into the overall review process of deliverables and data production.

Achieving FAIRness is predominantly accomplished by publishing all metadata pertaining to the data generated in the project on Zenodo (or similar long-term repository finally designated by the project). Moreover, as a general practice, all data produced in the project will be released as Open Data (unless indicated the contrary), either as part of the Zenodo record or through a separate Open Data service or repository with ensured long-term sustainability independent of the project.

Lastly, societal responsibility is addressed by establishing a mechanism to evaluate any commercial, ethical, legal, or security concerns regarding data publication for each SEDIMARK deliverable.

10 References

[1] European Commission, “Horizon 2020 Online Manual, Cross-cutting issues, Open access & Data management,” European Union, [Online]. Available: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm.

[2] European Commission, “Open access & Data management. Open access to research data,” EC H2020 Online manual, [Online]. Available: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm.

[3] EUR-Lex, “Regulation (EU) 2016/679, General Data Protection Regulation,” 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[4] Cornell University, “arXiv,” [Online]. Available: https://arxiv.org/.

[5] European Commission, “Data Management Plan Template for Horizon Europe.,” 11 05 2022. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan_he_en.docx.

[6] ownCloud GmbH, “ownCloud - share files and folders, easy and secure,” [Online]. Available: https://owncloud.com/. [Accessed 03 2023].

[7] M. D. M. A. I. e. a. Wilkinson, ““The FAIR Guiding Principles for scientific data management and stewardship,” Sci Data, vol.3,” Sci Data 3, 160018 (2016). [Online]. Available: vol. https://doi.org/10.1038/sdata.2016.18.

[8] “FAIR Principles,” GO FAIR, [Online]. Available: https://www.go-fair.org/fair-principles/. .

[9] European Commission, “Guidelines on FAIR Data Management in Horizon 2020,” 2016. [Online]. Available: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

[10] W3C, “DCAT Application Porfile for data portals in Europe - Version 3.0,” February 2024. [Online]. Available: https://semiceu.github.io/DCAT-AP/releases/3.0.0/. [Accessed 05 03 2024].

[11] W3C, “Resource Description Framework (RDF),” [Online]. Available: http://www.w3.org/standards/techs/rdf.

[12] W3C, “Web Ontology Language (OWL),” [Online]. Available: http://www.w3.org/standards/techs/owl.

[13] Context Information Management (CIM) ETSI Industry Specification Group, “ETSI GS CIM 006 V1.2.1. NGSI-LD Information Model,” June 2023. [Online]. Available: https://www.etsi.org/deliver/etsi_gs/CIM/001_099/006/01.02.01_60/gs_CIM006v010201p.pdf. [Accessed 05 03 2024].

[14] CERN Data Centre & Invenio., “Zenodo,” 09 02 2023. [Online]. Available: https://zenodo.org/.

[15] “git,” Software Freedom Conservancy, [Online]. Available: https://git-scm.com/.

[16] International DOI Foundation, “The DOI® System ISO 26324,” [Online]. Available: http://www.doi.org.

[17] creative commons, “Attribution-NonCommercial 4.0 International (CC BY-NC 4.0),” [Online]. Available: https://creativecommons.org/licenses/by-nc/4.0/deed.en. [Accessed 09 02 2023].

[18] N. Labs, “Neptune.ai - Log, organize, compare, register, and share all your ML model metadata in a single place,” 2022. [Online]. Available: https://neptune.ai/.

[19] I. Pavlov, “7-Zip,” [Online]. Available: http://www.7-zip.org/.

[20] European Commision, «Adequacy decisions. How the EU determines if a non-EU country has an adequate level of data protection.,» [En línea]. Available: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en#adequacy-decisions-latest.

[21] A. H.-L. E. Group, “‘Ethics guidelines for trustworthy AI’,” B-1049, Brussels, 2019.

[22] E. Commission, «Adequacy decisions. How the EU determines if a non-EU country has an adequate level of data protection.,» [En línea]. Available: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en#adequacy-decisions-latest.