http://www.everest-h2020.eu

dEsign enVironmEnt foR Extreme-Scale big data analyTics on heterogeneous platforms

D1.3 — Interim Data Management Plan

The EVEREST project has received funding from the European Union's Horizon 2020 Research & Innovation programme under grant agreement No 957269

Project Summary Information

Project Title	dEsign enVironmEnt foR Extreme-Scale big data analyTics on heterogeneous platforms
Project Acronym	EVEREST
Project No.	957269
Start Date	01/10/2020
Project Duration	36 months
Project Website	http://www.everest-h2020.eu

Copyright

Num.	Partner Name	Short Name	Country
1 (Coord.)	IBM RESEARCH GMBH	IBM	CH
2	POLITECNICO DI MILANO	PDM	IT
3	UNIVERSITÀ DELLA SVIZZERA ITALIANA	USI	CH
4	TECHNISCHE UNIVERSITAET DRESDEN	TUD	DE
5	Centro Internazionale in Monitoraggio Ambientale - Fondazione CIMA	CIMA	IT
6	IT4Innovations, VSB – Technical University of Ostrava	IT4I	CZ
7	VIRTUAL OPEN SYSTEMS SAS	VOS	FR
8	DUFERCO ENERGIA SPA	DUF	IT
9	NUMTECH	NUM	FR
10	SYGIC AS	SYG	SK

Project Coordinator: Christoph Hagleitner – IBM Research – Zurich Research Laboratory

Scientific Coordinator: Christian Pilato – Politecnico di Milano

Disclaimer

Deliverable Information

Work-package	WP1
Deliverable No.	D1.3
Deliverable Title	Interim Data Management Plan
Lead Beneficiary	NUM
Type of Deliverable	ORDP: Open Data Research Plan
Dissemination Level	Public
Due Date	31/03/2022

Document Information

Delivery Date	31/03/2022
No. pages	24
Version | Status	0.4 | Final
Responsible Person	Fabien Brocheton (NUM)
Authors	Fabien Brocheton (NUM); Antonella Galizia (CIMA), Katerina Slaninova (IT4I), Radim Cmar (SYG), Guido Rusca (DUF), Riccardo Cevasco (DUF), Dionysios Diamantopoulos (IBM)
Internal Reviewer	Riccardo Cevasco (DUF)

The list of authors reflects the major contributors to the activity described in the document. All EVEREST partners have agreed to the full publication of this document. The list of authors does not imply any claim of ownership on the Intellectual Properties described in this document.

Revision History

Date	Ver.	Author(s)	Summary of main changes
19/01/2022	0.1	Fabien Brocheton (NUM); Antonella Galizia (CIMA); Katerina Slaninova (IT4I); Radim Cmar (SYG); Guido Rusca (DUF); Riccardo Cevasco (DUF); Dionysios Diamantopoulos (IBM)	Initial draft
19/03/2022	0.4	Dionysios Diamantopoulos (IBM)	Added data management for IBM ZYC2 infrastructure

Quality Control

Approved by Internal Reviewer	March 28, 2022
Approved by WP Leader	March 28, 2022
Approved by Scientific Coordinator	March 30, 2022
Approved by Project Coordinator	March 31, 2022

Table of Contents

1 Executive Summary

The document is an update of the Deliverable D1.2. It describes the current status of the Data management plan (DMP) at M18, specially in regard to research data available outside the project.

Indeed, EVEREST project has chosen to be part of the Open Research Data (ORD) pilot of the H2020 program[1]. The ORD Pilot aims to improve and maximize access and re-use of research data generated by Horizon 2020 projects and considers the need to balance openness and protection of scientific information, commercialization and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions. The ORD pilot applies:

- Primarily to the data needed to validate the results presented in scientific publications.
- Other data can also be provided by the beneficiaries on a voluntary basis

There are two main pillars in the Pilot:

- Develop (and keep up to date) a Data Management Plan (DMP)
- Provide open access to research data (i.e., implement the DMP):
  - Deposit our data in a “research data repository”.
  - Ensure third parties can freely access, mine, exploit, reproduce, and disseminate our data.
  - Provide related information and identify (or provide) the tools needed to use the raw data to validate our research.

Allowing data to be Findable, Accessible, Interoperable and Reusable corresponds to the **FAIR data concept** requested by the ORD pilot.

This document describes the current decisions and the plans for the next months, and the partners plan to update it regularly during the project when new data will come in.

1.1 Structure of this Document

The document is organized as follows:

- Section 3 presents EVEREST data management policies.
- Section 4 presents methodology for access of data.
- Section 5, Section 6, and Section 7 describe the research data managed in the three pilot use cases that we plan to make available for research outside the project.

1.2 Related Document

D2.1 and D6.1 – More details on the EVEREST use cases
D2.3 – More details on the data managed inside the project
D3.1 – More details on EVEREST data management techniques inside the project.

[1] All details are available on https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

2 Introduction

The next figure presents the EVEREST Data Lifetime (EDL). EDL is a high-level flowchart, depicting how data are managed. Since "data management" may have multi-dimensional aspects, we divide EDL in three main categories:

- Data gathering: The process of collecting data from various sources to process in the following stage.
- Experimentation: The main process of performing calculation on data to derive useful insights for the EVEREST applications.
- Data sharing: The process of offering the results of the previous stage to other interested parties, either in confidential, open-access policy or as an input to the gathering stage in the form of a feedback loop (reintegration).

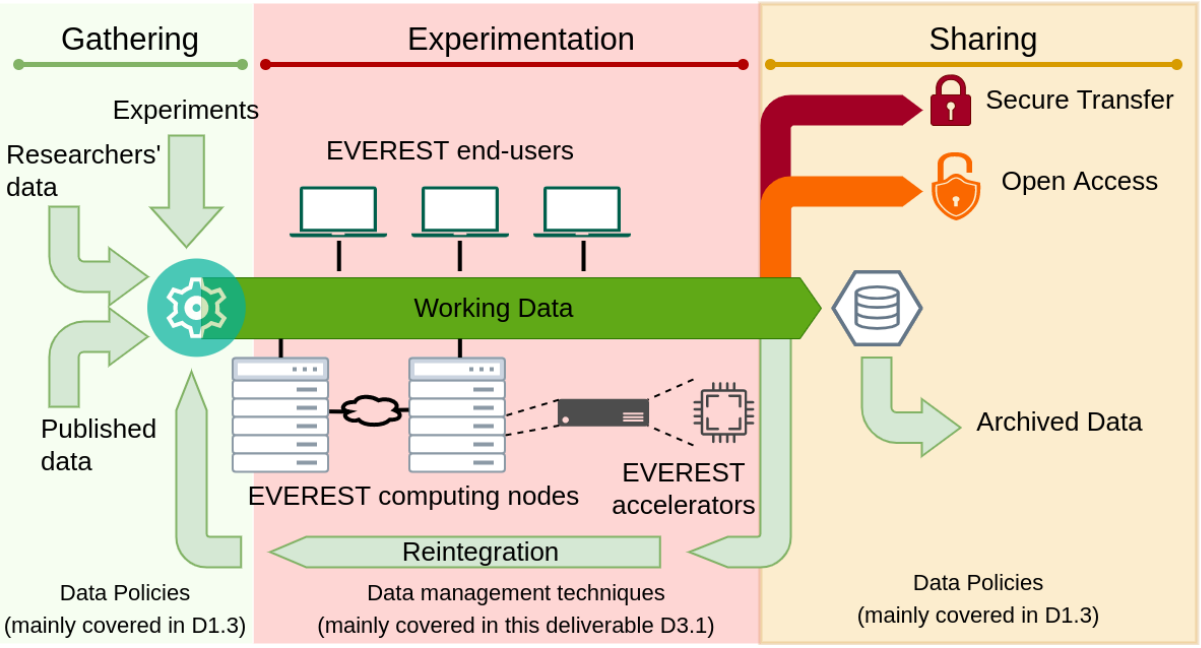

Figure 1 – Representation of the data management plan in EVEREST

In D1.3, we mainly focus on "gathering" and "sharing" parts, whereas EVEREST data management techniques are described in D3.1 Section 3 describes the data policies. Section 4 describes how the open data for research will be published and accessible, as well as how data inside the project are shared and store. Section 5, Section 6, and Section 7 describe the data used and produced by each use case, with a focus on the open research data.

3 General Data Management on EVEREST Servers

The EVEREST infrastructure is divided in two parts:

- Servers shared between some EVEREST partners to execute the workflow. This concerns servers from **IBM** and **IT4I**.
- Partner's servers where some input data are processed before to be ingested inside the workflow, and where some output data produced by the workflow are treated (visualisation, transfer to third parties, ...). This concerns **NUM**, **CIMA**, **IT4I**, and **SYG**.

3.1 Data management and security on servers shared between partners

3.1.1 IBM Servers

In the EVEREST project IBM is providing access to the cloudFPGA research platform through the the Zurich Yellow security zone Compute Cluster (ZYC2) of the IBM Zurich Research Laboratory. The access is granted based on the "international agreement for early release of IBM SaaS and hosted program offerings". Such an agreement needs to be signed by all EVEREST user upon the "process of granting access permissions" (Section 3.1.1.4).

3.1.1.1 Content

IBM provides only services for Content. IBM is not the publisher of Content transmitted within IBM's cloudFPGA platform. EVEREST users have sole responsibility for the following:
a. ensuring the adequacy of any IBM cloudFPGA elements to satisfy any EVEREST user requirements:
b. all Content including, without limitation, its selection, creation, design, licensing, installation, accuracy, maintenance, testing, backup and support;
c. having all necessary authorizations to allow IBM and its subcontractors to host, cache, record, copy, and display Content, and Customer represents that it has and will keep in effect during its use of IBM cloudFPGA platform all such authorizations and approvals necessary to grant IBM and its subcontractors these rights, and that such rights are provided at no charge to IBM. EVEREST users retain all right, title, and interest in and to its Content; and
d. the selection and implementation of procedures and controls regarding access, security, encryption,use, transmission, and backup and recovery of Content.

3.1.1.2 Personal Data

In EVEREST no personal data will be collected, as informed by D8.2. For any other applications that use IBM infrastructure which might also be used by EVEREST (e.g. VPN service), any personal data provided as part of those applications will be collected, processed and/or utilized exclusively for the purposes it has been provided for. No personal data will be transmitted to third-party organizations and only aggregate statistics, without the possibility to identify the single individuals, may be made public. Any users of that IBM infrastructure have the right to ask for a copy of the information held in IBM's records. Such users are entitled to request the correction and, if the legal requirements are fulfilled, the disabling or deletion of their personal data.

3.1.1.3 Customer Data and Databases

The IBM cloudFPGA platform is not intended for the storage or receipt of any Sensitive Personal Information or Protected Health Information (as defined below), in any form. EVEREST users will not send or provide IBM

access to any Sensitive Personal Information or Protected Health Information, whether in data or any other form and EVEREST users will be responsible for reasonable costs and other amounts IBM may incur relating to any such information provided to IBM or the loss or disclosure of such information by IBM, including those arising out of any third party claims.

3.1.1.4 Process of granting access permissions

The following steps are required to grant access permissions to ZYC2

Step-1: Registration of a ZYC2 account

- send an email to ZYC2's administrator
  - In the subject line
    - Indicate " ZYC2 Access Request ".
  - In the email body
    - Give a short description of your purpose for using IBM's EVEREST compute infrastructure,
    - Declare your acceptance of the " ZYC2 Access Agreement " and attach the file to your email,
    - Declare your acceptance of the " ZYC2 Data Usage Agreement " and attach the file to your email.
- an email will be provided by the ZYC2 administrator with
  - User's credentials (login and password) to access ZYC2,
  - an OpenVPN configuration file,
  - some instructions for getting started in ZYC2.

Step-2: Setup of a Virtual Private Network (VPN) connection

- Access to ZYC2 is provided to EVEREST users via OpenVPN. The following steps are needed
  - Downloading and installation of an OpenVPN client (must be version 2.4 or higher),
  - Installation of the OpenVPN configuration file provided in Step 1.

Step-3: Setup and Management of ZYC2 Virtual Resources

- ZYC2 runs OpenStack to provide an Infrastructure-as-a-Service Cloud. As such, it lets EVEREST users to create and manage their own virtual networks and machines in ZYC2. A ZYC2 User Guide is provided to EVEREST users with detailed following-up instructions.

3.1.1.5 Data storage areas

In general, IBM is not providing any long-term storage or backup service in the EVEREST project, but it offers only ephemeral storage for the necessities of the execution of the EVEREST application workflows. As such, it provides a single storage area named HOME for the user crated VMs in the ZYC2. HOME storage is designed for low-term to mid-term storage of data. Depending on the VM configuration selection, the users may choose the desired HOME storage area from a predefined set of available storage space.

3.1.1.6 Data access

Remote access and electronic security.

- All external access to IBM resources is provided only through encrypted data channels (SSH, SFTP, SCP and OpenVPN).

- Control of permissions on the operating system level is done via standard Linux facilities – classical UNIX permissions (read, write, execute granted for user, group or others) and Extended ACL mechanism (for a more fine-grained control of permissions to specific users and groups).

3.1.1.7 Data lifecycle

- Transfer of data to IBM: User transfers data from his facility to IBM only via safely encrypted and authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.
- Data within IBM: Once the data are at IBM data storage, access permissions apply.
- Transfer of data from IBM: User transfers data to facility from IBM only via safely encrypted and authenticated channels (SFTP, SCP). Users are strongly advised not to initiate unencrypted data transfer channels (such as HTTP or FTP) to remote machines.

3.1.2 IT4I servers

In the EVEREST project, IT4I will provide its HPC infrastructure (Karolina and Barbora) and experimental cloud (LEXIS OpenStack) in the extent necessary for the project, see more details in D6.2. The access to the HPC infrastructure is provided according to the Allocation policy through the Open access Competitions, Directors Discretion, PRACE (DECI) program or EuroHPC JU Access Calls[2].

3.1.2.1 Human roles and administration process

IT4I System Administrators are full-time internal employees of IT4I, department of Supercomputing Services. The system administrators are responsible for safe and efficient operation of the computer hardware installed at IT4I. Administrators have signed a confidentiality agreement.

User access to IT4I supercomputing services is based on projects — membership in a project provides access to the granted computing resources (accounted in core-hours consumed). Each project will have one Primary Investigator, a physical person, who will be responsible for the project, and is responsible for approving other users' access to the project. At the beginning of the project, **Primary Investigator** will appoint one Company Representative for each company involved in the project.

Company Representatives will be responsible for approving access to **Private Storage Areas** belonging to their company. Private Storage Areas are designated for storing sensitive private data. Granting access permissions to a Private Storage area must be always authorized by the respective Company Representative and Primary Investigator. Available on request.

Users are physical persons participating in the project. Membership of users to EVEREST project is authorized by Primary Investigator. Users can log in to IT4I compute cluster, consume computing time and access shared project storage areas. Their access to Private Storage Areas is limited by permissions granted by Company Representatives.

User data in general can be accessed by:

1. IT4Innovations System Administrators.
2. The user, who created them (i.e. the UNIX owner).
3. Other users, to whom the user has granted permission and at the same time have access to the particular Private Storage Area (in the case of data stored in the Private Storage Area) granted via the "Process of granting of access permissions".

[2] IT4I Documentation: https://docs.it4i.cz

3.1.2.2 Process of granting access permissions

All communication with participating parties is in the manner of signed email messages, digitally signed by a cryptographic certificate issued by a trusted Certification Authority. All requests for administrative tasks must be sent to IT4I HelpDesk. All communication with Help Desk is archived and can be later reviewed.

Access permissions for files and folder within the standard storage areas (HOME, SCRATCH) can be changed directly by the owner of the file/folder by respective Linux system commands. The user can request Help Desk for assistance on how to set the permissions.

Access to Private Storage Areas is governed by the following process:

1. A request for access to Private Storage Area for given user is sent to IT4I HelpDesk via a signed email message by a user participating in the project.
2. HelpDesk verifies the identity of the user by validating the cryptographic signature of the message.
3. HelpDesk sends a digitally signed message with request of approval to the respective Company Representative and to the Primary Investigator.
4. Both the Company Representative and the Primary Investigator must reply with a digitally signed message with explicit approval of the access to the requested Private Storage Area.
5. System administrator at HelpDesk grants the requested access permission to the user.

Company representative or Primary Investigator can also send a request to HelpDesk to revoke access permission for a user.

3.1.2.3 Data storage areas

There are five types of relevant storage areas: HOME, SCRATCH, PRIVATE and PROJECT Data Storage. HOME, SCRATCH, and PROJECT Data Storage are standard storage areas provided to all users of IT4I supercomputing resources (file permissions apply).

HOME storage is designed for long-term storage of data. SCRATCH is a fast storage for short- or mid-term data, with no backups. PRIVATE storages are dedicated storages for sensitive data, stored outside the standard storage areas. The PROJECT data storage is a central storage for projects' /users' data on IT4I and is accessible from all IT4I clusters and allows to share data amongst clusters.

HOME Storage The HOME filesystem is an HA cluster of two active-passive NFS servers.

This filesystem contains users' home directories /home/username. By default, the permissions of the home directory are set to 750, and thus it is not accessible by other users.

Accessible capacity is 31 TB, shared among all users. Individual users are restricted by filesystem usage quotas, set to 25 GB per user. Should 25 GB prove insufficient, it is possible to contact support, the quota may be increased upon request.

The files on HOME filesystem will not be deleted until the end of the user's lifecycle.

SCRATCH storage The SCRATCH filesystem is realized as a parallel Lustre filesystem. It is accessible via the Infiniband network and is available from all login and computational nodes.

Extended ACLs are provided on the Lustre filesystems for sharing data with other users using fine-grained control.

The SCRATCH filesystem is mounted in directory /scratch. Users may freely create subdirectories and files on the filesystem. Accessible capacity is 1000 TB, shared among all users. Users are restricted by PROJECT quotas set to 20 TB. The purpose of this quota is to prevent runaway programs from filling the entire filesystem and deny service to other users. Should 20 TB prove insufficient, it is possible to contact support, the quota may be increased upon request.

The Scratch filesystem is intended for temporary scratch data generated during the calculation as well as for high-performance access to input and output files. All I/O intensive jobs must use the SCRATCH filesystem as their working directory.

Users are advised to save the necessary data from the SCRATCH filesystem to HOME filesystem after the calculations and clean up the scratch files.

Files on the SCRATCH filesystem that are not accessed for more than 90 days will be automatically deleted.

PRIVATE storage In order to provide additional level of security of sensitive data, we will setup dedicated storage areas for each company participating in the project. PRIVATE storage areas will be setup in a separate storage and will be not accessible to regular IT4Innovation users. IT4I can additionally provide encryption of PRIVATE storage; the particular solution will be discussed with regards to security and performance considerations.

PRIVATE BACKUP storage It is possible to setup dedicated backups of PRIVATE storage. In this case we can guarantee secure removal of data archived in PRIVATE BACKUP.

PROJECT Data Storage The PROJECT Data storage is a central storage for projects'/users' data on IT4I. The PROJECT Data storage is accessible from all IT4I clusters and allows to share data amongst clusters. The storage is intended to be used throughout the whole project's lifecycle.

All aspects of allocation, provisioning, accessing, and using the PROJECT storage are driven by project paradigm. Storage allocation and access to the storage are based on projects (i.e. computing resources allocations) and project membership.

A project directory (implemented as an independent fileset) is created for every active project. Default limits (quotas), default file permissions, and ACLs are set. The project directory life cycle strictly follows the project's life cycle.

The PROJECT storage is not primarily intended for computing. The project directory is removed after the project's data expiration. Data on the PROJECT storage is not backed up.

3.1.2.4 Data access

Physical security All data storage is placed in a single room, which is physically separated from the rest of the building, has a single entry door and no windows. Entry to the room is secured by electromechanical locks controlled by access cards with PINs and non-stop alarm system. The room is connected to CCTV system monitored at reception with 20 cameras, recording and backup. Reception of the building has 24/7 human presence and external security guard during night. Reception has a panic button to call a security agency.

Remote access and electronic security All external access to IT4I resources is provided only through encrypted data channels (SSH, SFTP, SCP and FortiGate VPN).

Control of permissions on the operating system level is done via standard Linux facilities – classical UNIX permissions (read, write, execute granted for user, group or others) and Extended ACL mechanism (for a more fine-grained control of permissions to specific users and groups). PRIVATE storage will have another level of security that will not allow mounting the storage to non-authorized persons.

3.1.2.5 Data lifecycle

1. **Transfer of data to IT4I:** User transfers data from his facility to IT4I only via safely encrypted and authenticated channels (SFTP, SCP). Unencrypted transfer is not possible.
2. **Data within IT4I:** Once the data are at IT4I data storage, access permissions apply.
3. **Transfer of data from IT4I:** User transfers data to facility from IT4I only via safely encrypted and authenticated channels (SFTP, SCP). Users are strongly advised not to initiate unencrypted data transfer channels (such as HTTP or FTP) to remote machines.

4. **Removal of data:** On SCRATCH file system, the files are immediately removed upon user request and if not accessed for more than 90 days. PRIVATE storage will be securely deleted upon request or when the project ends.

3.1.2.6 Data in a computational job life cycle

When a user wants to perform a computational job on the supercomputer the following procedure is applied:

1. User submits a request for computational resources to the job scheduler
2. When the resources become available, the nodes are allocated exclusively for the requesting user and no other user can login during the duration of the computational job. The job is running with same permissions to data as the user who submitted it.
3. After the job finishes, all user processes are terminated, and all user data is removed from local disks (including ram-disks).
4. After the clean-up is done, the nodes can be allocated to another user, no data from the previous user are retained on the nodes.

All Karolina and Barbora computational nodes are disk-less and cannot retain any data.

There is a special SMP server Superdome Flex accessible via separate job queue, which has different behaviour from regular computational nodes: it has a local hard drive installed and multiple users may access it simultaneously.

3.1.2.7 ISO certification

IT4I has established and continually improves an internationally recognized information security management system, manages risks, and has established processes and regulations to secure information against misuse, unauthorized changes, and loss. Since December 2018, IT4I has been an Information Security Management System certificate holder according to the international ISO/IEC 27001:2013 standard. This certificate has been awarded for the following areas: provision of national supercomputing infrastructure services, high-performance computing problems solutions, the performance of advanced data analysis and simulations, and processing of large data sets.

3.2 Data security for communication with private partner's servers

In EVEREST, data are exchanged between different location and different partners. Data transfer happens mostly over the internet. Because of this, it is fundamental to ensure that data are handled and accessed only by the legitimate users.

To achieve this, EVEREST uses state of the art secure transfer protocols. In particular, the communication between the premises of different partners happens only in encrypted form (using SSH or a VPN system) and the data are exchanged using the SFTP or the SCP protocols. To further guarantee the security of the whole infrastructure, the systems used are regularly updated, so to ensure protections from known vulnerabilities. Credential to access the servers are given to partners upon request and verification.

4 Access to datasets

4.1 Access to open dataset

Concerning the open dataset associated to each use case, we decided to use the Zenodo platform (https://zenodo.org) in order to provide access to them. Publication on Zenodo will be linked for each dataset to OpenAIRE platform (https://www.openaire.eu/).

The EVEREST community has been already created inside Zenodo platform (https://zenodo.org/communities/everest/). At this stage, we have identified datasets from each use case (see Section 5, Section 6, and Section 7) which will be published on Zenodo.

Open data provided by the EVEREST project will meet the requirements for the FAIR data policy. To make them accessible, metadata will be defined as much as possible for each dataset.

We propose to follow international standards, especially DataCite (DataCite - https://en.wikipedia.org/wiki/DataCite, https://schema.datacite.org/meta/kernel-4.2/index.html, https://schema.datacite.org/meta/kernel-4.2/doc/DataCite-MetadataKernel_v4.2.pdf - Table 4) which defines a minimun set of information to provide (description, producer, ...).

There are 6 mandatory fields for DataCite:

- Identifier (normally a DOI – Digital Object Identifier),
- Creator (main researchers involved in producing the data),
- Title (of the data set),
- Publisher (entity which holds/archives/publishes/produces data),
- PublicationYear (year when data became/become/will become public),
- ResourceType.

Each of these fields contains very few subfields for further description. Additional metadata can be added, especially in relation to scientific discipline (road traffic science, weather science, ...) of the dataset which can request specific details/keywords.

We can mention that publication on Zenodo requests to fill fields in agreement with this specification. Specially, Zenodo attributes a specific adn persistent DOI to each published dataset). Additionally, publication on Zenodo requests to define access/usage rights by choosing an exploitation license.

For the two use cases based on WRF simulation, the open dataset will be limited to the dataset used for AI process. But there is an open question to be able to publish the full WRF outputs over simulated domain (France and Italy). The size is around 2 To for one domain and one year which is nearly impossible to manage with Zenodo with limitation of 50 Go. In the next months, an evaluation will be performed if it is possible to publish such large dataset in a sustainable way after the project.

Concerning source code, we focus on this first part of the project to share internally the different applications codes and EVEREST SDK. In the next months we will start also to identify the codes will be published also on Zenodo.

4.2 Internal access to dataset and storage

Inside the EVEREST infrastructure, we decided to rely of the IT4I storage to store permanent data (input and output of applications).

Within the course of the project, it is envisaged that the IT4I Distributed Data Infrastructure (DDI), based on iRODS and EUDAT (https://eudat.eu) B2SAFE will be used. Not being meant as a competitor to platforms

like Zenodo or GitHUB, it serves the special purpose of an efficient distributed data management within the project. Publication of the data can be realised through EUDAT B2 services (including PID). Data access is provided via the LEXIS web portal, and Interoperability and Reproducibility are fostered by a common metadata standard (apart from measures within the pilot use cases themselves), as we will describe in the following.

Concerning exchange of source code between partners a GitLab repository has been created and is accessible on IT4I servers (https://code.it4i.cz/).

CI/CD infrastructure is available as well, including runners with custom images or shell executors. The runners are supposed to be used for compilation and performing a set of tests. It is not supposed to run a heavy computation pipeline.

5 Air Quality Use Case Data Management Plan

In this section, the air quality pilot data management plan will be described.

5.1 Data Summary

The context of the use case is the forecast of air pollution peak (or not peak) due to the emission of an industrial site in order for this site to manage its impact and prevent such pollution event on the population. The objective is to improve the quality of the forecast by avoiding false pollution peaks (financial loss for the site) or miss pollution peaks (health impact on population).

This use case relies on three main steps for each air quality forecast simulation:

Step-1: Compute deterministic (short range) and/or probabilistic (nowcasting) meteorological forecast with the WRF model[3]. During each simulation, an assimilation procedure in order to force computation by observations is activated. This part is executed typically on HPC server or can be executed also on FPGA-accelerated platform for selected kernels.

Step-2: Combine this deterministic (short range) and/or probabilistic (nowcasting) meteorological forecast with another weather forecasts and local measurement in order to obtain, by machine learning approach, a better weather forecast. This part is executed on a cloud server.

Step-3: Compute deterministic air quality forecast with the ADMS5 model[4], based either on the weather forecast from the Step 1 or the Step 2. This part is executed on Windows cloud server.

Step 1 requires the following data sets:

- Global forecast dataset (GFS) produced by NCEP (US National Center for Environmental Prediction) or Integrated Forecasting System (IFS) produced by ECMWF (European Centre for Medium-Range Weather Forecasts)
- For the assimilation and validation procedures:
  - Surface weather observation data (hydrometeorological variables, e.g. temperature, water vapour, wind speed and direction) provided by authoritative (e.g. ECOMET) and personal weather (e.g. underground) stations network.
  - Other hydrometeorological variables (e.g. reflectivity, soil moisture, land surface temperature, sea surface temperature, integrated water vapor content) retrieved from ground-based (radar) and space-borne remote sensing system (Sentinel data from the European Copernicus service).

The outputs of the Step 1 are meteorological parameters for the simulated domain on a 3D grid.

Step 2 requires also to download external input data:

- Another dataset of weather forecast: forecast produced by NUMTECH at different scales (Europe/France); an extension of the GFS dataset used at Step 1, eventually a forecast from Meteo France.
- For the learning phase, local weather surface observation at industrial site.

The outputs of the Step 2 are meteorological parameters at the location of the industrial site (where local observation data are provided).

For the Step 3, the outputs of Step 1 and Step 2 are intermediate data used as input data. Another input data are emission data for the forecast period. The outputs are 2D (at surface) grid of pollutants concentrations.

Most of the datasets describe above are open-source data available freely, except:

[3] https://www.mmm.ucar.edu/weather-research-and-forecasting-model
[4] http://cerc.co.uk/environmental-software/ADMS-model.html

- **NUM** forecast data which are normally commercial ones.
- Local weather observation data at industrial site which are private data.
- Some third-party observation data used for assimilation (radar data or personal surface station) which are commercial ones.

In the framework of the EVEREST project, these datasets could be used freely internally.

The output data relevant to the air quality pilot are proprietary model outputs.

We identify two datasets which can be share for open research:

- The WRF model output from Step 1 described above, associated to the other numerical weather forecasts and the weather observations used as input in Step 2.
- The ADMS model output from Step 3 described above.

5.2 FAIR data

Making data findable, including provisions for metadata The data will be stored on data repositories with digital object identifiers. We will choose in priority public repositories as long as they allow us to comply with the constraints on the datasets access. Datasets will all have a metadata description, and, in the case of datasets with access restrictions, their metadata will be publicly available. A semantic versioning scheme will be used to track versions of the datasets. The partner responsible for generating that data will be the point of contact for requesting an access to the data.

Making data openly accessible For the air-quality use case, there will be no restrictions on the use of dataset for only research activity in any domain. No control on the use of the dataset will be done, except at the downloading step where applicants in order to use dataset must declare for which research activity they want to do it, and accept to quote EVEREST in acknowledgements in case of publication.

In order to make data interoperable and increase the data re-use, the research data will be provided on an open format and not in a proprietary model format.

5.3 Plan of the outputs

The research outputs produced by the use case are:

- Weather dataset: WRF model output from EVEREST simulation associated to additional weather forecast from external providers and local weather measurement. the dataset will concern wind speed, wind direction and temperature at surface for different locations in France. Such dataset can be used by researchers who want to work on local weather data assimilation or ensemble approach; or people who want to develop and test method of bias correction of numerical simulation compared to observation. User community can be then any people who exploit weather dataset and need at the same location "real" weather forecast and weather measurement for testing methods or applications in various sectors (agriculture, sport, industry, mobility, ...).
- The ADMS model output: External researchers will have access to real temporal and spatial variations of atmospheric impacts of some industrial sites which can be the input of other applications as for example health impact assessment protocols or design measuring devices.

  The dataset can be also exploited by people who work on fusion/assimilation of air-quality measurement and simulation, especially if the objective is to propagate a correction not only locally but also spatially. To develop and test such approaches, this requires to have coherent 1D measurement and 2D simulations over a domain, which is not so easy. One way is to exploit 2D simulation outputs (as provided) to create complementary datasets by extracting values for some grid points and introducing random variation in order to emulate measurement at these grid points.

For each dataset, a specific table provides additional information. The dataset will be shared on Zenodo when they are produced, so mainly during the last 6 months of the project.

Table 1 – Research dataset for AIR QUALITY use case.

ID	ITEM	DESCRIPTION
D1	Dataset name and reference	Local weather forecast and weather observation (zenodo reference will be updated after publication)
	Dataset description	1D meteorological fields, over different locations (France) at hourly temporal resolution for a selected period. Dataset use for AI application
	Standards, format and metadata	Csv format
	Is dataset confidential? Must be encrypted?	Not confidential
	Data sharing/access inside EVEREST	Yes
	Data sharing/access outside EVEREST for research	Yes (Free Access)
	Is dataset reusable?	Yes (research activity only)
	Archiving and preservation (including storage and backup)	During the project: Outputs will be stored for the duration of the EVEREST project, with a focus on parameters used for air quality application. After the project: Outputs will be stored on **NUM** storage system.
D2	Dataset name and reference	Air-quality forecast output (zenodo reference will be updated after publication)
	Dataset description	2D air-quality concentrations simulated by ADMS over different locations (France) at hourly temporal resolution, for a selected period.
	Standards, format and metadata	Csv format
	Is dataset confidential? Must be encrypted?	Spatial coordinates will be “anonymized” (relative coordinates) in order to not have a link to an industrial site
	Data sharing/access inside EVEREST	Yes
	Data sharing/access outside EVEREST for research	Yes (Free Access)
	Is dataset reusable?	Yes (research activity only)
	Archiving and preservation (including storage and backup)	During the project: Outputs will be stored for the duration of the EVEREST project, with a limited set of pollutants. After the project: Outputs will be stored on the **NUM** storage system.

6 Renewable Energy Production Use Case Data Management Plan

This section describes the DMP for the renewable energy production pilot.

6.1 Data Summary

The context of the use case is the prediction of the renewable energy produced by a wind farm reducing the risks related to severe meteorological ramp-up/down events. The objective is to improve the quality of the forecast by reducing the related uncertainty minimizing the prediction error for renewable energy trading avoiding false production peaks or miss production peaks.

The following table shows the wind farms dispatched by **DUF** that will be considered in this study:

Wind Farm code	Electricity Zone	Power capacity MW	Yearly production [MWh]
UP_SARD	SARD	29,75	55.000
UP_SICI	SICI	33,15	66.000
UP_CSUD	CSUD	17,85	26.600
UP_SUD	SUD	18,00	47.827
UP_CALA	CALA	34,00	80.000

This use case relies on three main steps for each energy production prediction simulation:

Step-1: Compute deterministic (short range) and/or probabilistic (nowcasting) with the WRF model. Data assimilation procedures are applied to force computation through atmosphere observations in order to improve weather prediction. Due to the high computational and memory requirements, HPC resources are mandatory to run the simulations. Some execution parts can be executed also on EVEREST FPGA-based systems.

Step-2: Concerns with deterministic prediction of hourly energy generation, in the site-specific meteorological conditions forecasted by WRF model.

Step-3: Compute energy production forecast, based either on the results obtained from the Step 1 or the Step 2 and with machine learning approach for implementation of site-specific data

Step 1 requires different input data the main ones are:

- Global forecast dataset (GFS) produced by NCEP (US National Center for Environmental Prediction) or Integrated Forecasting System (IFS) produced by ECMWF (European Centre for Medium-Range Weather Forecasts)
- Some surface observation data (sea surface temperature, humidity of soils, etc.).
- For the assimilation procedure:
  - Additional Surface weather observation data (hydrometeorological variables, e.g., temperature, water vapours, wind speed and direction) provided by authoritative (e.g., ECOMET) and personal weather (e.g., underground) stations network.
  - Another hydrometeorological variables (e.g., reflectivity, soil moisture, land surface temperature, sea surface temperature, integrated water vapor content) retrieved from ground-based (radar) and spaceborne remote sensing system (Sentinel data from the European Copernicus service).

The outputs of the Step 1 are meteorological parameters for the simulated domain on a 3D grid.

Step 2 requires site specific and technical additional data for energy production estimation (e.g., turbine power curve)

For the Step 3, the outputs of Step 1 and Step 2 are used as input data. The outputs are horizontal wind field in the lowest part of the atmosphere on a 2D grid of spacing around 2-3 km and hourly power generation predicted by deterministic model.

In this step site specific observation data and historical data are provided by **DUF**.

Most of the dataset describe above are open-source data available freely, except:

- **DUF** site specific historical and observation data (e-g. hourly basis generation, wind speed...), confidential and to not share publicly.
- **DUF** forecast data, if necessary, to model training or assessment, which are commercial ones, used freely in the framework of this research project, but not possible to share publicly.
- no implicit or explicit references to the wind farm or the owner companies must be published
- Local weather observation data at the wind farm location site which are private data (to keep confidential inside the EVEREST project).
- Some third-party observation data (radar data or personal surface station) which are available only for research purposes within the scope of the project and available after approval of the owner (e.g., Italian Civil Protection Department). In the framework of the EVEREST project, these datasets could be used freely, but not possible to share publicly.

Most of the output data relevant to the energy production pilot are proprietary model output. Categories of the research outputs produced by the use case are listed as follows:

- The WRF model output from Step 1 described above.
- The energy prediction model output from Step 3 described above.

6.2 FAIR data

Making data findable, including provisions for metadata The data will be stored on data repositories with digital object identifiers. We will choose in priority public repositories as long as they allow us to comply with the constraints on the datasets access. Datasets will all have a metadata description, and, in the case of datasets with access restrictions, their metadata will be publicly available. A semantic versioning scheme will be used to track versions of the datasets. The partner responsible for generating that data will be the point of contact for requesting an access to the data.

Making data openly accessible For the energy production use case, there will be no restrictions on the use of dataset for only research activity in any domain. No control on the use of the dataset will be done, except at the downloading step where applicants which want to use a dataset must declare for which research activity and accept to quote EVEREST in acknowledgements in case of publication.

In order to make data interoperable and increase the data re-use, the research data will be provided on an open scientific format largely used in the meteorology applications, and not in a proprietary model format. The format is NetCDF (Network Common Data Form) devoted for storing multidimensional scientific data including inside its metadata description. Open-source tools are available in order to read easily this format.

6.3 Plan of the outputs

The research outputs produced by the use case are:

- The WRF model output: External researchers will have access to a new source of local weather forecast over Italy for their own applications in order to evaluate the impact of assimilation and procedures applied in the framework of EVEREST, without any limitation on the application domain except the list of weather parameters provided in the dataset.
- The energy production model output: External researchers will have access of energy prediction from windfarms which can be input requested by another application models, such as for example models to determine when maintenance could be applied, etc.

For each category, a specific table provides additional information. A first version of the Dataset D3 will be published before the mid-term review meeting.

Table 2 – Research dataset for ENERGY PRODUCTION use case.

ID	ITEM	DESCRIPTION
D3	Dataset name and reference	Local weather forecast (zenodo reference will be updated after publication)
	Dataset description	1D meteorological fields simulated by WRF, over different locations at hourly temporal resolution, for a selected period. Weather forecast used for AI application.
	Standards, format and metadata	NetCDF format
	Is dataset confidential? Must be encrypted?	Not confidential
	Data sharing/access inside EVEREST	Yes
	Data sharing/access outside EVEREST for research	Yes (Free Access)
	Is dataset reusable?	Yes (research activity only)
	Archiving and preservation (including storage and backup)	During the project: Outputs will be stored for the duration of the EVEREST project, with a focus on parameters used for air quality application. After the project: Outputs will be stored on the **DUF** storage system.
D4	Dataset name and reference	Energy prediction output (zenodo reference will be updated after publication)
	Dataset description	Energy production prediction over different locations (wind farms set in Italy) at hourly temporal resolution, for a selected period.
	Standards, format and metadata	Csv format
	Is dataset confidential? Must be encrypted?	Spatial coordinates will be “anonymized” (relative coordinates) to avoid any link to the wind farm sites
	Data sharing/access inside EVEREST	Yes
	Data sharing/access outside EVEREST for research	Yes
	Is dataset reusable?	Yes (research activity only)
	Archiving and preservation (including storage and backup)	During the project: Outputs will be stored for the duration of the EVEREST project. After the project: Outputs will be stored on **DUF** storage system.

7 Traffic Modeling Use Case Data Management Plan

This section describes the DMP for the traffic modelling use case.

7.1 Data Summary

Traffic modelling and prediction is a critical component for smart cities to build their intelligent traffic management system (ITS). The goal is to find a true traffic data model representation of the city, which is used for providing precise traffic predictions.

Our computation ecosystem starts with reading big raw sensory data, both real-time and long history records. Traffic simulator subsequently converts the sensory data into a traffic model as well as into rich training sequences for prediction model training. Next, a traffic prediction model is learnt from the training data set, finally being exploited by route calculation service.

The use case consists of four computationally and data intensive steps:

Step-1: Pre-processing of FCD data into road speed profiles using map matching algorithm. The computation should be accelerated with FPGA.

Step-2: Compute traffic 3D data model from road speed profiles and O/D matrix with the help of Traffic simulator. The computation is typically executed on the HPC server.

Step-3: Train and regularly update the prediction model based on traffic data model. The computation is typically executed on HPC server, Microsoft Windows Cloud platform, or can be executed also on FPGA-accelerated platforms.

Step-4: Online routing utilizing the trained prediction model. The computation is typically executed on HPC server but should also be accelerated with FPGA.

For Step 1, the data input is a large historical data set of floating car data (FCD). FCD is represented by geo-positions and the raw and noisy speeds of vehicles sensed approximately each 5 seconds from navigation devices, that is from millions of devices every day over the period of several years worldwide. However, our model will operate on selected cities only (like Vienna) counting thousands of vehicles daily and with the data from the time window of a limited period. The output is the road speed profiles processed from FCD data. The data is in the form of aggregated speeds over 10- or 15-minute intervals across the week period.

For Step 2, the data input is a provisioned origin-destination matrix (O/D) and the road speed profiles. The calculated 3D traffic model is represented by the three macroscopic variables, speed, flow, density, provided for main roads in a given city. The calculation is done by means of Traffic simulator, which also generates training sequences for traffic prediction learning. Additional dimensions are a seasonality attribute (month of the year) and weather condition factors.

For Step 3, the training sequences are used for training the prediction models. The large amount of vector samples can be obtained for each road element under prediction. The result of the step is a trained model (e.g., coefficients of the neural network) ideally for each road. The envisioned number of road elements typically relate to the number of main crossings in a city. As an example, the city of Vienna counts several thousand crossings.

For Step 4, the trained prediction models will be incorporated into the online routing **SYG** platform/simulator. A set of experiments based on different scenarios will be performed, e.g., online routing on the level of cities (smart-city routing) and routing on the level of country. The collected data will be provided in a raw format as well as a post-processed statistical overview including the used methodology.

The datasets used in our use case are with respect to the access scheme classified as follows:

- Sygic FCD data used in this use case will be available to EVEREST project, and in a limited scope can be publicly available as open source.

- O/D destination matrix is a purchased data under license terms, thus are private data with the possibility to keep confidentially available to EVEREST project.
- Historical weather data will be supplied by 3rd party so will be confidential, possibly open to EVEREST project
- Road speed profiles calculated by map matching of FCD data can be publicly available in a limited scope.
- Calculated 3D traffic model in the limited scope be shared publicly to the research projects.
- Training sequences for training road traffic prediction models can be publicly available in a limited scope
- Learned Prediction model with its test dataset results can be provided as publicly available benchmarks to research projects.

7.2 FAIR data

Making data findable, including provisions for metadata The data will be stored on data repositories with digital object identifiers. We will choose in priority public repositories as long as they allow us to comply with the constraints on the datasets access. Datasets will all have a metadata description, and, in the case of datasets with access restrictions, their metadata will be publicly available. A semantic versioning scheme will be used to track versions of the datasets. The partner responsible for generating that data will be the point of contact for requesting an access to the data.

Making data openly accessible For the traffic modelling use case, there will be no restrictions on the use of dataset for research activity. The only restriction might be the purchased data (e.g., origin-destination matrix and historical weather data), which will follow the license term of the provider. No control on the use of the dataset will be done, except the downloading step where applicants must declare for which research activity they want to use the dataset, and accept to quote EVEREST in acknowledgements in case of publication.

In order to make data interoperable and increase its reusage, the research data will be provided in an easy usable format ready for download and immediate use, typically vector data in CSV format.

7.3 Plan of the outputs

There are five research outputs produced by the use case as described in the following table. A first version of the Dataset D5 will be published before the mid-term review meeting.

Table 3 – Research dataset for TRAFFIC MODELING use case.

ID	ITEM	DESCRIPTION
D5	Dataset name and reference	FCD data sample (zenodo reference will be updated after publication)
	Dataset description	Floating car data (speeds on GPS locations) for a small geographical region (part of a city) for a given time interval (e.g., one day or one week)
	Standards, format and metadata	CSV
	Is dataset confidential? Must be encrypted?	Not confidential
	Data sharing/access inside EVEREST	Yes
	Data sharing/access outside EVEREST for research	Yes (Free Access)
	Is dataset reusable?	Yes (research activity only)

	Archiving and preservation (including storage and backup)	Data sample will be archived on EVEREST premises during project duration and for the future use after the project end possibly moved to **SYG** premises.
D6	Dataset name and reference	Road speed profiles
	Dataset description	Annotated speed labels and other metadata on a sample grid road network (part of a city on a higher-class roads)
	Standards, format and metadata	CSV
	Is dataset confidential? Must be encrypted?	Not confidential
	Data sharing/access inside EVEREST	Yes
	Data sharing/access outside EVEREST for research	Yes (Free Access)
	Is dataset reusable?	Yes (research activity only)
	Archiving and preservation (including storage and backup)	Data sample will be archived on EVEREST premises during project duration and for the future use after the project end possibly moved to **SYG** premises
D7	Dataset name and reference	Traffic prediction training sequence (zenodo reference will be updated after publication)
	Dataset description	Data sample for ML learning of prediction algorithm
	Standards, format and metadata	CSV
	Is dataset confidential? Must be encrypted?	Not confidential
	Data sharing/access inside EVEREST	Yes
	Data sharing/access outside EVEREST for research	Yes (Free Access)
	Is dataset reusable?	Yes (research activity only)
	Archiving and preservation (including storage and backup)	Data sample will be archived on EVEREST premises during project duration and for the future use after the project end possibly moved to **SYG** premises.
D8	Dataset name and reference	Simulated probabilistic speed profiles
	Dataset description	Simulated probabilistic speed profiles for the selected city/area and period.
	Standards, format and metadata	CSV
	Is dataset confidential? Must be encrypted?	Not confidential
	Data sharing/access inside EVEREST	Yes
	Data sharing/access outside EVEREST for research	Yes (Free Access)
	Is dataset reusable?	Yes
	Archiving and preservation (including storage and backup)	Data sample will be archived on EVEREST premises during project duration and for the future use after the project end possibly moved to **SYG** premises
D9	Dataset name and reference	Benchmark dataset for the simulation of routing in Smart City (zenodo reference will be updated after publication)

	Dataset description	The configuration of the traffic simulator and the setting of input parameters for the simulation of routing in a selected city (selected use cases).
	Standards, format and metadata	CSV
	Is dataset confidential? Must be encrypted?	Not confidential
	Data sharing/access inside EVEREST	Yes
	Data sharing/access outside EVEREST for research	Yes
	Is dataset reusable?	Yes (research activity only)
	Archiving and preservation (including storage and backup)	Data sample will be archived on EVEREST premises during project duration and for the future use after the project end possibly moved to **SYG** premises.