D1.6 Data Management Plan - update

Date: 23 December 2024

Document Identification			
Status	Final	**Due Date**	31/12/2024
Version	1.0	**Submission Date**	23/12/2024

Related WP	WP1	**Document Reference**	D1.6
Related Deliverable(s)	D1.5	**Dissemination Level (*)**	PU
Lead Participant	SZE	**Lead Author**	Zoltán Horváth, SZE
Contributors	PSNC, UNISTRA, MTG, FAU	**Reviewers**	Harald Köstler, FAU
			Kyriaki Daskaloudi, FN

Keywords:
Data management, datasets, global challenges, FAIR principles, national legislation, EU AI Act

Document Information

List of Contributors	
Name	**Partner**
Zoltán Horváth	SZE
László Környei	SZE
Marcin Lawenda	PSNC
Dennis Hoppe	USTUTT
Christophe Prud'homme	UNISTRA
Harald Köstler	FAU
Ravi Ayyala Somayajula	FAU
Michal Kulczewski	PSNC
Luis Torres	MTG
Kyriaki Daskaloudi	FN

Document History			
Version	**Date**	**Change editors**	**Changes**
0.1	27/11/2024	Zoltán Horváth (SZE)	Initial version of the document
0.15	28/11/2024	Marcin Lawenda (PSNC) Harald Köstler(FAU)	ToC, timeline and responsibilities approved
0.2	30/11/2024	Zoltán Horváth (SZE), Marcin Lawenda (PSNC)	Update of the methodology
0.3	30/11/2024	Zoltán Horváth (SZE), Dennis Hoppe (USTUTT)	Update of the methodology, completed text
0.5	12/12/2024	Christophe Prud'homme (UNISTRA), Michal Kulczewski (PSNC), László Környei (SZE), Luis Torres (MTG), Kyriaki Daskaloudi (FN)	Update the existing use case application data reports and the other data
0.55	13/12/2024	Kyriaki Daskaloudi (FN)	Create the report for the communication, dissemination, and exploitation
0.6	18/12/2024	Ravi Ayyala Somayajula (FAU)	Create the data report of MTW, the new use case application,
0.7	19/12/2024	Zoltán Horváth (SZE)	Preparation for submission and submission for internal review
0.8	21/12/2024	Zoltán Horváth (SZE)	Changes after the internal review
0.95	23/12/2024	Harald Köstler	Quality assurance check

1.0	23/12/2024	Marcin Lawenda	Final check and improvements

Quality Control		
Role	**Who (Partner short name)**	**Approval Date**
Deliverable leader	Zoltán Horváth (SZE)	21/12/2024
Quality manager	Harald Köstler (FAU)	23/12/2024
Project Coordinator	Marcin Lawenda (PSNC)	23/12/2024

Table of Contents

List of Tables

List of Acronyms

Abbreviation / acronym	Description
AI	Artificial Intelligence
Cemosis	Modelling and Simulation Centre of Strasbourg (UNISTRA)
CO2	Carbon dioxide
CoE	Centre of Excellence
CKAN	Comprehensive Knowledge Archive Network
CSTB	Scientific and Technical Centre for Building (France)
CSV	Comma-separated values (text file format)
DMP	Data Management Plan
Dx.y	Deliverable number y belonging to WP x

EB	Ethics Board
EC	European Commission
FAIR	Findable, accessible, interoperable, re-usable
FAU	Friedrich-Alexander-Universität Erlangen-Nürnberg (HiDALGO2 partner)
FN	Future Needs (HiDALGO2 partner)
GC	Global Challenges
GIS	Geographic Information System
GUI	Graphical User Interphase
HPC	High Performance Computing
JSON	JavaScript Object Notation
MTG	MeteoGrid (HiDALGO2 partner)
NOx	Nitrogen oxides (NO, NO2)
Projekty	HiDALGO2's project repository at https://projekty.drive.man.poznan.pl
PSNC	Poznan Supercomputing and Networking Center (HiDALGO2 partner)
RES	Renewable Energy Systems
SZE	Széchenyi István University (HiDALGO2 partner)
Tx.y	Task number y belonging to WP x
UAP	Urban Air Project
UBM	Urban Building Modelling
UNISTRA	Université de Strasbourg (HiDALGO2 partner)
USTUTT	University of Stuttgart (HiDALGO2 partner)
VM	Virtual Machine
WF	Wildfire pilot
WP	Work Package

Executive Summary

The current document is the update of the Data Management Plan of the HiDALGO2 project (HiDALGO2 DMP), which was issued as the deliverable D1.5 of the project.

In the current document, first the methodology is revised. We added some attributes to the dataset description tables that better describe metadata information, storage, data sharing and security. We analysed in the compliance section how the changes in the European legislation and operational background, the very recently introduced AI Act and the AI Office effect on the HiDALGO2 DMP.

Following the revised methodology, the dataset reports were updated in the document for the pilots, including the full dataset report of the MTW pilot, which belongs to FAU, the new member of the consortium.

In addition, reports on datasets for the other information stored during the project are reported, namely the data used for communication, dissemination and exploitation.

As a conclusion, the need of a thorough continuous revision of the methodology was formulated to be compliant with the foreseen policy changes due to AI-related regulations.

1. Introduction

1.1 Purpose of the document

This document provides the updated Data Management Plan (DMP) for the HiDALGO2 project as of December 2024. The DMP outlines the data management methodology, while the dataset reports and their status as of June 2023 are detailed in D1.5 [1]. The methodology implements the FAIR concepts to the HiDALGO2 project and the datasets report the datasets of the pilot applications, which are the main drivers of data production in the project.

The main goals of this document are as follows:

- Update the methodology:
  - added more features to datasets to better describe data production, security, and metadata information,
  - a compliance section was added to the DMP in alignment to the significantly new relevant directive of the EU, the appointment of the AI Act. The AI Act emphasizes the need for robust data governance and management practices, including ensuring high-quality, unbiased, and transparent datasets to minimize risks and align with compliance and accountability standards for AI systems. This Act and relevant other operational conditions are analysed in the compliance section.
- update the already reported datasets for UAP, UBM, RES, WF,
- create the dataset report for the new pilot, MTW, which is a new pilot application of HiDALGO2; the MTW datasets are reported in this document for the first time,
- create dataset reports on other data, including data for communication, dissemination and exploitation,
- formulate the next steps in the DMP management.

In this way, the DMP has been completed for all data sources relevant to the HiDALGO2 project as of current state.

1.2 Relation to other project work

Data management pervades the project. Each services, pilot applications, dissemination, and project management activities of the project generate data of which management is coordinated in the current document.

According to the decision of the Ethics Board, all management data are handled, stored and reported by following the EuroHPC JU regulations for its projects, thus the DMP does not report management data, which are generated by the activities of WP1.

Technology-driven data of the HiDALGO2 services and pilots are generated by activities coordinated in WP2, WP3, WP4 and WP5 work packages and data generated by communication, dissemination, exploitation work package are results of WP6 work package's activities – all of these data are reported in the DMP. A large set of deliverables reports the DMP-related work, which can be found in the project repository and at the official sites designated by the project.

1.3 Structure of the document

This document is organized as follows:

- Chapter 1 presents the purpose and structure of the document, and makes a position of the document in the project work,
- Chapter 2 presents the update of the methodology for the HiDALGO2 data management plan.
- Chapter 3 contains, by following the template updated in Chapter 2,
    - the updates of the dataset reports for four pilot applications,
    - the new report for the MTW pilot,
    - the reports for other data: communication, dissemination, exploitation,
- Chapter 4 contains the ethical and legal compliance analysis,
- Chapter 5 contains the conclusions and outlines the next steps.

2. Update of the Methodology of the Data Management in HiDALGO2

2.1 Extension of the scope of the HiDALGO2 DMP

In addition to reporting the data management of the HPC-related data, the scope of the DMP has been extended with reporting of datasets produced by dissemination activities.

2.2 The applied Open Science practices

HiDALGO2 is committed to Open Science and adopts the best practices for that. Their implementation respects the FAIR principles as is explained in Section 2.2 of D1.5. In addition to the existing attributes, the representation of the following topics were specified more rigorously in the HiDALGO2 DMP by following the discussions of the Ethics Board:

- data storage, infrastructure for storage and backups,
- method of data generation, addressing whether data are dummy, synthetic, from numerical computation or observations,
- metadata for each reported data when available,
- information on data preservation and long-term storage,
- data security.

2.3 Dataset management documentation templates

The HiDALGO2 DMP templates for reporting on the datasets are provided in D1.5 and were approved by the EC.

Table 1. Template of the dataset summary

<Dataset> Data Summary	
Brief overview of the dataflow	
Data purpose, metadata, types, formats, and origin (for existing data)	
Data collection: methods for data generation, instruments, ethical considerations	

Expected size of the data	
Data utility: to whom might the data be useful outside the project	

Table 2. Template of the FAIR principles description

<Dataset> FAIR principles	
Findability: Directory structure, name convention, persistent identifiers	
Findability: Metadata: standards, keywords, indexing opportunities	
Accessibility: Infrastructure for storage and backup, Repository for preservation and sharing the data, Preservation and long-term storage	
Accessibility: Data availability: open or restricted, access protocol, security	
Accessibility: Metadata: open availability, duration of access of metadata	
Interoperability: Support of data exchange (e.g., with qualified references)	
Re-usability: Documentation (e.g., readme files with information on methodology,	

codebooks, data cleaning, analyses, variable definitions, and units of measurements)	
Re-usability: Roles, responsibilities for data management within the project team	
Re-usability: Licenses	

Table 3. Template of remaining dataset aspects description

<Dataset> Other outputs, FAIR costs, security, legal and ethics issues	
Management of other research outputs (software, workflows, protocols, models)	
Allocation of resources: research data/output management costs for enabling FAIR	
Security: provisions for data security (data recovery, secure storage/archiving, and transfer of sensitive data)	
Ethics, legal, and other issues	

2.4 Data management plan review and update procedures

The DMP is updated periodically in every quarter of the project's lifetime; this document is a publication of the current status of the DMP. This work is supervised by the Ethics Board of the project on their regular meetings, which is held regularly in each quarter of the years, and when needed, the methodology is amended and the reports updated, see also Chapter 4 below on compliance considerations.

3. Dataset reports for the pilots

The data management procedures for the pilot applications are presented in consecutive sections of this chapter. The pilots and their abbreviations are as follows:

- Urban Air Project (UAP)
- Urban Building Modelling (UBM)
- Renewable Energy Sources (RES)
- Wildfire simulation (WF)
- Material Transport in Water (MTW)

The tables that belong to the MTW are new; the others are updated versions of the already reported databases.

For the sake of the HiDALGO2 DMP self-contained, the full descriptions are provided.

3.1 Reports on the UAP datasets

Table 4. Dataset summary of UAP use case

UAP Data Summary	
Brief overview of the dataflow	UAP input data describe the geometry of the domain of interest, mesh for the geometry, weather on a coarse mesh, and optionally emission data directly, or traffic network data for emission data generation. Optionally, the geometry is obtained from the OpenStreetMap database and its mesh by an automated meshing procedure; weather information is obtained from measurement data and weather forecast services; traffic data are derived from loop and camera data by traffic simulation and emission calculation. Output data consists of raw data from the simulation, extracted and processed data for analytics, and visualization data. In addition, for each UAP component parameter files contain information on the physics to be treated in that component and configuration files assist in running the component.
Data purpose, metadata, types, formats, and origin (for existing data)	Purpose of the data is to store simulation and measurement data and intermittent or final processed data. Most small-scale data, like measurements, boundary conditions, traffic data, emissions are stored in CSV table format. CFD mesh models are stored in fluent MSH format. Simulation outputs are stored in EnSight or VTK formats, which are processed into 3D objects and 2D images. Quantitative evaluations results are stored in CSV format. Weather forecast data is stored in GRIB format. Tags are used to include metadata describing data type, project, software version, city name, etc.
Data collection: methods for data generation,	All data collection and most of the processing are done by in-house tools, that are tied into the UAP workflow. Some meteorological data is obtained from the Hungarian Meteorology Service and is available online at https://odp.met.hu. For some investigations, random or systematic weather data is generated. This data is freely available. The

instruments, ethical considerations	source of the data is given in the header of the data file and in a readme file of each dataset.
Expected size of the data	The total expected size of simulation input data is about 10 to 100 MB per simulation. Output data size greatly varies from simulation parameters and can go from 1GB to 10TB.
Data utility: to whom might the data be useful outside the project	Meteorological data is useful for everyone, who is interested in weather forecasts. Report data of simulations and applications are useful for policy makers.

Table 5. FAIR principles description of UAP use case

UAP FAIR principles	
Findability: Directory structure, name convention, persistent identifiers	All used data is currently available on https://datarepo.mathso.sze.hu/ and https://ckan.hidalgo2.eu. Public datasets are available without logins. No directory structure is used currently. Data can be found with searching for tags.
Findability: Metadata: standards, keywords, indexing opportunities	Data is tagged. The following tags are used: inputs (for inputs), 2.0 (for workflow version 2.0), OPENFOAM (for data with OpenFOAM), fluidsolver (for data with FluidSolver), gyor (for data with the city of Győr, or any other city name), svd (for input data for singular value decomposition). Format and licences are also given: csv, mesh, tar, text, zip, application, png, json for format and creative commons or not-open for license.
Accessibility: Infrastructure for storage and backup, Repository for preservation and sharing the data, Preservation and long-term storage	Our current datasets are hosted at https://datarepo.mathso.sze.hu/ and https://ckan.hidalgo2.eu. These sites are hosted at SZE and PSNC, respectively. Storage service is based on CKAN running on a Linux based VM system. Data may also be stored on clusters and local workstations if being worked on or being used for simulation. Archived simulations are stored on local clusters and workstations.
Accessibility: Data availability: open or restricted, access protocol, security	As most tools are developed with funding from many projects, all data access is negotiated beforehand. All data that are public can be accessed without login from https://datarepo.mathso.sze.hu/. Any other data can be accessed only by a valid login that can be obtained by contacting the pilot providers at math@sze.hu.
Accessibility: Metadata: open availability, duration of access of metadata	Accessibility can be set as public or private at the data store by the data uploader. No additional metadata is stored regarding this point. Currently, it is not possible to set any time limit for access.

Interoperability: Support of data exchange (e.g., with qualified references)	Data to be shared can be uploaded to repository and be shared with partners with a sharable link.
Re-usability: Documentation (e.g., readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, and units of measurements)	Most CSV data is self-explanatory, with descriptive column names. All special internal data structures are described in detail, and a description is provided for shared datasets.
Re-usability: Roles, responsibilities for data management within the project team	A team member is responsible for managing the CKAN repository. Other team members get member access to CKAN database. The member who shares the data is responsible for providing proper description, visibility settings and tags.
Re-usability: Licenses	All tools and have SZE license, as most tools are developed across many projects. Each access needs to be clarified. All data and tools that do not need access clarification are available without login at https://datarepo.mathso.sze.hu/.

Table 6. Remaining dataset aspects of UAP use case

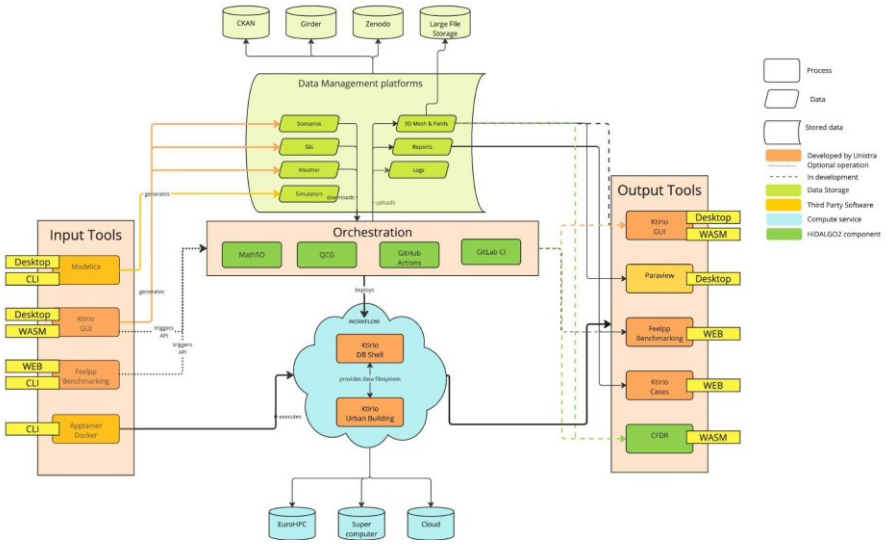
UAP other outputs, FAIR costs, security, legal and ethics issues	
Management of other research outputs (software, workflows, protocols, models)	All research output is considered by its creator and the project owner for publications, use and dissemination.
Allocation of resources: research data/output management costs for enabling FAIR	Costs are included in day-to-day work and not tracked separately.
Security: provisions for data security (data recovery, secure storage/archiving, and transfer of sensitive data)	Currently, there are no sensitive data in the workflow. Data recovery is handled by reproducibility. Data storage currently on HPC systems and CKAN. Security is handled by these systems. Archiving of results is needed more consistently.

Ethics, legal, and other issues	Ethics: We pay sufficient attention to this matter and are open to suggestions to the Ethics Board if an issue arises. Additional effort is put into steady clarifying of per request data and tool access.

3.2 Reports on the UBM datasets

Table 7. Dataset summary of UBM use case

UBM Data Summary	
Brief overview of the dataflow	The workflow begins with Ktirio GUI and Modelica, which act as the primary tools for generating simulation input data. By leveraging OpenStreetMap as a core data source, the system extracts and processes geospatial information, including vegetation, terrain, and building geometries. This data is used to create GIS and 3D representations, which are further refined through partitioning to support efficient execution on high-performance computing (HPC) systems. The workflow also includes scenario generation tailored to specific building types, weather conditions, and usage patterns, along with the computation of solar mask data to analyse the impact of sunlight on building surfaces. The prepared GIS and 3D meshes, along with other simulation inputs, are automatically processed into models for 3D building simulations and multizone analyses. All data, including intermediate files, is uploaded to data management platforms such as CKAN, Girder, or Zenodo for centralized storage and accessibility. Simulations are executed on EuroHPC systems, supercomputers, or cloud platforms, with support for coupling urban air pollution (UAP) models to enable integrated analyses. This execution generates detailed outputs, including 3D visualizations of temperature distribution, heat flux, and CO2/NOx concentrations on building surfaces. Additionally, comprehensive reports and dashboards provide insights into the energy performance, emissions, and key statistics of selected buildings within the city, as identified through Ktirio GUI. This streamlined workflow ensures seamless data handling, scalable simulations, and actionable results for urban and building energy modelling.



Data purpose, metadata, types, formats, and origin (for existing data)

Geometry: from OpenStreetMap

We start by sourcing data from OpenStreetMap (OSM) as the primary geospatial information provider. This includes extracting data for terrain, buildings, vegetation, and urban layouts. These datasets are enriched and processed for various simulation and modelling purposes. Below is a detailed breakdown:

1. **Data Purpose**:

The data is used for generating GIS and 3D representations, preparing watertight meshes, modelling urban energy scenarios, and creating simulation-ready input files for building energy and environmental modelling. It supports both multizone and 3D building simulations, solar mask computation, and coupling with urban air pollution (UAP) models for advanced analyses.

2. **Metadata (JSON)**:

Metadata is stored in **JSON** format and includes:

- **Location Information**: Latitude, longitude, and area bounds.
- **Building Attributes**: IDs, construction types, usage categories (residential, commercial, etc.), and energy-related parameters (insulation, materials).
- **Vegetation**: Types (trees, shrubs), canopy size, and placement coordinates.
- **Weather Data**: Source (e.g., OpenMeteo), granularity, and format.
- **Scenario Settings**: Schedules, thermal properties, heating/cooling setpoints, and ventilation rates.

3. **Types of Data**:

- **Geospatial Data**: Urban layouts, terrain elevation, and vegetation.
- **Building Geometry**: 2D and 3D models, surface meshes, and structural properties.
- **Weather Data**: Historical, observational, and forecasted weather parameters.

	● **Scenario Data**: Customizable energy-use schedules and thermal zone configurations. ● **Simulation Outputs**: Temperature maps, heat fluxes, CO2/NOx concentrations, and statistical reports. 4. **Formats**: ● **Input**: ○ Vector formats: GeoJSON, Shapefiles. ○ Raster formats: MVT ○ Scenario configurations: JSON. ○ Mesh files: MSH. ● **Output**: ○ 3D visualization data: ParaView-readable formats. ○ Reports: JSON, HTML (dashboards and statistics). ○ Logs and metadata: JSON. 5. **Data Origin**: ● **Existing Data**: ○ OpenStreetMap: Core geospatial information for terrain, buildings, and vegetation. ○ OpenMeteo or similar: Weather observation and forecast data. ○ Local datasets: Optional high-resolution GIS and building data when available. ● **Generated Data**: ○ Meshes and partitioned models are produced using internal tools like **Ktirio GUI** and mesh generators. ○ Scenario and weather data are customized for energy simulations.
Data collection: methods for data generation, instruments, ethical considerations	Mapbox Vector Tiles (.mvt format): tiles of city downloaded from Mapbox Gmsh mesh format: meshes of buildings and cities, produced by workflow JSON: metadata describing the buildings, produced by workflow and database scraping **Data Collection: Methods, Instruments, and Ethical Considerations** 1. **Methods for Data Generation**: ● **OpenStreetMap (OSM)**: We source urban layouts, building geometries, and vegetation details directly from OSM. This data is then enhanced with external datasets when high-resolution local data is available. ● **Mesh Generation**: To produce conforming watertight meshes, we process OSM data using the **CGAL library**, which repairs geometry, eliminates inconsistencies, and ensures mesh conformity. ● **Simulation Input Preparation**: The generated meshes are integrated with weather and scenario data to create simulation-ready files. Metadata describing building

properties and urban configurations is derived through database scraping and processed into JSON format.

2. **Instruments and Tools**:
- **CGAL Library**: Repairs geometry and generates watertight meshes, ensuring that the output is ready for computational modelling.
- **Gmsh**: Converts repaired geometries into mesh formats suitable for numerical simulations.
- **Mapbox API**: Retrieves vector tiles (.mvt) of city areas for geospatial visualization and integration into workflows.
- **Python Tools and GIS Libraries**: Handle metadata extraction and transformation into JSON formats for workflow integration.

3. **Ethical Considerations**:
- **Data Privacy**: OpenStreetMap data is publicly available and adheres to global privacy standards, ensuring no sensitive personal data is included.
- **Attribution**: All sourced data complies with OSM's license requirements and Mapbox terms of use, with proper attribution provided in documentation and outputs.
- **Transparency**: Metadata and generated data are openly documented to ensure reproducibility and ethical use of results.
- **Environmental Impact**: The workflow prioritizes efficiency in data processing and simulation to reduce unnecessary computational resource consumption.

Data Formats and Their Roles

1. **Mapbox Vector Tiles (.mvt)**:
- **Purpose**: These tiles provide detailed geospatial information about city layouts and are essential for defining the area of interest.
- **Usage**: Downloaded via Mapbox, they form the basis for extracting geometries and preparing inputs for mesh generation.

2. **Gmsh Mesh Format**:
- **Purpose**: Represents buildings and cities in a format optimized for computational simulations, including energy and environmental modelling.
- **Generation**: Produced through the workflow using CGAL-repaired geometries, ensuring compliance with simulation requirements.

3. **JSON Metadata**:
- **Purpose**: Describes building attributes, urban scenarios, and weather configurations.
- **Generation**: Created by scraping databases and integrating workflow outputs. Metadata ensures traceability and supports scenario customization.

These formats collectively enable seamless integration across the workflow, from data collection to simulation and analysis.

Data collection: methods, instruments, ethical considerations	Mesh data are automatically retrieved from online databases like OpenStreetMap, by providing the geographical coordinates and the radius of the zone to be retrieved. Building data (materials, usage), are also automatically retrieved from online databases, when the latter are available (CSTB database in France, for instance). All data come from publicly available services.
Expected size of the data	The data depends on the size of the city (number of buildings) and topology of the terrain, the complexity of the buildings and the level of fidelity of the buildings. Its size can range from few GB to tens of GB.
Data utility: to whom might the data be useful, outside the project	Retrieved data could be useful to land and city managers, to identify regions where temperatures are higher or where energy losses are more important. They can also interest companies working in building energy simulation, insurance, diffuse energy demand response, as well as individuals concerned by the environment of their home or companies concerned by the environment of their building assets.

Table 8. FAIR principles description of UBM use case

UBM FAIR principles	
Findability: Directory structure, name convention, persistent identifiers	On Girder (or other data management systems) The overall structure is as follows using the step ids and names to identify uniquely the workflow steps: <location-name>/<workflow-step-id>-<workflow-step-name>/<data>
Findability: Metadata: standards, keywords, indexing opportunities	Each <location-name> contains a JSON file describing the overall dataset that has been generated. This can then be used for searching and indexing purposes.
Accessibility: Infrastructure for storage and backup, Repository for preservation and sharing the data, Preservation and long-term storage	**Infrastructure for Storage and Backup** Our primary infrastructure for data storage and backup is built around **Girder 3.2.6**, hosted locally at the University of Strasbourg. Girder provides a secure and scalable platform for organizing and managing datasets during their active development phase. Data backups are automated and stored in secure locations to mitigate risks of data loss or corruption. **Repository for Preservation and Sharing** To ensure seamless data sharing and long-term accessibility, our strategy leverages the following repositories: 1. **Girder (https://girder.math.unistra.fr)**: ● The core repository for managing active datasets during the project lifecycle. ● Enables efficient handling of datasets with metadata annotations, access control, and integration with workflows.

	2. **CKAN (Hidalgo2 Data Management Platform)**: ● CKAN will serve as the primary data repository during the active phase of the **Hidalgo2 project**. ● Facilitates easy sharing, searching, and reuse of datasets among project collaborators and external stakeholders. ● Provides tools for versioning and collaboration, ensuring datasets remain well-organized and accessible during the project. 3. **Zenodo (https://zenodo.org)**: ● Zenodo will be the platform of choice for **long-term storage and public dissemination** of datasets. ● After project completion or at significant milestones, datasets will be periodically released on Zenodo, ensuring they are preserved and accessible indefinitely. ● Zenodo assigns Digital Object Identifiers (DOIs) for datasets, ensuring they are citable and easily discoverable by the broader research community. ● Depending on the size and scope of the city-scale datasets, data may be published as individual entries or grouped collections. **Preservation and Long-term Storage** The dual-platform approach ensures robust data preservation and sharing: ● **During the Project Lifecycle**: ○ Girder and CKAN are the main platforms for managing datasets. Girder handles secure storage and backups, while CKAN supports project-wide collaboration and dataset dissemination. ● **Post-project and Long-term Storage**: ○ Zenodo takes over as the long-term archival platform, ensuring datasets adhere to FAIR (Findable, Accessible, Interoperable, Reusable) principles. ○ Zenodo's integration with ORCID and other metadata standards ensures datasets remain visible and reusable for future research. This strategy balances the need for active data management during the project's life and reliable, long-term preservation and accessibility afterward.
Accessibility: Data availability: open or restricted, access protocol, security	**Data Availability: Access Protocol and Security** The data managed within our infrastructure is classified as **open or restricted**, depending on its sensitivity, relevance, and intended audience. The access protocol and security measures ensure the data is securely managed and appropriately shared. 1. **Open Access Data**: ● Public datasets are made openly accessible during the project's active phases via the **CKAN platform** and periodically released on **Zenodo** for long-term availability.

	• These datasets are enriched with detailed metadata to ensure transparency and adherence to FAIR (Findable, Accessible, Interoperable, and Reusable) principles. 2. **Restricted Data**: • Restricted datasets, such as sensitive or proprietary information, are stored and managed on the **Girder platform**. Access to these datasets is controlled through **User and Group permissions**, ensuring that only authorized individuals or teams can interact with the data. • Access rights can be configured dynamically to align with evolving project needs and data-sharing policies. **Access Protocol** The **Girder platform** simplifies access management through: • **User Authentication**: Secure login for registered users. • **Group Management**: Fine-grained access control by assigning permissions at the group level, enabling collaborative workflows. • **API Integration**: Allows programmatic access to datasets for automated workflows and external integrations, ensuring scalability and efficiency. **Security** The platform benefits from the robust security infrastructure provided by the **University of Strasbourg and its affiliated laboratory**. These security measures include: • Network-level protection via institutional firewalls and secure access points. • Regular system updates and monitoring to mitigate potential vulnerabilities. • Encrypted data storage and secure transfer protocols (e.g., HTTPS) to ensure the confidentiality and integrity of datasets. This combination of secure platforms and institutional support ensures that both open and restricted datasets are managed responsibly, fostering collaboration while safeguarding sensitive information.
Accessibility: Metadata: open availability, duration of access of metadata	The data that can be open (generated from public data) during and after the project in particular metadata. Intermediate steps may not be retained to alleviate storage cost after the project.
Interoperability: Support of data exchange (e.g., with qualified references)	Cross-links and other information between datasets will be specified in case a dataset is complemented, built on or depends on other datasets. The persistent identifier will also be provided to connect them.
Re-usability: Documentation (e.g., readme files with information on	Data collection methodology and other important information will be available via a website which is associated with the use case and created with the site generator Antora. PDF exports of the website

methodology, codebooks, data cleaning, analyses, variable definitions, and units of measurements)	pages will be enabled, as well as export in Jupyter notebook format, when applicable.
Re-usability: Roles, responsibilities for data management within the project team	Data management responsibilities will be shared among the team members.
Re-usability: Licenses	The data is licensed through the framework of the HiDALGO2 project unless the initial data origin is not public such as a high-fidelity representation of a building or a building set. The license will then depend on the owner of the initial data.

Table 9. Remaining dataset aspects of UBM use case

UBM other outputs, FAIR costs, security, legal and ethics issues	
Management of other research outputs (software, workflows, protocols, models)	The research and development software are managed at github.com, some of the models are stored on Cemosis data management platform.
Allocation of resources: research data/output management costs for enabling FAIR	Data is currently stored at girder.math.unistra.fr The cost is taken care of by the hosting lab of Cemosis regarding data management which includes storage and maintenance of the Girder
Security: provisions for data security (data recovery, secure storage/archiving, and transfer of sensitive data)	**Ethics, Legal, and Other Issues** No significant ethical concerns are anticipated based on the current workflow and data management practices. The generated data is derived from openly available or research-permissible sources, and appropriate precautions are taken to ensure compliance with relevant legal and ethical guidelines. 1. **Ethical Issues**: ● **Absence of Sensitive Data**: The data generated does not involve personal or sensitive information, mitigating concerns about privacy violations or misuse. ● **OpenStreetMap Compliance**: OpenStreetMap data is used under its open data license (ODbL), ensuring ethical use of crowdsourced information. 2. **Legal Issues**: ● **Permissible Use**: All data sources used (e.g., OpenStreetMap, Mapbox) are legally permissible for research purposes. Licensing terms are respected,

	especially for tools like CGAL and data formats such as Gmsh and JSON. ● **Restricted Data**: When initial data is not public (e.g., proprietary GIS data or sensitive information), access is appropriately restricted, and sharing is limited to authorized users in compliance with legal agreements or institutional policies. 3. **Other Considerations**: ● **Attribution Requirements**: Certain datasets or software tools require attribution under their respective licenses. These attributions are documented and included in the project's dissemination materials. ● **Data Sharing Protocols**: Open data made available via Zenodo and CKAN ensures compliance with data-sharing policies, but proprietary data will not be included in public releases. Urban Building workflow adheres to ethical and legal standards, with appropriate measures to handle restricted data and respect license agreements.
Ethics, legal, and other issues	There are no ethical issues associated to the generated data. Concerning legal issues, used data is available for research purposes unless the initial data is not public.

3.3 Reports on the RES datasets

Table 10. Dataset summary of RES use case

RES Data Summary	
Brief overview of the dataflow	RES is a framework for multiscale weather modelling which uses and produces different data during the entire workflow. In the first step, it requires global weather data to produce mesoscale or regional weather prediction. This produced data is then used by the small-scale model to enhance the prediction. The outcome is used afterwards, in combination with on-site data sensors, to estimate renewable energy production.
Data purpose, metadata, types, formats, and origin (for existing data)	The data used as an input to the workflow is following: - shape and location of urban buildings. Format: Esri Shapefile. Source: OpenStreetMap, geoportal.gov.pl, local institutions. - land cover data. Format: GeoTIFF. Source: Copernicus. - weather prediction at global level. Format: grib2. Source: NOAA's FTP server. - digital elevation model. Format: GeoTIFF. Origin: Copernicus. - static geographical data for WRF model. Origin: UCAR. Data produced during the workflow contains detailed information about weather predictions and is stored in NetCDF format.
Data collection: methods for data generation,	- manual collection for fixed data (e.g., land cover) - automatic collection via scripts for in-time changing data (e.g., global weather prediction)

instruments, ethical considerations	
Expected size of the data	- one-time download: up to hundreds of GBs for each region. - several GBs for weather prediction initial condition each time RES workflow is executed - size of generated data depends on the area of interest, grid resolution and the length of simulated time. Currently used workflows produce from several to hundreds of GBs.
Data utility: to whom might the data be useful, outside the project	The intermediate data may be useful for citizens and institutions which require detailed weather prediction. The final product output data is suitable for individuals owning and Distribution System Operators (DSO) owning/operating wind turbines and/or solar panels to predict energy production. For the DSOs the data may be used to stabilise the grid or to predict damages to the infrastructure due to weather conditions.

Table 11. FAIR principles description of RES use case

RES FAIR principles	
Findability: Directory structure, name convention, persistent identifiers	To differentiate between different workflows runs following directory structure is used at the moment: - <scenario_name>_<date_and_time>/WRF - scenario_name>_<date_and_time>/EULAG_domain_XY where XY stands form domain number It will be extended though to support information on different model setup, ensembles, etc.
Findability: Metadata: standards, keywords, indexing opportunities	Produced data is described with a metadata containing information on run – data, time, mesh size, grid resolution, forecasted period. It is planned to tag it with more keywords and metadata for search and sharing purposes.
Accessibility: Infrastructure for storage and backup, Repository for preservation and sharing the data, Preservation and long-term storage	Produced data is stored on PSNC Altair and Proxima systems, where computations take place. It is also stored in HiDALGO2 CKAN with private access.
Accessibility: Data availability: open or restricted, access protocol, security	Produced data is available only to project partners by the means of direct (but secured) access to PSNC Altair and Proxima system, and HiDALGO2 CKAN repository.

Accessibility: Metadata: open availability, duration of access of metadata	No metadata is stored now.
Interoperability: Support of data exchange (e.g., with qualified references)	The intermediate created data is available in one of the common and open formats – NetCDF – but it can be shared in other formats as well, such as HDF5, which may be used by other users and/or applications/services.
Re-usability: Documentation (e.g., readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, and units of measurements)	Data stored in NetCDF files is grouped into self-explanatory fields. Detailed documentation and data description will be provided before data is publicly shared.
Re-usability: Roles, responsibilities for data management within the project team	Data management responsibilities are shared among the team members.
Re-usability: Licenses	The license depends on the owner of the data. For the generated data, it is licenced through the framework of the HiDALGO2 project. The existing and produced software related to RES will be available under AGPLv3 license.

Table 12. Remaining dataset aspects of RES use case

RES other outputs, FAIR costs, security, legal and ethics issues	
Management of other research outputs (software, workflows, protocols, models)	Software and workflows are currently available at internal repository to be shared with project partners.
Allocation of resources: research data/output management costs for enabling FAIR	Costs are included in daily work and not tracked separately.
Security: provisions for data security (data recovery, secure storage/archiving, and	For the external available data, we rely on security provided by the data owners.

transfer of sensitive data)	Internally generated data is stored now at PSNC Altair and Proxima machines, which provides secure data access and backup/recovery capabilities. No sensitive data is considered at the moment.
Ethics, legal, and other issues	The are no ethical or other issues. As for the legal issues, used data is available for research purposes, unless initial data is not publicly available. In such case, data can be used internally by PSNC for research purposes, but can't be shared unless proper agreement.

3.4 Reports on the WF datasets

Table 13. Dataset summary of WF use case

WF Data Summary	
Brief overview of the dataflow	The wildfire pilot uses as initial input data the meteorological predictions of a global model, in our case outputs from the European ERA5 reanalysis initially and in a second stage outputs from the ECMWF forecast model. In addition to these data, static data of digital terrain model, land use and forest fuels are needed. These data are used to produce a high-resolution forecast using WRF coupled with LES-SFIRE-CHEM, that allows to evaluate the progress of the fire, its interaction with the atmosphere, the emission of smoke and polluting particles and finally the dispersion of these particles in the surrounding areas. The output NetCDF files from the coupled model can in some cases be used as inputs for the wildland fire scenario in wildland-urban interface areas, in which case they are adapted as inputs to OpenFoam. The air and smoke movement files generated by OpenFoam are used to feed the Openfire model that generates outputs of fire movement in urbanized areas.
Data purpose, metadata, types, formats, and origin (for existing data)	Initial conditions Land cover data. Format: GeoTIFF. Origin: Copernicus. One-time download. Weather forecasts at a global level. Format: NetCDF. Origin: Copernicus or ECMWF. Digital elevation model. Format: GeoTIFF. Origin: Copernicus. One-time download. -Static geographical data for WRF model. Origin: UCAR. One-time download Fuel models: Format: GeoTIFF, Origin: Spanish Environmental Ministry, Regional Forest Fire Services. Outputs of WRF-SFIRE model are in NetCDF format. Some of the fields of interest are:

	CAN_TOP Height of tree canopy m CANHFX Heat flux from crown fire W/m^2 CANQFX Moisture flux from crown fire W/m^2 CANWAT Canopy water kg m-2 CUF U-wind at canopy top m/s CVF V-wind at canopy top m/s F_INT Fire reaction intensity for risk rating, without fire J/m^2/s F_LINEINT Byram fireline intensity for risk rating, without fire J/m/s F_ROS Max spread rate in any direction m/s F_ROS0 Base rate of spread in all directions F_ROSX X component of the spread vector driven by wind and slope m/s F_ROSY Y component of the spread vector driven by wind and slope m/s FCANHFX Heat flux from crown fire W/m^2 FCANQFX Moisture flux from crown fire W/m^2 FFWIDTH Fire front width FGRNHFX Heat flux from ground fire W/m^2 FIRE_AREA Fraction of cell area on fire FIRE_HFX Observed fire heat flux W/m^2 FLINEINT Fireline intensity W/m FLINEINT2 Alternative fireline intensity J/m/s^2 FMC_G Ground fuel moisture contents FMC_TEND Fuel moisture contents by class time lag (diagnostics only) FUEL_FRAC_BURNT Fraction of fuel burnt in timestep (per 1) FXLAT latitude of midpoints of fire cells, degrees FXLONG longitude of midpoints of fire cells, degrees HFX Upward heat flux at the surface w m-2 HGT Terrain Height m LAI Leaf Area Index m-2/m-2 LU_INDEX Land Use Category NDVI Normalized Difference Vegetation Index NFUEL_CAT Fuel data, fuel categories (BEHAVE, S&B) PM10 Pm10 dry mass ug m^-3 PM2_5_DRY Pm2.5 aerosol dry mass ug m^-3 RH_FIRE Relative humidity at the surface (per 1) ROS Rate of Spread m/s W Z-wind component m/s Outputs of Openfire model are ASCII or VTK files describing the fire behaviour.

Data collection: methods for data generation, instruments, ethical considerations	The external data are collected from: COPERNICUS, ERA5, Land cover, DEM UCAR: Static geographic data ECMWF: Global weather forecasts Output data are generated by WRF-Sfire and OpenFOAM in NETCDF and VTK format. There are no ethical considerations related with the generated data.
Expected size of the data	Static Input data: 5 GB Daily forecast data: 3 GB Output daily data: 20 GB 2 to 4 TB for the total pilot runs
Data utility: to whom might the data be useful, outside the project	Output data from WRF-SFIRE-CHEM may be used for forensic use in past events, Local and regional administrations can use output data for forest management and wildfire preparedness and prevention, and operational groups during complex fires in areas of rugged terrain where pyrocumulus clouds may develop affecting the evolution of the fire. OpenFoam - Openfire data can be used by local administrations and firefighting task forces for awareness raising in potentially affected areas, for personnel training and for simulation and rehearsal exercises.

Table 14. FAIR principles description of WF use case

WF FAIR principles	
Findability: Directory structure, name convention, persistent identifiers	Still under consideration
Findability: Metadata: standards, keywords, indexing opportunities	Still under consideration
Accessibility: Infrastructure for storage and backup, Repository for preservation and sharing the data, Preservation and long-term storage	No repository is available now. Produced data is stored on PSNC Altair system or in the HPC-JU servers where computations take place. Data will be stored in the project infrastructure: CKAN Hadoop. In order to make data publicly available, there will be periodical releases on the Zenodo platform.
Accessibility: Data availability: open or restricted,	Still under consideration a project solution. Produced data is available only to project partners by the means of direct (but secured) access to PSNC Altair system at this stage.

access protocol, security	
Accessibility: Metadata: open availability, duration of access of metadata	Still under consideration. At this stage CKAN is foreseen for accessibility to data and metadata but its implementation is still on progress.
Interoperability: Support of data exchange (e.g., with qualified references)	The intermediate created data is available in one of the common and open formats, NetCDF, which may be used by other users.
Re-usability: Documentation (e.g., readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, and units of measurements)	The documentation is the subject of work. At this stage a GitLab repository is used internally, and operational procedure documentation is stored in the HiDALGO2 platform.
Re-usability: Roles, responsibilities for data management within the project team	We are responsible for data generated by WRF-SFIRE and OpenFoam-WFDS coupled models.
Re-usability: Licenses	All modules are available through open-source licences and expected results will be of public access.

Table 15. Remaining dataset aspects of WF use case

WF other outputs, FAIR costs, security, legal and ethics issues	
Management of other research outputs (software, workflows, protocols, models)	Software, workflows, and protocols, will be put available for public access through the foreseen project tools Open Project and GitHub soon.
Allocation of resources: research data/output management costs for enabling FAIR	Now there are no costs for enabling FAIR

Security: provisions for data security (data recovery, secure storage/archiving, and transfer of sensitive data)	For the external available data, we rely on security provided by the data owners. Internally generated data is stored now at PSNC Altair machine, which provides secure data access and backup/recovery capabilities. No sensitive data is considered now.
Ethics, legal, and other issues	The are no ethics or other issues. As for the legal issues, used data is available for research purposes.

3.5 Reports on the MTW datasets

In this section we report the current status of the pilot application MTW. This pilot is new in the HiDALGO2 project and is under development, thus some parts of the data tables are not available yet.

Table 16. Dataset summary of MTW use case

MTW Data Summary	
Brief overview of the dataflow	Simulation input parameters are stored in json format. These are also available as python config files.
Data purpose, metadata, types, formats, and origin (for existing data)	Simulation outputs are stored in VTK formats, which are processed into 3D objects. Metadata of the simulation is stored in SQLite and json format.
Data collection: methods for data generation, instruments, ethical considerations	besides available experimental data, a large amount of data will be obtained through numerical simulations which are used for surrogate model training.
Expected size of the data	The total expected size of simulation input data ranging from small parameter files and geometry descriptions to several GB for the surrogate model training. Output data size greatly varies from simulation parameters and can go from 10GB to several TB.
Data utility: to whom might the data be useful outside the project	For environmental activists, industrial organizations who are assessing and working towards mitigating problems in the rivers especially with changing temperatures and spread of pollutants .

Table 17. FAIR principles description of MTW use case

MTW Datasets FAIR principles

Findability: Directory structure, name convention, persistent identifiers	HiDALGO2 CKAN
Findability: Metadata: standards, keywords, indexing opportunities	Storage location within FAU local repository: waLBerla/<application>/<cluster>/<execution-script>-<unix-timestamp>/metadata.<suffix>
Accessibility: Infrastructure for storage and backup, Repository for preservation and sharing the data, Preservation and long-term storage	Local git repositories, CKAN is also a potential option. Other options should be discussed internally
Accessibility: Data availability: open or restricted, access protocol, security	Open access
Accessibility: Metadata: open availability, duration of access of metadata	https://i10git.cs.fau.de/lss-rdm/benchmarking-data Access requires manual grant of permission by repository maintainer. For HiDALGO2, might probably be changed to a HiDALGO specific storage place, that is handled by ReFrame.
Interoperability: Support of data exchange (e.g., with qualified references)	Support could be arranged with appropriate contact and meetings.
Re-usability: Documentation (e.g., readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, and units of measurements)	A suitable README file can be provided with detailed explanation of how the data has been generated and how to use this data. Apart from this if there are any Python scripts, the description of the code can be done in suitable comments in the script.
Re-usability: Roles, responsibilities for data management	FAU is responsible for data generated especially for training purposes, surrogate modelling, ensemble scenarios and visualization.

within the project team	
Re-usability: Licenses	N/A

Table 18. Remaining dataset aspects description of MTW use case

MTW Datasets Other outputs, FAIR costs, security, legal and ethics issues	
Management of other research outputs (software, workflows, protocols, models)	Software and workflows are currently available at internal repository to be shared with project partners.
Allocation of resources: research data/output management costs for enabling FAIR	Costs are included in daily work and not tracked separately.
Security: provisions for data security (data recovery, secure storage/archiving, and transfer of sensitive data)	For the external available data, we rely on security provided by the data owners. No sensitive data is considered at the moment.
Ethics, legal, and other issues	The are no ethical or other issues.

3.6 Dataset reports for other data

In this section we report data for the communication, dissemination and exploitation data by following the methodology presented in Section 2, thus three new tables were created.

Table 19. Dataset summary of communication, dissemination and exploitation data

Dissemination, Communication, Exploitation Data Summary	
Brief overview of the dataflow	FN handles key aspects of outreach, community engagement, and exploitation plans. Data originates from stakeholder data flow (already described in more detail in the previous version of the deliverable (D1.5)), market analysis, and project outputs, feeding into communication strategies and sustainability roadmaps.
Data purpose, metadata, types, formats, and origin (for existing data)	The purpose is to develop market insights, refine exploitation strategies, monitor awareness, and boost stakeholder engagement. Types of data such as surveys, community engagement metrics, IPR tracking documents, market data, newsletter subscriptions, analytics, demographics and project dissemination materials. Formats include reports (PDF, DOCX), engagement logs (XLSX), graphical materials (PNG, JPG, JPEG, SVG). Data origin is mostly stakeholder workshops, HiDALGO2 solution evaluations, and online dissemination tools.
Data collection: methods for data generation, instruments, ethical considerations	Methods: Surveys, interactive workshops, data from online awareness campaigns, and project performance assessments. Instruments: Feedback forms, and communication tools like websites and newsletters. Ethical considerations: Ensure transparency in data collection and compliance with GDPR.
Expected size of the data	The expected size of the data will be moderate—likely ranging from megabytes to gigabytes depending on stakeholder engagement intensity and outreach activities.
Data utility: to whom might the data be useful outside the project	Researchers studying HPC solutions. Policy-makers leveraging project insights. Organizations developing ethical HPC applications.

Table 20. FAIR principles of communication, dissemination and exploitation data

Dissemination, Communication, Exploitation FAIR principles	
Findability: Directory structure, name convention, persistent identifiers	All used data is currently available on the project's shared space on "Projekty" and follows the agreed by the project structure and name convention.

Findability: Metadata: standards, keywords, indexing opportunities	N/A
Accessibility: Infrastructure for storage and backup, Repository for preservation and sharing the data, Preservation and long-term storage	N/A
Accessibility: Data availability: open or restricted, access protocol, security	N/A
Accessibility: Metadata: open availability, duration of access of metadata	N/A
Interoperability: Support of data exchange (e.g., with qualified references)	N/A
Re-usability: Documentation (e.g., readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, and units of measurements)	N/A
Re-usability: Roles, responsibilities for data management within the project team	FN ensures adherence to FAIR principles within dissemination and exploitation tasks.
Re-usability: Licenses	License for Dissemination materials: non-distribution without reference

	Licence for IPRs: Creative Commons Attribution-NoDerivatives 4.0 International

Table 21. Other outputs of communication, dissemination and exploitation data

Dissemination, Communication, Exploitation Other outputs, FAIR costs, security, legal and ethics issues	
Management of other research outputs (software, workflows, protocols, models)	Communication strategies. Market analysis reports. IPR tracking tools
Allocation of resources: research data/output management costs for enabling FAIR	Costs are included in day-to-day work and not tracked separately.
Security: provisions for data security (data recovery, secure storage/archiving, and transfer of sensitive data)	Currently, there is no sensitive data in the workflow. Data recovery is handled by reproducibility. PSNC stores and manages data in the “Projekty” repository at https://projekty.drive.man.poznan.pl/ .
Ethics, legal, and other issues	There are no ethical issues associated with the generated data. Concerning legal issues, used data is available for research purposes unless the initial data is not public.

4. Ethical and legal compliance

4.1 National and EU data regulations

The way each partner manages the data in the project complies first with the national law/regulation. Some countries of the project consortium and the stakeholders have adopted special rules for data management, which have to be taken into account when managing the data flow and writing the DMP reports.

Although the project has not yet identified any need to collect sensitive data (e.g. personal, health-related), the project members are aware of the general principles for collecting and storing them in case such a need arises. At a general level, these principles can be formulated in the following points:

- Lawfulness, Fairness, and Transparency - the processing of personal data must adhere to legal standards, be conducted fairly, and be transparent in relation to the individual concerned.
- Accountability - data controllers bear the responsibility for ensuring compliance with law principles and must be able to demonstrate such compliance.
- Integrity and Confidentiality - the processing of personal data must be conducted in a way that guarantees adequate security, safeguarding against unauthorized or unlawful processing, as well as against accidental loss, destruction, or damage.
- Purpose Limitation - data collection should be limited to specific, clear, and legitimate objectives, and any further processing must not conflict with these objectives.
- Accuracy - appropriate measures must be implemented to ensure the accuracy of personal data and, when necessary, to keep it current.
- Storage Limitation - data must be retained in a manner that allows for the identification of individuals only for as long as is necessary to fulfil the purposes for which the data is processed.

Below are detailed an exemplary legal acts (mostly related to the GDPR legislation but not only) issued in partners' countries (involved in pilots' development) along with EU acts.

Table 22. National and EU data acts

Country	Act name	Link
Hungary	Hungarian Info Act, May 25, 2018	https://njt.hu/jogszabaly/2011-112-00-00

France	French Data Protection Act Law No. 2018-493 of June 20, 2018	https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037085952
Poland	Data protection Act Guidelines for applicants to complete the DATA MANAGEMENT PLAN in the research project provided by National Science Center Poland	https://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU20180001000/U/D20181000Lj.pdf https://www.ncn.gov.pl/sites/default/files/pliki/regulaminy/wytyczne_zarzadzanie_danymi.pdf
Spain	Organic Law 3/2018 of December 5 on the Protection of Personal Data and Guarantee of Digital Rights (Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales)	https://www.boe.es/buscar/pdf/2018/BOE-A-2018-16673-consolidado.pdf https://en.wikipedia.org/wiki/Organic_Law_on_Protection_of_Personal_Data_and_Guarantee_of_Digital_Rights?utm_source=chatgpt.com
Germany	Federal Data Protection Act (Bundesdatenschutzgesetz, BDSG)	https://www.gesetze-im-internet.de/bdsg_2018/BDSG.pdf
EU	The General Data Protection Regulation (EU) (2016/679) (GDPR) EU Data Governance Act (DGA)	https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

4.2 Investigation of the AI Act and the AI Office operations

In June 2024, the EU AI Act [2] was officially adopted, marking a significant development highly relevant to the HiDALGO2 project. The EU AI Act imposes obligations on AI systems based on their risk level and usage context. For research purposes, specific exemptions and considerations apply, but they are not blanket exclusions. For the innovation actions of HiDALGO2, the requirements seem more severe. These requirements will come into action in 2025. Under the actual implementation, including national implementations of the AI Act, HiDALGO2 project

will investigate the development of the legislation and will amend the data management procedures accordingly if necessary.

In the framework of the HiDALGO2 project, a critical element of the EU AI Act appears to be its focus on the quality and management of data concerning their reuse. Article 10 is particularly significant, as it delineates the requirements for data utilized in high-risk AI systems. Essential components of Article 10 encompass data management responsibilities for providers of high-risk AI systems, who are required to adopt suitable data management and governance practices. This entails a thorough examination of data collection methods, their sources, preparation processes, and strategies to identify and mitigate bias. Regarding data quality standards, it is crucial to ensure that data used for training, validation, and testing is relevant, representative, free from errors, and comprehensive in relation to its intended application. Furthermore, it should accurately reflect the specific characteristics of the environment in which the AI system will be deployed. In certain instances, the processing of special categories of personal data is permissible for the purpose of identifying and rectifying bias, provided that adequate safeguards are in place. Additionally, the directive on copyright in the digital single market offers exceptions for text and data mining (TDM), thereby promoting the reuse of data for AI training. For instance, Article 4 permits commercial TDM, as long as the rights holders have not explicitly reserved their rights.

Also, HiDALGO2 follows closely, occasionally as participant of the AI Office activities [3].

5. Conclusions

In this document the HiDALGO2 consortium has reported the current status of the Data Management Plan of the HiDALGO2 project. The document contains the slightly modified templates for the data management reports. It is complete in scope and aligns with the current status. Continuous investigation of the European AI legislation and operational status have to be done by the Ethics Board, which considers a task of updating the methodology and publish the new DMP version when the reporting regulations according to the European data management changes significantly. The change in the regulations are foreseen to 2025 according to the activities of the European AI Office and related bodies.

References

[1] HiDALGO2: D1.5 Data Management Plan, August 2023, http://dx.doi.org/10.13140/RG.2.2.31134.51521

[2] AI Act, REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 June 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689. Retrieved on 30.11.2024

[3] European AI Office, https://digital-strategy.ec.europa.eu/en/policies/ai-office. Retrieved in 06.12.2024.