http://www.everest-h2020.eu

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

EVEREST

D1.4 – Final 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	42 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.	D1.4
Deliverable Title	Final Data Management Plan
Lead Beneficiary	NUM
Type of Deliverable	ORDP: Open Data Research Plan
Dissemination Level	Public
Due date	31/01/2024

Document Information

Delivery date	23/03/2024
No. pages	34
Version | Status	0.3
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), Raphael Polig (IBM)
Internal Reviewer	Radim Cmar (SYG)

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
08/01/2024	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
15/03/2024	0.2	Fabien Brocheton (NUM), Antonella Galizia (CIMA), Katerina Slaninova (IT4I), Radim Cmar (SYG), Guido Rusca (DUF), Riccardo Cevasco (DUF), Raphael Polig (IBM)	updates
19/03/2024	0.3	Radim Cmar (SYG)	Internal Review

Quality Control

Approved by internal reviewer	19/03/2024
Approved by WP leader	25/03/2024
Approved by Scientific Coordinator	23/03/2024
Approved by Project Coordinator	25/03/2024

Table of Contents

1. Executive summary

This document is an update of Deliverable D1.3. It describes the final status of the Data management plan (DMP) at M40, especially regarding research data available outside the project.

Indeed, the 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 the document

The document is organized as follows:

- Section 2 presents EVEREST data management policies.
- Section 3 presents methodology for access of data.
- Sections 4 to 6 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

D2.4 and D6.3 - details on the EVEREST use cases with respect to data use.

[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

D2.3 - More details on the data managed inside the project.
D3.2 - More details on EVEREST data management techniques inside the project.

2. Introduction

Figure 1 presents the EVEREST Data Lifetime (EDL). EDL is a high-level flowchart that depicts how data are managed. Since "data management" may have multi-dimensional aspects, we divide EDL into 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).

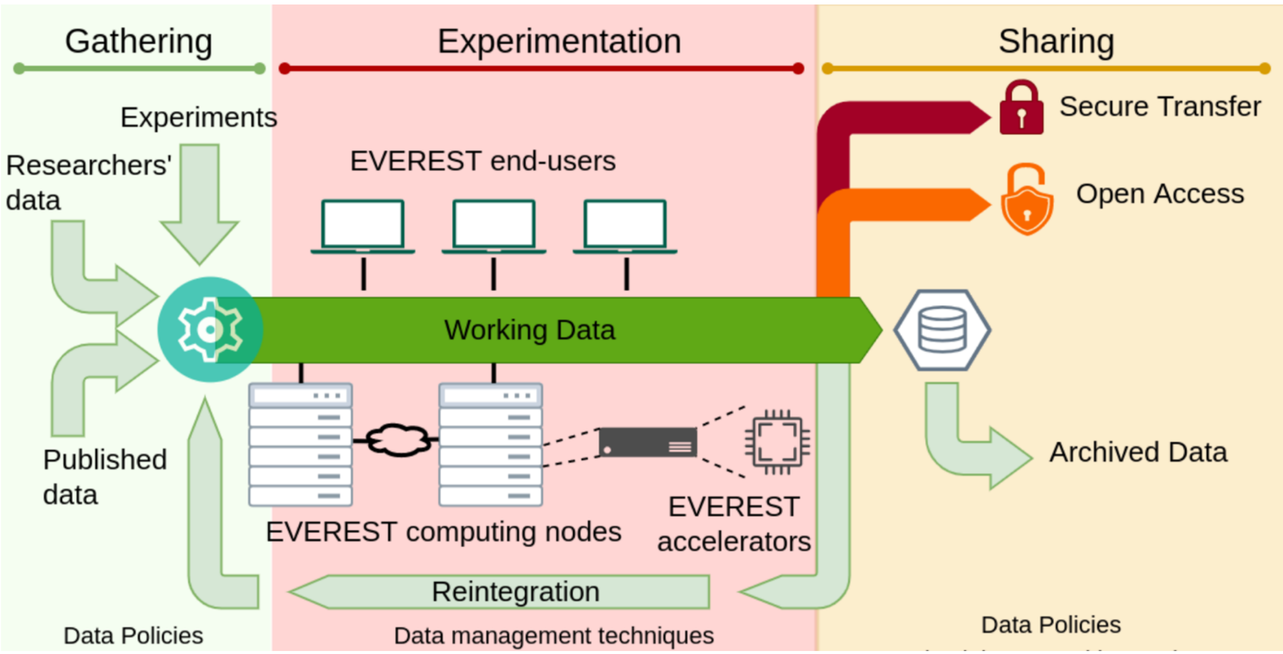

Figure 1: EVEREST Data Lifetime, gathering and sharing of data are mostly covered in this deliverable D1.4, while deliverable D3.2 addresses the data management techniques used for experimentation.

In D1.4, we mainly focus on "gathering" and "sharing" parts, whereas EVEREST data management techniques are described in D3.2.

Section 2 describes the data policies.

Section 3 describes how the open data for research will be published and accessible, as well as how data inside the project are shared and store.

Sections 4, 5 and 6 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 into 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 Zurich Yellow security zone Compute Cluster (ZYC2) of the IBM Zurich Research Laboratory. 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 users upon the "process of granting access permissions" (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 keeps 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 is collected, as informed by D.8.1. 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 is collected, processed and/or utilized exclusively for the purposes it has been provided for. No personal data is 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 do 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 are 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-created 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 select 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 provided 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 was provided according to the Allocation policy through the Open access Competitions, Directors Discretion, and EuroHPC JU Access Calls[2].

3.1.2.1. *Human roles and administration process*

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

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

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.

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.

User data in general can be accessed by:

1. IT4Innovations System Administrators.
2. The user, who created them (i.e. the LINUX owner).
3. Other users, to whom the user has granted permission.
4. Users, who at the same time have access to the particular Project Storage Area granted via the "Process of granting of access permissions".

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.

3.1.2.3. Data storage areas

There are four types of relevant storage areas: HOME, SCRATCH, PROJECT Data Storage, and PROJ4 Object Storage. HOME, SCRATCH, 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. 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. PROJ4 Object Storage is a highly scalable, distributed object storage system designed to store and retrieve large amounts of unstructured data.

HOME storage

The HOME filesystem is an HA cluster of two active-passive NFS servers. This filesystem contains users' home directories /home/username. This filesystem contains users' home directories /home/username. By default, 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. The filesystem is backed up. However, this backup is not intended to restore old versions of user data or to restore deleted files.

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 compute 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 the /scratch/project/PROJECT_ID directory created automatically with the PROJECT_ID project. 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, 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.

PROJECT Data Storage

The PROJECT data storage is a central storage for projects' and users' data at IT4I that is accessible from all 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 the project paradigm. Storage allocation and access to the storage are based on projects (i.e., computing resources allocations) and project membership.
A project directory (actually 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 directory is removed after the project's data expiration.
The PROJECT storage utilizes a fully redundant design, redundant devices, highly available services, data redundancy, and snapshots. For increased data protection, disks in each disk array are connected in Distributed RAID6 with two

hot-spare disks, meaning the disk array can recover full redundancy after two simultaneous disk failures.
However, the storage does not provide data backup, so it is strongly recommend using other storage facility for making independent copies of data.

PROJ4 Object Storage
OpenStack Swift is a highly scalable, distributed object storage system designed to store and retrieve large amounts of unstructured data. It is an open-source project that provides a simple, scalable, and durable storage system for applications and services. Swift is built to be highly available, fault-tolerant, and scalable, making it an ideal choice for storing large amounts of data.
Swift is designed to be highly modular, with a simple API that allows developers to easily integrate it into their applications. It provides a RESTful API that can be accessed using a variety of programming languages, making it easy to integrate with existing applications.
One of the key features of Swift is its ability to scale horizontally, allowing it to handle large amounts of data and high levels of traffic. It is also designed to be highly durable, with data being replicated across multiple nodes to ensure that it is always available.
Overall, OpenStack Swift is a powerful and flexible object storage system that is well-suited for a wide range of applications and use cases.

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 and 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 a 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 file Access Control List mechanism ACL (for a more fine-grained control of permissions to specific users and groups). Default file permissions and ACLs are set by IT4I during project directory provisioning.

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. PROJECT storage will be securely deleted after 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 also a special Data Analytics Compute Node oriented on supporting huge memory jobs by implementing a NUMA SMP system with large cache coherent memory. This Superdome Flex server is 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. Credentials to access the servers are given to partners upon request and verification.

4. Access to datasets

4.1. Access to open datasets

Concerning the open datasets associated with each use case, we decided to use the Zenodo platform (https://zenodo.org) to provide access to them. Publications on Zenodo are linked for each dataset to the OpenAIRE platform (https://www.openaire.eu/).

The EVEREST community has been created on the Zenodo platform (https://zenodo.org/communities/everest/). During the project, we identified datasets in each use case (see sections 5, 6, and 7), which are published on Zenodo.

The open data provided by the EVEREST project meet the requirements for the FAIR data policy. To make them accessible, metadata was defined for each dataset as much as possible.
We followed international standards, especially DataCite[3], which defines a minimum set of information to provide (description, producer, etc.)

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),
- Publication Year (year when data became/become/will become public),
- Resource Type.

Each of these fields contains very few subfields for further description. Additional metadata could 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 and 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 datasets are limited to the dataset used for AI process.

[3] 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

Concerning source code, we focused in the first part of the project to share internally the different applications codes and EVEREST SDK (see section 4.2).

4.2. Internal access to dataset and storage

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

The storage is realised as iRODS zone accessible through the LEXIS Distributed Data Infrastructure DDI[4], which is inspired partially by the EUDAT B2SAFE. The DDI implements asynchronous data transfers between locations and metadata indexing features. These features are used by the workflows executed by the LEXIS Platform which automatically stage the results of a computation in the DDI and create a new dataset. These datasets are by default accessible only internally by the project partners, but can be also set as public with PID assignment. The DDI partially enforces DataCite standard and collects information about workflow and execution, which contributes to the data provenance and to the overall alignment with the FAIR principles.

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 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 (https://www.mmm.ucar.edu/weather-research-and-forecasting-model). During each simulation, an assimilation procedure in order to force

[4] https://docs.lexis.tech/_pages/data_system/data_system.html

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 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 (http://cerc.co.uk/environmental-software/ADMS-model.html), 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). In the final demonstration of EVEREST, it is IFS which is used on EVEREST infrastructure.
- 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.

The outputs of 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); the GFS dataset used (or not) at step 1, eventually a forecast from Meteo France.
- For the learning phase, local weather surface observation at industrial site.

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

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

Most of the datasets described above are open-source data available freely, except:
- **NUM** forecast data, which are commercial ones.
- Local weather observation data at the industrial sites are private data.
- IFS forecast, which can be available freely only in the framework of a research project.

- 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 internally with some restrictions (in terms of diffusion, etc.).

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

We identify one dataset which can be shared for open research:

- The numerical weather forecasts (including WRF model output from step 1 described above) and the weather observations used as input in step 2 of the AI approach.

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 must be provided with an open scientific format largely used in the meteorology applications, and not in a proprietary model format. We choose csv format which is a format easily readable and manage by many tools and software libraries.

5.3. Outputs

Among the possible public outputs, we decide to select data in relation with the main outputs of the air-quality user case, outputs which can be used (i) to develop similar research as those done in EVEREST and (ii) to compare / reproduce the results obtained during EVEREST project. case.

This public output concerns the weather data used to construct and validate the local ensemble aggregation approach. The provided dataset corresponds to a set of different numerical weather forecasts: **NUM** forecast and GFS forecast. Associated to these forecasts, local weather observations are also provided.

When this report is written, the EVEREST WRF forecast is not already inside the dataset. The reason is to provide a dataset on a large period which can be exploitable for AI research. At the end of the project, when EVEREST WRF outputs cover a larger period, the dataset will be replaced.
The dataset 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 following table provides additional information on this dataset.

Table 1 – Research dataset from AIR QUALITY use case.

ID	ITEM	DESCRIPTION
D1	Dataset name and reference	Local weather forecast and weather observation (DOI: 10.5281/zenodo.6528866)
	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
	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.

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.
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**: it concerns with pre-processing methods based on a deterministic calculation of hourly energy generation, in the site-specific historical meteorological conditions forecasted by the WRF model, using the power curve of the Wind Turbine.
- In **Step 3**, historical data pre-processed by **Step 2**, are used for training machine learning model, combined with last hourly generation data.

Step 1 requires different input data the main ones are:

- Global forecast dataset (GFS) produced by NCEP (US National Center for Environmental Prediction).
- Some surface observation data (sea surface temperature, humidity of soils, etc.).
- For the assimilation procedure:
  - Additional Surface weather observation data (i.e., temperature) 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) were retrieved from ground-based (radar).

The outputs of 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 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 described above are open-source data available freely, except:

- **DUF** site-specific historical and observation data (e-g. hourly basis generation, wind speed...) is confidential and to not be shared 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 outputs. 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 must be provided with an open scientific format largely used in the meteorology applications, and not in a proprietary model format. We choose csv format which is a format easily readable and manage by many tools and software libraries.

6.3. 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.

Table 2 – Research dataset from ENERGY PRODUCTION use case

ID	**ITEM**	**DESCRIPTION**
D2	Dataset name and reference	Local weather forecast (DOI:10.5281/zenodo.10592939)
	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	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	During the project: Outputs stored for the duration of the EVEREST project, with a focus on

	(including storage and backup)	parameters used for air quality application. After the project: Outputs will be stored on the **DUF** storage system.
D3	Dataset name and reference	Energy prediction output (DOI:10.5281/zenodo.10592071)
	Dataset description	Energy production prediction (wind farms set in Italy) at hourly temporal resolution, for a selected period and wind farm.
	Standards, format and metadata	CSV
	Is dataset confidential? Must be encrypted?	Spatial coordinates are "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 was operated on selected cities only (like Vienna, Prague, or Bratislava) 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 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 factor (optional).

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 used for traffic prediction incorporated into the online routing **SYG** platform/simulator. A set of experiments based on different scenarios were performed, e.g., online routing on the level of cities (smart-city routing) and routing on the level of country. The collected data were 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 were available to EVEREST project, and in a limited scope are publicly available as open data.
- 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.
- Road speed profiles calculated by map matching of FCD data are publicly available in a limited scope.
- Probabilistic road speed profiles calculated by traffic simulator for a given time period can be shared publicly to the research projects.
- Training sequences to infer road traffic prediction models are publicly available in a limited scope

7.2. FAIR data

Making data findable, including provisions for metadata

The data are stored on data repositories with digital object identifiers. We selected Zenodo as a public repository which allows us to comply with the constraints on the datasets access. All datasets have a metadata description, and, in the case of datasets with access restrictions, the metadata are publicly available. A semantic versioning scheme was used to track versions of the datasets. The partner responsible for generating that data is 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 followed 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 are provided in an easy usable format ready for download and immediate use, typically vector data in CSV format.

7.3. Outputs

There are five research outputs produced by the traffic use cases open to public.

- FCD data sample (D5) can help research community to develop their own algorithm to infer speeds on the road network and challenge the road speed profiles calculated in EVEREST (D6)
- Training sequences as road speed time-series (D7) can be used by research community to experiment with training models for traffic prediction on major roads for 1 hour ahead prediction
- Simulated probabilistic time-dependent routing (PTDR) profiles (D8) with road speeds over different times of a day for major roads in Prague within the span of one week.
- Configuration of open-source Traffic simulator and input data to run a simulation for optimising traffic flow within a city (D9) with the goal to get research community up to speed using the simulator.

Table 3 - Research dataset from TRAFFIC MODELING use case

ID	ITEM	DESCRIPTION
D5	Dataset name and reference	FCD data sample (DOI: 10.5281/zenodo.6373586)
	Dataset description	Floating car data (noisy speeds on noisy GPS locations) for a part of Prague city for a single day
	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 archived on ZENODO repository.

D6	Dataset name and reference	Road speed profiles (DOI: 10.5281/zenodo.10663408)
	Dataset description	Calculated speed information across a full day on Prague city road network for major 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 archived on ZENODO repository.
D7	Dataset name and reference	Traffic prediction training sequence (DOI: 10.5281/zenodo.10818506)
	Dataset description	Data samples for ML prediction algorithm for hundreds of roads in Prague city provided with 2 hours' time-series speed sequences for a single day
	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	Data sample archived on ZENODO repository.

	(including storage and backup)	
D8	Dataset name and reference	Simulated probabilistic speed profiles (DOI: 10.5281/zenodo.10701789)
	Dataset description	Simulated probabilistic speed profiles for Prague, 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
	Archiving and preservation (including storage and backup)	Data sample archived on ZENODO repository.
D9	Dataset name and reference	Benchmark dataset for the simulation of routing in Smart City (DOI: 10.5281/zenodo.6985345)
	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 archived on ZENODO repository.

References

LEXIS project https://lexis-project.eu/web/