D8.1

Data Management Plan

IT4I@VSB

Public

Funded by
the European Union

Deliverable Properties

Version	v1.0
Dissemination level	PU
Deliverable type	DMP
Work package	WP8
Task	T8.2
Due date	30 June, 2023
Submission date	30 June, 2023
Deliverable lead	IT4I@VSB
Authors Name	Kateřina Slaninová (IT4I@VSB), Stephan Hachinger (BADW-LRZ), Piyush Harsh (TERRAVIEW), Vojtěch Mlýnský (IT4I@VSB), Josef Šivic (CVUT), Antonín Vobecký (CVUT), David Hurych (VALEO), Jan Zahradník (Valeo), Jérôme Freani (ALTRNATIV)
Reviewers	Pinar Karagoz (METU), Marc Derquennes (EURAXENT)

Abstract

The Data Management Plan lays out our planning for handling main aspects of the life cycle of the project data (data organisation and long-term storage, access, preservation, and sharing). This document also includes a preliminary specification of outputs (what data will be generated during the project). It is a living document and will be continuously updated during the project.

Keywords

Data management plan; FAIR; Metadata; Repository

Document Revision History

Version	Date	Description of Change	Contributor(s)
v0.1	10 May, 2023	Table of Content	Kateřina Slaninová (IT4I@VSB)
v0.2	15 May, 2023	1st version of the deliverable	Kateřina Slaninová (IT4I@VSB)
v0.3	25 May, 2023	Included first input from application cases	Kateřina Slaninová (IT4I@VSB), Piyush Harsh (TERRAVIEW), Vojtěch Mlýnský (IT4I@VSB), Josef Šivic (CVUT), Antonín Vobecký (CVUT), David Hurych (VALEO), Jan Zahradník (Valeo)
v0.4	29 May, 2023	Included input from SME application case - health	Kateřina Slaninová (IT4I@VSB), Jérôme Freani (ALTRNATIV)
v0.5	5 June, 2023	Updated Sections 1, 2, and inputs from application cases	Stephan Hachinger (BADW-LRZ), Vojtěch Mlýnský (IT4I@VSB), Antonín Vobecký (CVUT), Piyush Harsh (TERRAVIEW)
v0.6	9 June, 2023	Updated Sections 2, 5 and inputs from application cases	Jérôme Freani (ALTRNATIV), David Hurych (VALEO)
v0.7	20 June, 2023	Updated Section 1.2 and Section 6 by comments from Ethical advisor	Kateřina Slaninová (IT4I@VSB)
v1.0	30 June, 2023	Final check by the coordinator	Jan Martinovič (IT4I@VSB)

Disclaimer

Information, documentation, and Figures available in this deliverable are provided by the EXA4MIND project's consortium funded by a European Union's Horizon Europe Research and Innovation programme under grant agreement No **101092944**. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Copyright Notice

Dissemination Levels

PU	Public	Public, fully open, e.g. website
SEN	Sensitive	Confidential to EXA4MIND project and Commission Services

Deliverable Types

R	Document, report (excluding periodic and final reports)
DEM	Demonstrator, pilot, prototype, plan designs
DEC	Websites, patent filings, press and media actions, videos, etc.
OTHER	Software, technical diagrams, etc.

Table of Contents

Executive Summary **9**

1 Introduction **10**
1.1 Integration with the European FAIR Data Ecosystem 10
1.2 Project Hardware and Storage 11
1.3 Allocation of Other Resources 12

2 Relevant Data - Summary **12**
2.1 WP4 - Scientific Application Case 13
2.2 WP5 - Industry Application Case 13
2.3 WP6 - SME Application Case - Agriculture 14
2.4 WP6 - SME Application Case - Health 15

3 FAIR Approach **16**
3.1 Repository 16
3.2 FAIR Data 16
3.2.1 Findability 17
3.2.2 Accessibility 17
3.2.3 Inter-operability 20
3.2.4 Re-usability 20

4 Other Outputs **20**

5 Security **21**

6 Ethics **22**

7 Data Outputs Plan **22**
7.1 Scientific Application Case 22
7.2 Industry Application Case 23
7.3 SME Application Case - Agriculture 24
7.4 SME Application Case - Health 26

8 Conclusion **27**

9 References **28**

List of Figures

No figures are given.

List of Tables

Table 1 Making Data Findable in Application Cases - Metadata 18
Table 2 Making Data Accessible in Application Cases 19
Table 3 Scientific Application Case - Output DS01 23
Table 4 Industry Application Case - Output 01 24
Table 5 Industry Application Case - Output 02 24
Table 6 SME Application Case - Agriculture - Output 01 25
Table 7 SME Application Case - Agriculture - Output 02 25
Table 8 SME Application Case - Health - Output 01 26
Table 9 SME Application Case - Health - Output 02 26
Table 10 SME Application Case - Health - Output 03 27
Table 11 SME Application Case - Health - Output 04 27

Abbreviations

ADAS	Advanced Driving Assistance System
AI	Artificial Intelligence
API	Application Programming Interface
DBMS	Database Management System
DSSC	Data Spaces Support Centre
DMP	Data Management Plan
DoA	Description of Action
DOI	Digital Object Identifier
EDD	Extreme Data Database
EO	Earth Observation
EOSC	European Open Science Cloud
EUDAT	European Research Data Management and Storage Service
FAIR	Findable, Accessilbe, Interoperable, Re-usable
GDPR	General Data Protection Regulation
HPC	High Performance Computing
HPDA	High Performance Data Analysis
IP	Intelectual Property
IPR	Intelectual Property Rights
iRODS	Integrated Rule-Oriented Data System
MD	Molecular Dynamics
ML	Machine Learning
PID	Persistent Object Identifier
SME	Small and Medium Enterprise
WP	Work Package

Table of Partners

Short Name	Partner
IT4I@VSB	IT4Innovations at VSB – Technical University of Ostrava (IT4Innovations) https://www.it4i.cz/en
BADW-LRZ	Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities https://www.lrz.de/english/
METU	Middle East Technical University (METU) https://www.metu.edu.tr/
AUSTRALO	AUSTRALO Interinnov Marketing Lab https://www.australo.org/
EURAXENT	Euraxent https://www.linkedin.com/company/euraxent/
CVUT	Czech Institute of Informatics, Robotics and Cybernetics at the Czech Technical University (CIIRC CTU) https://www.ciirc.cvut.cz/
VALEO	Valeo https://www.valeo.com/en/valeo-ai/
TERRAVIEW	Terraview https://www.terraview.co/
ALTRNATIV	Altrnativ https://altrnativ.com/?lang=en

Executive Summary

This Data Management Plan (DMP) specifies the main aspects of the life cycle of the project data (organisation and their long-term storage, access, preservation, and sharing). This document also includes a preliminary specification of outputs (what data will be generated during the project). This document will be continuously updated during the project. The final version of DMP will be submitted in M36 as D8.3 (EXA4MIND 2025b). The DMP will be published on one of the trusted repositories designated for long-term publication in this project (see Section 3.1).

DMP contains a discussion of the most relevant aspects of the EXA4MIND Platform and of its role in relation to the FAIR (Findable, Accessible, Interoperable, Re-usable) approach to working with datasets in the project. The relation of this to the integration of the platform with the European FAIR data ecosystem is laid out. The main part of the deliverable contains a summary of datasets which will be used in particular application cases in the project (see Section 2). The project's approach to complying with the FAIR data principles is described, together with a recap of the principles themselves, in Section 3. Then, working with other outputs (mainly software), security issues and ethical aspects are briefly discussed in Sections 4, 5, and 6. The last section (Section 7) contains plans for data outputs which will be provided in the project. The outputs are again divided into sections according to particular application cases. Section 8 concludes the document.

1 Introduction

The goal of the EXA4MIND project is to build a platform for Extreme Data that enables advanced data analytics on Supercomputers and automated data management providing the support for integration with European Open Science Cloud (EOSC)[1] and European Data Spaces by design, with the FAIR principles (Findable, Accessible, Interoperable, Re-usable – Wilkinson et al. 2016) in mind. Therefore, appropriately working with large amounts of data is essential for the project.

The EXA4MIND project will create a modular **Extreme Data Database** (EDD) analytics environment. Driven by four representative Application Cases from the fields of molecular dynamics (Scientific application case), autonomous driving (Industry application case), smart agri-/viticulture (SME Agricultural application case) and health / social (SME Health application case), EXA4MIND will support data ingestion and versatile preprocessing, develop an **Advanced Query and Indexing System** with convenient interfaces, and leverage high-performance storage and computing infrastructure. Integration of databases with supercomputing and with European approaches to FAIR distributed data is a strategic part of the project, as indicated above.

1.1 Integration with the European FAIR Data Ecosystem

An essential feature of EXA4MIND, which will be reflected in our work with data and in the data management plan, is that proper data management and related developments are the central point of the project and not just a side aspect. Working with data in accordance with the FAIR principles is just as essential for the project as strong integration with the European FAIR Data Ecosystem (including EUDAT[2], EOSC, and European Data Spaces[3]).

The project deliverables and milestones, as they are planned, highlight the support for these goals. Apart from the deliverables D8.1 and D8.3 (EXA4MIND 2023c, EXA4MIND 2025b) on the Data Management Plan, we plan to publish also a “Report on FAIRification of Database Data” (EXA4MIND 2025a) in M36. There will be continuous work on the aspects discussed throughout the project, which is obvious from MS6 (M24, focused on the integration of distributed data sources, HPC and FAIR ecosystems with the EDD – EXA4MIND 2024c), MS10 (M20, focused on pre-processing and metadata for FAIR AI – EXA4MIND 2024b), or MS7 (M30, focused on the publication of operation guidelines of distributed FAIR data and EDD – EXA4MIND 2025c).

Within the project, we will on one hand facilitate data publication, secure, FAIR and open data handling via well-established systems. On the other hand, we will provide a bridge from the EDD to European Data Ecosystems in order to publish research data products.

[1] EOSC: https://eosc.eu
[2] EUDAT: https://eudat.eu
[3] Data Spaces Support Centre: https://dssc.eu/

The essential mechanisms employed for this are:

- Mechanisms which will allow for replication or transfer of all kinds of data to our EUDAT-B2SAFE[4]/iRODS-based (cf. EUDAT website and Xu et al. 2017) distributed data system as it is part of the EDD, and
- Mechanisms which will allow **direct access to published/immutable database data**, as far as possible. These will involve a management database, assigning unique identifiers to queries which retrieve data from the EDD's database management systems (DBMS).

In both cases, referencing datasets with a unique, persistent identifier is essential and will be guaranteed through the usage of EUDAT-B2HANDLE[5]. In selected cases (e.g. important output data), also DOIs will be assigned. All functionalities just sketched can be triggered via REST APIs in the final system, just as data-retrieval will be possible via our REST APIs and/or further standard interfaces (e.g., S3 – Amazon Web Services, Inc. and affiliates 2023).

These mechanisms guarantee interoperability with EUDAT and thus with the EOSC (cf. DICE project, Horizon Europe GA 101017207), and will be used to connect the EXA4MIND EDD with the European Data Spaces as well. We are currently starting discussions with the Data Spaces Support Centre (DSSC) in order to understand which European Data Spaces (Agriculture, Industry/Manufacturing, etc.) to target first and how we have to design the adaptors to interface with them. Our aim is a convenient and efficient exchange of data (sending and retrieval) between the EDD and these European data ecosystems.

1.2 Project Hardware and Storage

Within the EXA4MIND project consortium, we have two HPC centres (European and national level): BADW-LRZ and IT4I@VSB; both have significant experience in the operation of multi-petascale computational resources at Tier-0 and Tier-1 levels.

These support diverse hardware architectures from a range of vendors and will provide a range of environments within which to prepare the platform, the applications and the data for successful deployment, verification and validation. These sites also have significant experience in optimising application performance for their particular architectures. The HPC centres also have significant data storage capacity, which is essential to further support the efforts of the project. European HPC centres provide such resources through national and EuroHPC JU access calls. A summary of the key HPC resources and data storage is available at BADW-LRZ[6] and IT4I@VSB[7] websites. The EXA4MIND project builds upon a technical basis of existing data and computing infrastructure from LEXIS project and its outcome LEXIS Platform[8] (LEXIS 2023).

[4] EUDAT-B2SAFE: https://www.eudat.eu/b2safe
[5] EUDAT-B2HANDLE: https://eudat.eu/catalogue/b2handle
[6] BADW-LRZ documentation: https://doku.lrz.de
[7] IT4Innovations documentation: https://docs.it4i.cz
[8] LEXIS Platform: https://portal.lexis.tech

EXA4MIND is a research and innovation project and it is planned to have outputs on Technology Readiness Level 5 (Enspire.science 2023). Therefore, it is expected that EDD will be used for the validation purposes throughout the whole project duration including benchmarks in each application case. Other trusted repositories will be used for the publication of outputs (especially long-term extreme data) during the project where possible, see Section 3.1. The data-publication capabilities of the EDD as we expect them to be developed during the project will complement this.

1.3 Allocation of Other Resources

The appropriate effort for the creation of the data management plan was allocated within Task 8.2 “Quality Assurance and Data Management Plan”. Also, continuous updates will be covered during the project in cooperation with all partners. All partners will work with data, metadata, and other outputs according to the defined lifecycle and FAIR principles.

The definition of ethics requirements that the project must comply with will be set up within the work package WP9 “Ethics Requirements”. An external independent Ethics Advisor specialised in EU data protection law and the ethics of AI was appointed in M2. This deliverable, especially Section 6 was consulted with Ethics Advisor and will serve as an input for Deliverable D9.1 (EXA4MIND 2024a).

2 Relevant Data - Summary

This section describes input data as well as output data which will be used within the project to create and validate the platform. Different types of data will be used within the project according to the variety of use cases. The project is in its initial phase (M6). Currently, no datasets have been generated from the beginning of the project. Therefore, the following tables describe the initial plan of datasets which will be used in the project and will be updated during the project.

All work packages will work with data during the project; only WP8/9 will merely create management data. While WP4, WP5 and WP6 obviously have a thematically-oriented data input and output, also WP1, WP2 and WP3 work with relevant inputs and outputs to be managed. WP1 will provide documentation of co-design and technical architecture. WP2 and WP3 will generate software with appropriate documentation related to distributed data space management and data analytics and processing.

Project partners also consider that the research outputs will be published in peer-reviewed journals and disseminated at appropriate conferences. The journal and conference papers, as well as data pertaining to relevant experiments, will be stored at long-term repositories respecting the FAIR principles (see Section 3). Also, the Data Management Plan itself will be provided as a public document.

It is obvious from the description of application cases that they will work with extreme datasets and would like to use the EXA4MIND platform to solve their challenges in this. A variety of application cases has been chosen within WP4 - WP6, for the EDD to be put on the test as an ambitiously generalised solution for extreme data problems. Thus, the data used and generated within each of the application cases

bear some further discussion. The following subsections will accordingly describe each application case and the datasets planned to be processed within WP4 - WP6.

2.1 WP4 - Scientific Application Case

The scientific application case will re-use existing data from molecular dynamics (MD) simulations for additional (detailed) analysis, comparison with experimental datasets, and application of the re-weighting approach to design/test new force fields. Also, new datasets will be generated from MD simulations and HPDA as intermediate results and output datasets.

Purpose of the data generation or re-use

The purpose of data generation or re-use is the automatic tuning of force-field parameters, comparison and validation (against experimental datasets) of different force-field versions based on a huge database of MD simulations. The database will be continuously updated and expanded with new datasets.

Type and size of data

Scientific application cases will work with ASCII or BIN files. The files could be compressed to archives (for example zip/tar).

The expected size of data is the following:

- Simulation output (trajectory, topology, output file; up to 1TB of data for the specific system),
- Metadata - datasets from post-processing analysis of simulations (files with structural parameters; up to 1 GB of data for each parameter),
- Metadata - experimental datasets (up to 10 MB for each file), outputs from the re-weighting approach (database of weights; up to 10 MB for each file).

Origin of data and potential users

Potential users are members of the MD simulation community and force-field developers. Members of the MD simulation community are expected to upload simulation outputs, force field developers will be able to upload new force fields and launch the re-weighting process to validate them.

2.2 WP5 - Industry Application Case

VALEO will re-use public datasets in the application case as well as VALEO in-house recorded datasets. VALEO data will be iteratively reused for model learning and KPI (Key Performance Indicators) evaluation.

VALEO and CVUT will train their models on existing publicly available datasets, such as the nuScenes dataset[9] mainly for publications preparation, as well as on VALEO data for the final industry use case.

Purpose of the data generation or re-use

VALEO will use the data to develop and validate ADAS (Advanced Driving Assistance Systems), automated generation of precise Ground Truth and discovery of rare

[9] nuScenes dataset: https://www.nuscenes.org

situations. Also, VALEO will focus on AI methods for open-vocabulary image querying to allow smart data exploration. CVUT will generate pre/pseudo-annotations using the developed models.

Type and size of data

VALEO will use different types of data for Data containers (Elektrobit/ADTF DAT file, PCAP, MDF4, HDF5, npz - numpy format, etc.) and for Sensor specific formats & metadata (PCD, *.lvx, CSV, XML, JSON, MPEG4, JPEG, PNG, etc.). Regarding the AI part of the application case, CVUT will work with images (.jpg, .png) and point clouds (.pcd files) as the input. The outputs are trained models saved in the framework-specific format, e.g., .pth files for the PyTorch framework. The data size will be from 2.5 to 3 PB data for inference and some self-supervised models training and hundreds of TB at the remaining AI models development part. The resulting extracted metadata can easily double the overall amount of data.

Origin of data and potential users

Apart from the publicly available datasets, the data will be recorded in real-world scenarios by ADAS test vehicles (camera, LiDAR, GNSS, ...), mainly in the European Union countries. A small part will be generated from virtual data (for validation purposes). The AI part will generate data using generative models.

Potential users of this application case are within the computer vision and machine learning scientific and industrial community, especially focusing on automotive applications.

2.3 WP6 - SME Application Case - Agriculture

This application case intends to ingest new data and not work with existing data as existing data is linked to customers and TERRAVIEW is not allowed to work with them outside the scope of service delivery to them.

Purpose of the data generation or re-use

The generated datasets and the input datasets will prove the technical capabilities of EXA4MIND EDD and linked workflows which use the capabilities of EDD and HPC compute resources demonstrating the suitability of such an approach to solve large planet-level climate/agriculture linked challenges. This will be done although at a much smaller scale in the agriculture application use case.

Type and size of data

The project will generate EO datasets and post-processed linked metadata files. Alongside with them will be historical weather datasets which will be commercially acquired, and will generate the machine learning training input datasets and trained models. The expected approximate size of input+output data is 4TB of EO+Weather datasets for a chosen agricultural area in Europe.

Origin of data and potential users

EO dataset will be acquired from USGS and SentinelHub and the historical weather dataset will be purchased from OpenWeatherMap and Copernicus climate platform.

Potential users outside the project are all agriculture farmland managers who have to decide how best to irrigate their agricultural land.

2.4 WP6 - SME Application Case - Health

Existing relevant open data will be used in this application case to ensure the accuracy of the application case's results from real-world information as well as generated data (if needed).

Purpose of the data generation or re-use

The data used for this application case will emphasize the need for EDD tables, created for efficient data storage, and ultimately combine the results with satellite images for geolocation and pattern detection purposes on maps to ease the visualization of the analytics for the end users. The global objective of this application case is to transpose EXA4MIND capabilities into a concrete application.

The proposed application case is aimed at showcasing the easiness of secured health data management. It will emphasize how the platform can gather heterogeneous information from different sources (i.e. numerous hospitals), transform them and provide an exploitable database. This comprehensive approach offers valuable insights for decision-makers, allowing for better resource allocation and improved patient care.

Type and size of data

The application case will generate/re-use data of the following formats: XLS, JSON, SQLite or ElasticSearch, and Parquet. Also, satellite images and scanned documents will be used. The expected size of data (generated and re-used data combined) is up to 100TB.

Origin of data and potential users

The origin of re-used data is open-source, such as:

- Number of patients by pathology, sex, age group and territory (department, region)[10]
- Number of patients by pathology and by age group by sex - 2015 to 2020[11]
- Causes of death[12]
- Comorbidities associated with each pathology[13]
- FINESS Extraction of the establishments file[14]
- The Annual Establishment Statistics (SAE)[15]
- Vaccination data by pathology and department / region[16]

[10] https://data.ameli.fr/explore/dataset/effectifs/information/
[11] https://assurance-maladie.ameli.fr/etudes-et-donnees/cartographie-effectif-patients-par-pathologie-age-sexe
[12] https://data.drees.solidarites-sante.gouv.fr/explore/dataset/causes-de-deces/information/
[13] https://www.data.gouv.fr/fr/datasets/comorbidites-associees-a-chaque-pathologie/
[14] https://www.data.gouv.fr/fr/datasets/finess-extraction-du-fichier-des-etablissements/
[15] https://www.data.gouv.fr/fr/datasets/la-statistique-annuelle-des-etablissements-sae/
[16] https://www.data.gouv.fr/fr/datasets/donnees-vaccination-par-pathologie-et-departement-region/

- Health 2020 indicators[17]
- European Health for All database (HFA-DB)[18]
- Surveillance Atlas of Infectious Diseases[19]

Data could be useful for end-users, outside of the project for example for doctors, management, service directors & administrations who take care of exploitation of the data in concrete cases such as in global patient management, reinforcement of certain operational services, identifying procedure deficiency, etc. In summary, data will be used to accompany them in the decisions related to a territory/site strategy.

3 FAIR Approach

The project will support the FAIR principles, also benchmark data collections from each application case will be generated to allow the reproducibility of the validation tests. Private or intermediate-result data as they can appear in particular in WP5-WP6 including contributed IP (intellectual properties) will be made FAIR (and open) as far as possible, and protected as needed by IPR and GDPR regulations.

3.1 Repository

The EXA4MIND project develops EDD which will allow the hosting of volatile project data as well as research outputs together with appropriate metadata. It will leverage the FAIR data services of EUDAT, in particular B2HANDLE, to assign PIDs and B2FIND to publish a catalogue of our public data.

For safe long-term archival and backup, data can be transferred to archives of the participating computing centres, but also be published, e.g., on Portal of European data[20] or Zenodo[21]. For this purpose, the EXA4MIND community was already created at Zenodo[22] by AUSTRALO.

The long-term goal of EXA4MIND is to build a long-term repository after the end of the project. In the future, the platform could be used also for the storage of long-term preservation.

3.2 FAIR Data

In light of the multi-disciplinary nature of the project, we fulfil the domain-agnostic DataCite[23] metadata standard with our outputs (and intermediate data, where appropriate) as a minimum. This standard is the prerequisite for obtaining Digital Object

[17] https://gateway.euro.who.int/en/datasets/health-2020-indicators/
[18] https://gateway.euro.who.int/en/datasets/european-health-for-all-database/
[19] https://data.europa.eu/data/datasets/atlas_infectious_diseases?locale=en
[20] Portal of European data: data.europa.eu
[21] Zenodo: https://zenodo.org
[22] EXA4MIND community at Zenodo: https://zenodo.org/communities/exa4mind-heurope
[23] DataCite: https://datacite.org

Identifiers (DOIs) and allows for basic georeferencing as well as references to further (e.g. domain-specific or HPDA-relevant) metadata and basic provenance tracking by relating data products. The consortium is aware of consistent annotation needs.

3.2.1 Findability

Published data will be identified by a *persistent identifier* (PID), for example DOI, in addition to B2HANDLE PIDs used within EXA4MIND project EDD (depending on the repository used). B2FIND provides an appropriate web portal for finding research data within the scope of the EOSC. Our catalogue with data uniquely identified by B2HANDLE PIDs can be published via OAI-PMH and B2FIND. The OAI-PMH interface of our EDD and further interfaces will allow (meta-)data export to further relevant sites (e.g. data.europa.eu) so as to warrant an optimum publication of our data.

Metadata will be provided to allow the discovery of the published datasets in all application cases, see Table 1. Also, keywords will be provided in the metadata to optimise the possibility for discovery and then potential re-use. Appropriate attention will be put on the metadata to allow data harvesting and indexing.

3.2.2 Accessibility

For development, testing and validation purposes, application cases will use EXA4MIND EDD for selected datasets. In EXA4MIND EDD, data ingest and HTTP-based download mechanisms for published data will be provided. Datasets too large for these possibilities may be transferred via other appropriate protocols.

Long-term data will be published in trusted repositories, see Section 3.1. For all application cases, the Zenodo repository is convenient. Where appropriate, special domain repositories will be investigated (for example Cryptobox by Ercom/Thales for SME health application case), but the cost needs to be evaluated. Detailed information about accessibility to data published by application cases is provided in Table 2.

Scientific Application Case
Data will be structured in 4 layers (EXA4MIND 2023a). Among these layers, it is worth mentioning that the second layer contains knowledge extraction data files (post-processed files = metadata), which will be generated from the MD trajectory. Those files (which could be hundreds of different files) contain information about the evolution of distance, angle, dihedral, coordination etc. in time between specific atoms from the simulated system (i.e., the user defines structural parameters to be calculated from the trajectory). Simulation datasets will be structured base on systems (for example the name of the system or any other identifier will be used, i.e., PDB ID[24] of the starting structure). Metadata will allow indexing based on atom ID or atom labels.
Industry Application Case
Metadata will be generated within the scope of this project, for example, vehicle data (speed, odometry, steering, etc.), 2D/3D objects appearing in the scene, classes of objects, segmentation, weather conditions, road type, and others. Some metadata will be extracted directly from car signals and others will be extracted by developed AI models. Also, trained models will be accompanied by metadata. Metadata will include a list of tags to allow findability and easy re-use.
SME Application Case - Agriculture
This application case will provide the input datasets for a chosen area of interest (AoI), and the ML training and validation results publicly with linked persistent identifiers. However, the model training pipeline, input transformation done for training and the generated model itself will not be public. Metadata will be provided to allow the discovery of published datasets, but the correct taxonomy will be to researched. The metadata will contain the geo-markers for the chosen AoI, the period of historical data observed for that AoI, and the origin/source of the data. Also, appropriate keywords will be used.
SME Application Case - Health
The input datasets and AI models can be disseminated publicly and linked with a persistent identifier. However, the model training pipeline, data transformation done for training and the generated model itself will not be public.

Table 1: Making Data Findable in Application Cases - Metadata

Scientific Application Case	
Accessibility	All data will be freely available without any embargo and restrictions.
Identification of Users	There will be no restrictions for members of the MD community to upload datasets, only validation that the correct formats of the datasets were uploaded. For FF developers, an approved list of experienced users will be used and validated by the data access committee.
Industry Application Case	
Accessibility	Only a subset of all data will probably become publicly available (part of VALEO datasets). Regarding the AI part, publication depends on the copyright and restrictions of the used datasets. It is not planned to use an embargo for published data.
Identification of Users	If the dataset will be published, the users will be asked to provide information like email, name, company, address and purpose of use. The purpose of use must be research and they would have to agree to licensing terms. There are no personal/sensitive data within the output datasets, so there is no need for a data access committee.
SME Application Case - Agriculture	
Accessibility	Only the earth observation data and related metadata fetched from freely accessible sources for the test land area alongside the processed soil moisture content map tiles and associated metadata (SMC) will be made available openly. The historical weather dataset from OpenWeatherMap can not be shared (the license does not allow free distribution to 3rd parties). The identified datasets can be published as soon as they are available. There is no IP on datasets.
Identification of Users	As the output dataset contains no personal identifiable markers, and the output projections are not long-range in nature thereby rendering no tangible commercial exploitation of the test dataset, it is not necessary to track who accesses the datasets. There are no personal/sensitive data within the output datasets, so there is no need for a data access committee.
SME Application Case - Health	
Accessibility	All fetched data from publicly available sources will be openly available (except if special licensing conditions apply, see Section 5.1.4 in Deliverable D7.1, EXA4MIND 2023b). Transformed data will be accessible through different formats such as visuals (dashboard, maps). Access to the data must be controlled by a verification of the requester and the acknowledgement of proper usage to avoid malicious behaviours.
Identification of Users	A request for user identity and the motivations on how the data will be used will be required. No sensitive data will be processed or outputted. A simple evaluation/approval process might be sufficient, according to previously defined common rules.

Table 2: Making Data Accessible in Application Cases

In most application cases, the output metadata will be made openly available. The license can be CC-0. Data and metadata will be available as long as possible according to the policy of the selected repository. If documentation or any software will be needed to access or read the data, it will be included by a reference to the repository (for example GitHub) or a publication.

In the Industry application case, metadata will be available only with the published dataset. If the dataset will be published, it will come out under a licence that allows the use for research purposes only. The exact license was not prepared yet.

3.2.3 Inter-operability

Wherever possible, domain-specific best-practice standards for metadata and data storage will be used. For example, there are many industry standards relevant to Industry application cases, which require additional knowledge and specific libraries. If published, data and outputs of AI models will be published in widely common formats (JSON, CSV, AVI, JPG, mp4, NumPy NPZ, PKL, etc.). In case specific ontologies or vocabularies will be used, relevant mappings to more commonly used ontologies will be provided to allow reusing, refining or extending them. If relevant, the published data will include qualified references to other data.

In SME agricultural application case, the data format of elements in the output dataset will use well-known file types such as GeoJSON and GeoTIFFs. Regarding the individual data value within these files and the vocabulary of terms, appropriate naming standards will be investigated.

3.2.4 Re-usability

Results from EXA4MIND will be put under appropriate and clearly specified licences (for example CC-BY 4.0) to permit re-use wherever desirable, with clear conditions. With its significant expertise in computational science, the consortium will make sure that the data are reusable in a practical sense (whether stored within EXA4MIND or in another repository). Where needed, re-usage instructions will be provided within the metadata or appropriate documentation.

4 Other Outputs

Source code and relevant documentation will be maintained on the platforms operated by IT4I@VSB and BADW-LRZ (e.g. https://gitlab.lrz.de, https://opencode.it4i.eu). These platforms are complemented by open providers for persistent identifiers (e.g. ORCID), code repositories (e.g. GitHub), and data repositories (e.g. Zenodo, data.europa.eu).

The software systems in EXA4MIND will be made reproducible by appropriate automation tools. Configurations and set-up scripts will be managed with Ansible (and similar tools). All this will be deposited in the project's GitLab for versioning. Relevant versions (releases) will be published together with the project's research outputs (including necessary documentation) for transparency.

Regarding SME application cases, part of the software code will be maintained by TERRAVIEW and ALTRNATIV as part of their internal code repositories in-house. Some ML models will be public as ALTRNATIV wants to share most of the knowledge publicly, however, some others will remain private and maintained by ALTRNATIV or TERRAVIEW using internal IT infrastructure within the company or in private repositories maintained on GitHub.

5 Security

Regarding the EXA4MIND Platform and infrastructure of IT4I@VSB, internal processes and regulations for safe work with information that prevent misuse of information, unauthorized changes and data loss will be followed for data management and their safe transfer within individual locations. IT4I@VSB holds the ISO 27001 certificate (ISO/IEC 27001:2013, ČSN ISO/IEC 27001:2014). The certificate was awarded for providing services to the national supercomputer infrastructure, solving computationally demanding problems, advanced data analysis and simulation, and processing big data.

LRZ has been certified according to ISO 20000 and 27001 since 2019, ensuring consistent service management and quality as well as proper information security management. The service portfolio of BADW-LRZ with specific service levels and availability is laid out in the LRZ service catalogue [25]. Service management and information security management are handled with an integrated "I/SMS" (Information Security Management System (ISMS) + Service Management System (SMS)). This system, with the procedures and policies therein defined, has been the basis for BADW-LRZ's successful certification (and re-certification in 2022, according to ISO/IEC 20000:2018 and ISO/IEC 27001:2013). In 2022, BADW-LRZ's research unit – though usually not directly responsible for the operation of services – has been fully added to the scope of IT security management according to ISO 27001. This includes the usual aspects with everyday impact such as secure data and information handling (with encryption where necessary), consistent risk and security management based on security concepts, and proper encapsulation of research, test and production infrastructure.

Appropriate security and user-to-user isolation measures will be implemented in EXA4MIND such that high-performance analytics remain possible and users can effectively share data whenever it is allowed.

Regarding partners' application cases, they will mostly use the EXA4MIND platform. In the case of any test being made on the ALTRNATIV premises, the security rules of the project will be strictly followed. Also, ALTRNATIV follows state-of-the-art security standards to ensure the proper protection of data and privacy (internal company policies around data management, based on ISO 27001 and 9001 principles). Input data will be anonymised. In the case of a dataset containing sensitive data, it will immediately be anonymised before any storage or processing. Final data will be related to statistics and anonymised data.

Data intended for publication will be securely stored in trusted repositories for long-term data retention and maintenance, see Section 3.1.

[25] LRZ service catalogue: https://www.lrz.de/wir/regelwerk/dienstleistungskatalog.pdf

6 Ethics

The consortium will carry out the action in compliance with Article 14 Ethics in Annex 5 of the Grant Agreement.

In the EXA4MIND project, there is a special work package WP9, whose objective is to ensure compliance with the 'ethics requirements set out in the project. An external independent Ethics Advisor specialised in EU data protection law and the ethics of AI was appointed in M2 with the purpose to be consulted on issues related to data identifiability, data minimisation, and lawful bases for data processing, including for all four use cases. There is a special deliverable D9.1 OEI - Requirement No. 1 (type of sensitive, EXA4MIND 2024a) planned for M18, in which the ethical issues will be discussed in detail.

In general, the partners will work with data in accordance with ethical norms and standards. The consortium is aware of research community discussions around the ethics of data production and consumption and will manage the data with those ongoing discussions in mind (Advanced Analytics at North Carolina State University 2023). It is not assumed that there should be any ethical obstacles in the project that would have an impact on data sharing. Data will be anonymised where needed to comply with data privacy regulations (Industry and SME Health application cases).

In this phase of the project, it is assumed that there will be no need to store datasets containing personal data. Therefore, it will not be necessary to use questionnaires dealing with personal data, including requiring informed consent to data sharing and long-term storage of personal data. This will be regularly reassessed during the project.

7 Data Outputs Plan

We have identified outputs we can provide during the project for each application case. This is a preliminary plan which will be regularly checked and updated.

7.1 Scientific Application Case

Future force field development should be automated and assisted by ML approaches on top of Extreme Data acquired from HPC. We propose the intended reusing of existing datasets by the reweighing approach that enables efficient exploitation/mining from MD datasets. Efficient analysis and visualization (ADAMS4SIMS) will proceed via the datacube structure, which will organize multiple dimensions of the data and enable efficient and flexible access to any subset of the datacube, based on the user's needs and preferences, see Table 3.

Item	Description
Dataset ID	DSSCI01
Dataset Name and Reference	ADAMS4SIMS, reference to be specified later
Dataset Description	Database of Simulations
Standards, Format, and Metadata	Simulations, metadata (calculated parameters from simulations), experimental datasets, the database of weights.
Data Sharing (Including License)	Freely available (license will be specified later)
Is Dataset Accessible?	Not yet
Is Dataset Reusable?	Not yet
Archiving and Preservation (Storage and Backup)	Data will be available at one of the long-term repositories like Zenodo. To be specified later.

Table 3: Scientific Application Case - Output DS01

7.2 Industry Application Case

For VALEO company, there will be multiple useful outputs coming from this project. First, it will be a set of scientific publications from the joint research efforts of all partners. The publications in which VALEO will participate will cover mainly the following AI topics: self-supervised learning in multi-modal setup (camera and LiDAR data), domain adaptation, virtual validation via generative models etc.

Second, VALEO will use the AI models, trained and developed during the EXA4MIND project, for automatic ground truth extraction for reliable system validation and rich metadata discovery in captured data. This will be used for smart validation to increase the safety of advanced driving assistance systems. Last but not least, it will be the use of a database developed in cooperation with EXA4MIND partners. This database, with a new set of tools, will enable real-time interaction with extreme-size data enriched by outputs of AI models, see Table 4.

For CVUT, one of the outputs will be trained model weights stored in checkpoints for further use. Furthermore, the developed method will be described in a scientific paper, see Table 5.

Item	Description
Dataset ID	DSIND01
Dataset Name and Reference	Dataset name and reference to be specified
Dataset Description	ADAS test driving recording (Camera, LiDAR, GNSS) + generated metadata
Standards, Format, and Metadata	Formats of published data will be standard ones (jpeg, avi, pcd, npz, json, xml, etc.).
Data Sharing (Including License)	To be specified, for academic purposes only
Is Dataset Accessible?	Not yet
Is Dataset Reusable?	Not yet
Archiving and Preservation (Storage and Backup)	Dataset will be available at one of the long-term repositories like Zenodo. To be specified later.

Table 4: Industry Application Case - Output 01

Item	Description
Dataset ID	DSIND02
Dataset Name and Reference	Dataset name and reference to be specified
Dataset Description	Machine learning model
Standards, Format, and Metadata	pkl PyTorch format
Data Sharing (Including License)	To be specified, free for academic use
Is Dataset Accessible?	Not yet
Is Dataset Reusable?	Not yet
Archiving and Preservation (Storage and Backup)	Dataset will be available at one of the long-term repositories like Zenodo. To be specified later.

Table 5: Industry Application Case - Output 02

7.3 SME Application Case - Agriculture

The new datasets which will be generated as a result of this application use case experimentation with the EXA4MIND platform (see Table 6 and Table 7) will include the following:

- Historical soil moisture content map tiles including the numerical data associated with each pixel in the tile encoded within the corresponding GeoJSON file for the canonical agricultural land area under study - for relevant Landsat satellite scene available from USGS / SentinelHub endpoints since the operations commencement date of the satellite.
- ML model-assisted generated back-filled soil moisture content map tiles for the

days where original satellite scenes are not available due to unsuitable atmospheric conditions, or no overpass days. Also linked will be numerical values for every pixel encoded as appropriate GeoJSON files.

- Forward projection of soil moisture content values as map tiles for 2 weeks beyond the date of data generation and subsequent validation of the projection accuracy using the actual ground sensor data or satellite scenes once they become available in the future.

Item	Description
Dataset ID	DSSMEAGR01
Dataset Name and Reference	Test AOI outline
Dataset Description	The landmass area outline for the agricultural landmass under study in this application use case.
Standards, Format, and Metadata	GeoJSON: RFC 7946: The GeoJSON Format (rfc-editor.org).
Data Sharing (Including License)	Creative Commons
Is Dataset Accessible?	Not yet
Is Dataset Reusable?	Not yet
Archiving and Preservation (Storage and Backup)	Potentially via Zenodo.

Table 6: SME Application Case - Agriculture - Output 01

Item	Description
Dataset ID	DSSMEAGR02
Dataset Name and Reference	Test AOI SMC scene catalog
Dataset Description	The set of soil moisture content analysis map tiles and associated metadata files for the target AoI – both historical backfilled data as well as predicted moisture content map tiles and numeric values.
Standards, Format, and Metadata	GeoJSON: RFC 7946: The GeoJSON Format (rfc-editor.org) GeoTIFF map tiles at various zoom levels.
Data Sharing (Including License)	Creative Commons
Is Dataset Accessible?	Not yet
Is Dataset Reusable?	Not yet
Archiving and Preservation (Storage and Backup)	Potentially via Zenodo or in GitHub within the public repository.

Table 7: SME Application Case - Agriculture - Output 02

7.4 SME Application Case - Health

One of the outputs will be a compiled, processed and enriched database from the identified sources to obtain a global and exploitable database and generate statistics on medical data (patients, facilities, etc.), see Table 8). Another dataset will be built for predictions on diseases, facilities, and anomaly detection (see Table 9). Last, another expected output is a dataset targeted at emphasising patterns and trends on medical data (see Table 10).

Item	Description
Dataset ID	DSSMEHEA01
Dataset Name and Reference	Name and reference to be added later.
Dataset Description	Statistics on medical data (patients, facilities, etc.).
Standards, Format, and Metadata	CSV, JSON
Data Sharing (Including License)	Available (license will be specified later).
Is Dataset Accessible?	Not yet
Is Dataset Reusable?	Not yet
Archiving and Preservation (Storage and Backup)	In EXA4MIND EDD and potentially via Zenodo.

Table 8: SME Application Case - Health - Output 01

Item	Description
Dataset ID	DSSMEHEA02
Dataset Name and Reference	Name and reference to be added later.
Dataset Description	The dataset will be built for predictions on diseases, facilities, and anomaly detection.
Standards, Format, and Metadata	CSV, potentially GeoJSON
Data Sharing (Including License)	Available (license will be specified later).
Is Dataset Accessible?	Not yet
Is Dataset Reusable?	Not yet
Archiving and Preservation (Storage and Backup)	In EXA4MIND EDD and potentially via Zenodo.

Table 9: SME Application Case - Health - Output 02

Item	Description
Dataset ID	DSSMEHEA03
Dataset Name and Reference	Name and reference to be added later.
Dataset Description	Dataset targeted at emphasising patterns and trends in medical data.
Standards, Format, and Metadata	CSV, potentially GeoJSON
Data Sharing (Including License)	Available (license will be specified later).
Is Dataset Accessible?	Not yet
Is Dataset Reusable?	Not yet
Archiving and Preservation (Storage and Backup)	In EXA4MIND EDD and potentially via Zenodo.

Table 10: SME Application Case - Health - Output 03

Item	Description
Dataset ID	DSSMEHEA04
Dataset Name and Reference	Name and reference to be added later.
Dataset Description	Anonymised patient database.
Standards, Format, and Metadata	CSV, JSON
Data Sharing (Including License)	Available (license will be specified later).
Is Dataset Accessible?	Not yet
Is Dataset Reusable?	Not yet
Archiving and Preservation (Storage and Backup)	In EXA4MIND EDD.

Table 11: SME Application Case - Health - Output 04

8 Conclusion

This deliverable explained the special position of the EXA4MIND project regarding the development of the platform and the integration with the European FAIR Data Ecosystem. The document summarised the awareness of the consortium to respect the FAIR data approach in the work with datasets. The type and size of data to be used in all four application cases were described, including the preliminary specification of outputs. This document will be continuously updated during the project. The final version of DMP will be submitted at the end of the project.

9 References

Advanced Analytics at North Carolina State University, Institute for (2023). **The Data Ethics Repository**. URL: https://dataethicsrepository.iaa.ncsu.edu/topical-categories/industry-specific/ (visited on 06/27/2023).

Amazon Web Services, Inc. and affiliates (2023). **Cloud Object Storage - Amazon S3 - Amazon Web Services**. URL: https://aws.amazon.com/en/s3/ (visited on 03/23/2023).

Enspire.science (2023). **TRL Scale in Horizon Europe and ERC – explained**. URL: https://enspire.science/trl-scale-horizon-europe-erc-explained/ (visited on 05/15/2023).

EXA4MIND (2023a). **Deliverable D1.1 Report on FAIRification of Database Data**.

EXA4MIND (2023b). **Deliverable D7.1 Impact Master Plan**.

EXA4MIND (2023c). **Deliverable D8.1 Data Management Plan**.

EXA4MIND (2024a). **Deliverable D9.1 OEI - Requirement No. 1**.

EXA4MIND (2024b). **Milestone MS10: Paper on Preprocessing and Metadata for Fair AI**.

EXA4MIND (2024c). **Milestone MS6: Paper Related to the Integration of Distributed Data Sources, HPC, and FAIR Ecosystems with EDD**.

EXA4MIND (2025a). **Deliverable D2.4 Report on FAIRification of Database Data**.

EXA4MIND (2025b). **Deliverable D8.3 Data Management Plan (Updated)**.

EXA4MIND (2025c). **Milestone MS7: Distributed FAIR Data and Extreme Data Database Final Operation Guidelines are Published**.

LEXIS (2023). **LEXIS Portal documentation**. URL: https://docs.lexis.tech (visited on 05/15/2023).

Wilkinson, Mark D. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." In: **Scientific Data** 3. URL: https://www.nature.com/articles/sdata201618.

Xu, Hao et al. (2017). **iRODS primer 2: Integrated Rule-Oriented Data System**. Williston, VT: Morgan & Claypool Publishers. DOI: 10.2200/S00760ED1V01Y201702ICR057.

This is the version of the deliverable before the review by European Commission.