D1.5 Data Management Plan

Date: 03 August 2023

Document Identification			
Status	Final	**Due Date**	30/06/2023
Version	1.0	**Submission Date**	03/08/2023

Related WP	WP1	**Document Reference**	D1.5
Related Deliverable(s)		**Dissemination Level (*)**	PU
Lead Participant	SZE	**Lead Author**	Zoltán Horváth, SZE
Contributors	László Környei, SZE, Michal Kulczewski, PSNC, Christophe Prud'homme, UNISTRA, Luis Torres, MTG	**Reviewers**	Ioannis Kitsos FN Dennis Hoppe, USTUTT

Keywords:
Data management, datasets, global challenges, FAIR principles

Document Information

List of Contributors	
Name	**Partner**
Zoltán Horváth	SZE
László Környei	SZE
Michal Kulczewski	PSNC
Christophe Prud'homme	UNISTRA
Luis Torres	MTG

Document History			
Version	**Date**	**Change editors**	**Changes**
0.1	20/05/2023	Zoltán Horváth (SZE)	Initial version of the document by using electronic tools (e.g. ARGOS)
0.3	10/07/2023	Zoltán Horváth (SZE)	Redesigned version in HiDALGO2 formats
0.8	18/07/2023	Luis Torres (MTG), Michal Kulczewski (PSNC), László Környei (SZE), Christophe Prud'homme (UNISTRA)	Pilots' inputs.
0.9	20/07/2023	Zoltán Horváth (SZE)	Conclusions, refinement of the text. Submitted for internal review.
0.95	21/07/2023	Zoltán Horváth (SZE), Luis Torres (MTG), Michal Kulczewski (PSNC), László Környei (SZE), Christophe Prud'homme (UNISTRA)	Revision based on the review comments and requests.
1.0	02/08/2023	Zoltán Horváth (SZE)	FINAL VERSION TO BE SUBMITTED

Quality Control		
Role	**Who (Partner short name)**	**Approval Date**
Deliverable leader	Zoltán Horváth (SZE)	02/08/2023
Quality manager	Jesus Gorronogoitia (ATOS)	03/08/2023
Project Coordinator	Marcin Lawenda (PSNC)	03/08/2023

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
CSTB	Scientific and Technical Centre for Building (France)
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
GC	Global Challenges
GIS	Geographic Information System
HPC	High Performance Computing
NOx	Nitrogen oxides (NO, NO2)
RES	Renewable Energy Systems
Tx.y	Task number y belonging to WP x
UAP	Urban Air Project
UBM	Urban Building Modelling
WILDFIRES	Wildfire pilot
WP	Work Package

Executive Summary

This document is the deliverable D1.5 “Data Management Plan” of the HiDALGO2 project. The document describes the methodology and the initial status of the processes for data management in the project. The data management plan, based on the HiDALGO2 Grant Agreement, implements the official recommendation by the European Commission for data management plans of Horizon Europe projects.

The main pillar of the recommendations is the implementation of the FAIR – Findable, Accessible, Interoperable, Re-usable – principles of the European Commission for data and other research outcomes.

Following the main goals of the project, which are the development and exploitation of HPC and AI-based solutions for tackling global challenges, this document reports on the datasets used by the project pilots. The pilot applications address four main global challenges, and thus they are complex software solutions. The document represents the datasets of the project as of month 6 of the project’s lifetime and will be regularly updated following the methodology and guidelines of the current document.

Consequently, this document contains information for data management procedures and specific data for pilots. This information will be used also for establishing the project data repositories, the data access management procedures, and introduces the processes required for updating this deliverable. In the upcoming versions of the current data management plan, the data management of all kinds of project data will also be discussed.

1. Introduction

1.1 Purpose of the document

All Horizon Europe projects must establish and continuously update a Data Management Plan (DMP). The first edition of the DMP must reflect the initial status of the HiDALGO2 project, within the first half year of the project lifetime. The document describes the associated data sets of the pilot applications and provides the concise background and the mechanisms relevant for the HiDALGO2 data management in general.

1.2 Relation to other project work

Data management, including data processing in scientific computing, applications of software and data handling for scientific publications and all dissemination activities, permeates the entire project. Thus, each project activity must eventually make a footprint in the DMP. Since the main technological goals of the project are to develop HPC software solutions for tackling global challenges coordinated in WP5 and develop general technology solutions supporting these HPC applications in WP4, the closest relation of the DMP requested activities are aligned with the WP4 and WP5 activities.

In M24, an update of the document with the completed scope will be compiled and submitted as deliverable D1.6, part of work package WP1.

1.3 Structure of the document

This document is organized as follows:

- Chapter 2 presents the overall methodology for the HiDALGO2 data management. This chapter includes subsections for the Here application of the FAIR principles, addressing the open science requirements, and a concrete template is provided for the data set reporting,
- Chapter 3 contains the dataset descriptions for four pilot applications following the template introduced in Chapter 2,
- Finally, Chapter 4 concludes and outlines the next steps.

2. The Methodology of the Data Management in HiDALGO2

2.1 The scope of the HiDALGO2 DMP

The main goals of HiDALGO2 project are to develop and exploit HPC and AI-based solutions primarily aimed at tackling global challenges. To demonstrate the usability of the HPC and AI-based solutions, several pilots are (re-)developed from existing pilot applications that will use the HiDALGO2 solutions and at the same time they support the development of the HiDALGO2 solutions. The pilot applications are matured, existing solutions for solving challenges of, respectively, urban air quality and planning, urban buildings energy, wildfires, and green energy production. All the pilots and the HiDALGO2 solutions, as software products, have and will have their own dataflow, which must be embedded into the HiDALGO2 portal platform that orchestrate the workflow, data management, and postprocessing. In WP4, T4.1 coordinates the design and implementation of data management procedures; the first main milestone of this task is in M12, where the data management system of HiDALGO2 will be in operation and reported in D4.1.

Currently, the pilot applications have individually a stable dataflow; the HPC and AI solutions of the CoE are under development. Thus, in the current DMP document the pilots' data management will be reported. In the next phase of the project, all the solutions' DMP will be added to the present live document, and along with the next deliverable D1.6 the whole HiDALGO2 DMP will be reported.

2.2 The applied Open Science practices

HiDALGO2 is committed to Open Science and adopts the best practices for that. The HiDALGO2 Grant Agreement [2] contains the procedures to be followed. In this section, a summary is given as follows.

Open Science is crucial for HiDALGO2 as it directly impacts five main project objectives: data collection and usage, high-level services provision, and stakeholder involvement. HiDALGO2 addresses Open Science through five key areas: Open Access, Open Data, Open Reproducible Research, Open Research Practices, and Open Science Tools.

- Open Access: HiDALGO2 promotes open access for publications and encourages researchers to select this option whenever possible. Dissemination of collaborations, outcomes, publications, and presentations is considered good practice to engage stakeholders.
- Open Data: HiDALGO2 implements FAIR principles through a data catalogue that ensures data is Findable, Accessible, Interoperable, and Reusable. A data management plan is in place to guide the collection and generation of data across various project areas.

- Open Reproducible Research: HiDALGO2 focuses on enabling others to replicate experiments under the same conditions for comparison and validation. Open-source software and shared workflows are encouraged. Researchers publish experiment notes and reproducibility guidelines for third-party access.
- Open Research Practices: Transparency in processes and tools is emphasised. HiDALGO2 publishes methods used and shares information on research conduct. Reports are made public to allow external researchers to track progress and ensure open science practices.
- Open Science Tools: Materials and codes resulting from the research are made public. Technical and scientific research reports provide access to models, algorithms, and workflows. Open repositories and open-source licenses are preferred for code publication, facilitating testing, contributions, and collaboration.
- The coordination team, which is the Ethics Board (EB) of the project, will define detailed implementation plans for these practices, monitor adherence, and provide internal training on Open Science principles. These practices will be aligned with the data management plan to ensure seamless integration.

2.3 General considerations for the management of the research data and other outputs

The HiDALGO2 Grant Agreement contains several prescriptions for managing the data and other outputs of the project activities. In this section a summary of the relevant section of the Grant Agreement, namely Section 1.2.7 on pp. 53-54, is presented, which makes the current document self-contained and serves as a guide for the regular update of the DMP.

- Data Types/Research Outputs:
  - HiDALGO2 expects diverse datasets from climate and IoT domains, including weather conditions, city geometry (building models), and sensor data. These datasets comprise numerical and text data from simulations, calculations, physical elements, logs, and observational instruments. Various formats (STL, CSV, TXT, MDT, VTK) will be supported, with sizes ranging from KBs to GBs.
- Findability of Data/Research Outputs:
  - Storage infrastructure and public repositories (e.g., Zenodo) will be utilized, with datasets published in a data catalogue (using CKAN) with standard metadata formats like DCAT. DOI assignment and keyword-based search capabilities will enhance discoverability.
- Accessibility of Data/Research Outputs:
  - Publicly available data from repositories will be promptly offered. HiDALGO2 will make partner-provided and project-generated data public, indicating ownership and licenses, unless confidentiality requirements (e.g., GDPR)

apply. Access will be granted with associated publications or earlier if feasible.

- Interoperability of Data/Research Outputs:
  - Datasets will be published in standard formats relevant to weather, hydrology, solids, and IT domains (STL, CSV, MDT, VTK). The data catalogue will adhere to metadata standards (e.g., CKAN's DCAT format) and provide a REST API for external querying.
- Reusability of Data/Research Outputs:
  - Datasets from external sources will be provided in their original form. Project datasets will promote licenses like CC BY-SA, allowing reuse, adaptation, and commercial use with attribution while maintaining the original license. HiDALGO2 will ensure adherence to adequate standards and offer access to visualisation tools.
- Curation, Storage, and Preservation Costs:
  - The project coordination team, including the technical coordinator and quality manager, will handle data management and quality assurance. Free repositories like Zenodo, linked to OpenAIR, will be utilised for long-term storage. Other data portals and research infrastructures (e.g., EU Open Data Portal, BBMRI-ERIC, CESSDA) may be considered based on dataset content.

2.4 Dataset management documentation templates

In this section the HiDALGO2 templates for the management of datasets are provided. As noted above, it closely follows the EC recommendations for Horizon Europe DMPs stated in the Modal Grant Agreement [1]. Also, the best practices of the web-based tools for supporting the writing of the DMPs are exploited; here the guidelines of ARGOS [3] are taken into account. A dedicated plan for data storage after the project ends will be made and reported in the DMP in the next published version of the DMP.

Table 1. Template of the dataset summary

<Dataset> Data Summary	
Brief overview of the dataflow	
Data purpose, types, formats, and origin (for existing data)	

Data collection: methods, 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: Repository for preservation and sharing the data	
Accessibility: Data availability: open or restricted, access protocol	
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	

2.5 Data management plan review and update procedures

The DMP is updated periodically in every quarter of the project's lifetime. This work is supervised by the Ethics Board, which was set up in M5 following the HiDALGO2 Grant Agreement. At the end of the project's second year, the DMP will be published in the D1.6 deliverable.

3. The datasets of the pilots' dataset

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

- Urban Air Project (UAP)
- Urban Building Modelling (UBM)
- Renewable Energy Eources (RES)
- Wildfire simulation (WILDFIRES)

3.1 The data management of the UAP datasets

Table 4. Dataset summary of UAP use case

UAP Data Summary	
Brief overview of the dataflow	Traffic network and CFD geometry is obtained from OpenStreetMap database. Traffic loop and camera data is used for traffic simulation and emission calculation. Weather information is obtained from measurement data and weather forecast services. Simulation output is processed for visualization and quantitative evaluation.
Data purpose, 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.
Data collection: methods, instruments, ethical considerations	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.
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 forecast.

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/. 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: Repository for preservation and sharing the data	Our current datasets are hosted at https://datarepo.mathso.sze.hu/. Data may also be stored on clusters and local workstations if being worked on or being used for simulation.
Accessibility: Data availability: open or restricted, access protocol	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 by contacting the pilot providers at horvathz@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. Description is provided for datasets that are shared.
Re-usability: Roles, responsibilities for	A team member is responsible for managing the CKAN repository. Other team members get member access to CKAN database. The

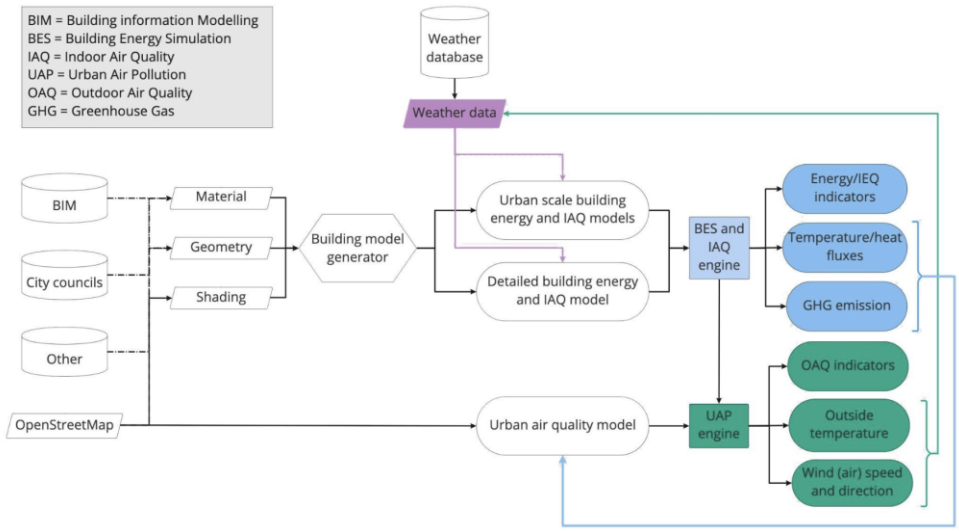
data management within the project team	member who shares the data is responsible for providing proper description, visibility settings and tags.
Re-usability: Licenses	All tools and have SZE license, as most tools are developed across many projects. Each access needs to be clarified. All data and tools 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 The data management of the UBM datasets

Table 7. Dataset summary of UBM use case

UBM Data Summary	
Brief overview of the dataflow	Here is an overview of the dataflow, starting from Location and area around the location, we produce (i) 3D and GIS representation that can be partitioned then fed to automatic model generators for 3D and multizone simulation as well as (ii) solar mask data associated to buildings, then these models and data are executed, possibly coupled with UAP to produce at the end: (i) 3D mapping of temperature, heat flux, CO2 and NOx concentration at the surface of the buildings and (ii) a report on the building energy simulations, overall dashboard and statistics of the city buildings selected by the location tool 
Data purpose, types, formats, and origin (for existing data)	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, 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 comes 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: Repository for preservation and sharing the data	For internal storage, the Girder platform (https://girder.math.unistra.fr) will be used. In order to make data publicly available, there will be periodical releases on the Zenodo platform. Depending on the size of the city under consideration, it will be either possible to publish a single dataset or multiple related datasets. https://girder.readthedocs.io/
Accessibility: Data availability: open or restricted, access protocol	Currently Girder repository for UBM is currently private within the project but it can easily made public after the first release of UBM
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 methodology, codebooks, data cleaning, analyses, variable definitions, and units of measurements)	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 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 Girder
Security: provisions for data security (data recovery, secure storage/archiving, and transfer of sensitive data)	Storage and archiving of intermediate steps of the UBM pipeline final report is the most important piece which needs to be archived. Intermediate results can be dropped or archived depending on the user's will. A retention policy may need to be setup on the intermediary data.
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 The data management of the RES datasets

Table 10. Dataset summary of RES use case

RES Data Summary	
Brief overview of the dataflow	RES is a multiscale model that 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 then used, in combination with on-site data sensors, to estimate renewable energy production.
Data purpose, 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 are stored in NetCDF format.
Data collection: methods, instruments, ethical considerations	- manual collection for fixed data (e.g., land cover) - automatic collection via scripts for in-time changing data (e.g., global weather prediction)
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 and grid resolution. 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 that require detailed weather prediction. The final product output data is 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 predict damages to the infrastructure.

Table 11. FAIR principles description of UBM 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: <area-of-interest>/<day>/<time>. It will be extended though to support information on different model setup, ensembles, etc.

Findability: Metadata: standards, keywords, indexing opportunities	Produced data is not described with a metadata at the moment. It is planned to tag it with keywords for search and sharing purposes.
Accessibility: Repository for preservation and sharing the data	The data is not stored at any repository at the moment. Produced data is stored on PSNC Altair system, where computations take place.
Accessibility: Data availability: open or restricted, access protocol	Produced data is available only to project partners by the means of direct (but secured) access to PSNC Altair system.
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 UBM 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 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 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 The data management of the WILDFIRES datasets

Table 13. Dataset summary of WILDFIRES use case

WILDFIRES 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 WFDS model that generates outputs of fire movement in urbanized areas.
Data purpose, 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 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: 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 WFDS model are ASCII files describing the fire behaviour.
Data collection: methods, instruments, ethical considerations	The external data are collected from: COPERNICUS, ERA5, Land cover, DEM UCAR: Static geographic data ECMWF: Global weather forecasts
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	Output data from WRF-SFIRE-CHEM may be used for forensic use in past events, Local and regional administrations can use output data for

useful, outside the project	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 - WFDS 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 WILDFIRES use case

WILDFIRES FAIR principles	
Findability: Directory structure, name convention, persistent identifiers	Still under consideration
Findability: Metadata: standards, keywords, indexing opportunities	Still under consideration
Accessibility: Repository for preservation and sharing the data	No repository is available now. Produced data is stored on PSNC Altair system, where computations take place.
Accessibility: Data availability: open or restricted, access protocol	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.
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,	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.

codebooks, data cleaning, analyses, variable definitions, and units of measurements)	
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 WILDFIRES use case

WILDFIRES 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.

4. Conclusions

HiDALGO2 provides tools and processes for users and customers, including simulation, analytics, visualisation, and smart control mechanisms to establish and improve HPC workflow efficiency. Data Management is crucial for the project since all the steps in the process require and produce data.

This document sets up the methodology for the HiDALGO2 DMP by following the guidelines of the European Commission and the HiDALGO2 Grant Agreement. The methodology is applied to all HiDALGO2 pilots and thus provides their dataset description. It is important to note that data sets may change over time, particularly when data-based AI methods are developed for the applications, so regular updates and continuous data management are necessary. HiDALGO2 understands this challenge and aims to use this document as a starting point for managing current and future data. Also, elaborated technical datasets, for example, for particular use cases of the pilots, and other non-directly software-related information will be added according to the methodology introduced in this document.

References

[1] HE Programme Guide: V3.0 – 01.04.2023, https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf, retrieved 2023-06-21.

[2] HiDALGO2 Grant Agreement. EuroHPC Joint Undertaking, Project 101093457

[3] ARGOS data management platform, https://argos.openaire.eu/, retrieved 2023-05-10.