D1.3 – Data Management Plan (First Release)

Deliverable ID	D1.3
Deliverable Title	Data Management Plan (First Release)
Work Package	WP1 – Project Management
Dissemination Level	PUBLIC
Version	0.6
Date	2021 – 08 – 31
Status	Submitted
Deliverable Leader	LINKS
Main Contributors	IT4I, CINECA, ATOS, AVIO, ECMWF, SINTEF

The ACROSS project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955648. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, France, Czech Republic, United Kingdom, Greece, Netherlands, Germany, Norway.

Published by the ACROSS Consortium

Document History

Version	Date	Author(s)	Description
0.1	2021-04-21	LINKS Foundation	First version ToC
0.2	2021-07-26	LINKS Foundation	In review
0.3	2021-07-30	MORFO, IT4I	Review completed by MORFO
0.4	2021-08-06	IT4I	Review completed by IT4I
0.5	2021-08-30	LINKS	Pre Submission Version
0.6	2021-08-31	LINKS	Submission Version

Table of Contents

List of figures

List of tables

Glossary

Acronym	Explanation
DMP	Data Management Plan
GDPR	General Data Protection Regulation
DCV	Data Catalogue Vocabulary
ORDP	Open Research Data Pilot
OA	Open Access
FAIR	Findable, Accessible, Interoperable, Reusable
NL- SAS	Near Line Serial Attached SCSI (Small Computer System Interface)
SGI DMF	SGI Data Migration Facility
SMP	Shared Memory Parallelism
HA	High Availability
NFS	Network File System
WMO	World Meteorological Organization
CFD	Computational Fluid Dynamics
AI	Artificial Intelligence
HPDA	High Performance Data Analytics
LSE	Large Eddy Simulation
URANS	Unsteady Reynolds Average Navier-Stokes
CISO	Chief Information Security Officer
ACL	Access control list
AdS	System Administrator

Executive Summary

The Deliverable D1.3 “Data Management Plan (First Release)” has the objective to establish the initial strategy for the Data Management Plan considering different conditions regarding the overall project and defining the policies about data which will be part of the ACROSS platform and coming out of the 3 pilots (Aeronautics, Weather and Climate, and Energy and Carbon Sequestration).

Position of the deliverable in the whole project context

This deliverable is the first release of the Data Management Plan focusing on the Data Management strategy of the ACROSS. Its main goal is to set up the guidelines and strategy for handling and clarifying the Data Managements approach during the project. The final release of the Data Management Plan will be include in the deliverable D1.4 “Data Management Plan (Final Release)” which will be delivered in M36 and will report all the results reached at the end of the project in term Data generated and gathered from pilots and the ACROSS platform.
The D1.3 “Data Management Plan (First Release)” is related to Task 1.3 “Quality assurance and Data Management Plan” of WP1 “Project Management”. In Figure 1 is represented the WP1 position in the ACROSS Project and its relation with the other WPs.

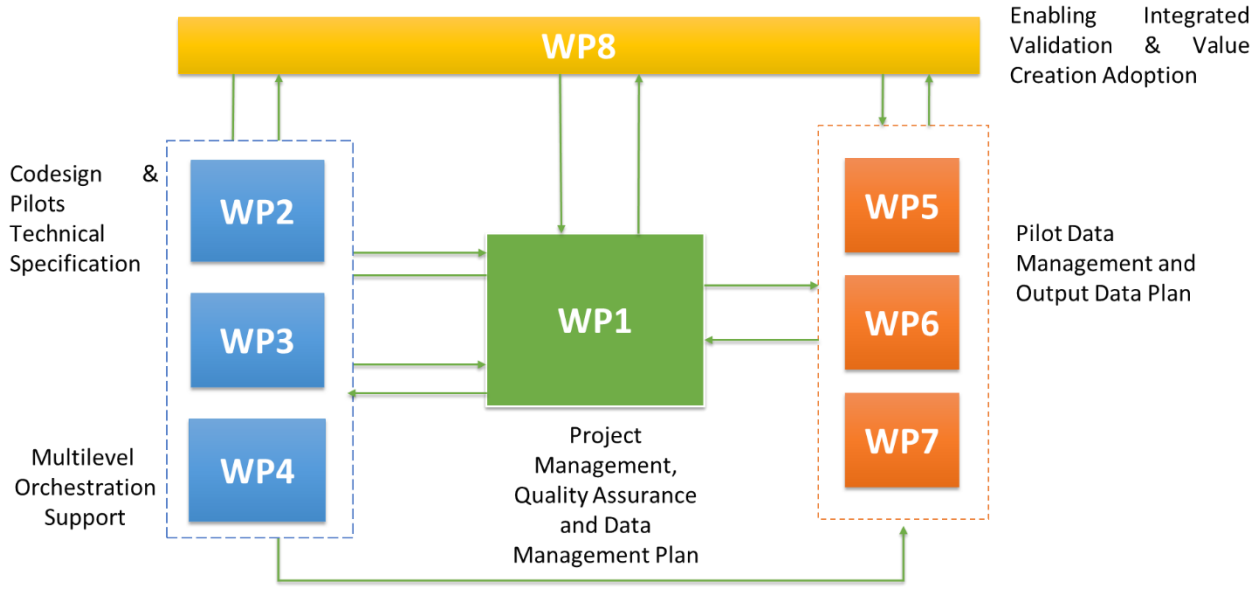

Figure 1 WP1 position in ACROSS project

Description of the deliverable

The Data Management Plan describes all the data management processes related to the ACROSS project. After the Introduction in Section 1, the document will define some general principles about data management policy and scientific publications in research context. Then Section 2, the Data Management Plan will describe the type of data will be generated or gathered during the project, the standard that will be used, the ways how data will be exploited for verification of reuse and preservation with particular focus to be GDPR compliant. This section includes data management policies of the supercomputing centres as the main contributors of the ACROSS overall infrastructure and the data management plans of the three pilots.

1 Introduction

This document is the first release of the Data Management Plan (DMP), presenting an overview of data management processes, as agreed among ACROSS project's partners. This DMP will first establish some general principles in terms of data management and Open Access.
Subsequently, it will include Data Management Policies of the Supercomputing Centres.
Then for each Pilot it will be structured as proposed by the European Commission in H2020 Programme – Guidelines on FAIR Data Management in Horizon 2020 [1] , covering the following aspects:

- Data Summary;
- FAIR Data;
- Allocation of resources;
- Data security;

1.1 Scope

This DMP describes how research data are managed both throughout the life cycle of ACROSS and after the end of the project. It identifies procedures and minimum requirements to collect, store, analyzed, and publish data in a consistent way according to the FAIR principles.
The DMP is a "living" document outlining how the research data collected or generated will be handled during ACROSS project. The DMP is updated over the course of the project whenever significant changes arise. At M36 will be released the final version of the DMP.

1.2 Related documents

ID	Title
[RD.1]	Grant Agreement No. 955648
[RD.2]	Consortium Agreement

2 Initial Data Management Plan

The Data Management Plan of ACROSS will include the data management process, which will identify the data generated during the project and the data exploited or made accessible for reuse. A well-defined data management process is an important task for the project. A key element of data management is a well-defined process for the handling of research data. For transparency reasons, this process needs to be clearly defined and accessible for all potential stakeholders of the data. The DMP describes how a research project processes research data and it provides answers to all important questions about the data processing, including data security, licensing, origin of data, format. Since these answers may change during the run time of a project, the DMP will be regularly updated and revised.
During the initial phase of the project have been identified the pilot related datasets that will be analysed during the project lifetime.

2.1 ACROSS Data Management Overview

ACROSS project is part of the Open Research Data Pilot (ORDP) and the data management process will follow a strategy to apply the Open Access in compliance with the Horizon 2020 guideline as represented in the Figure 2.The policy reflects the ACROSS CA [RD.2] regarding the data management and is in line with the exploitation and protection of results.

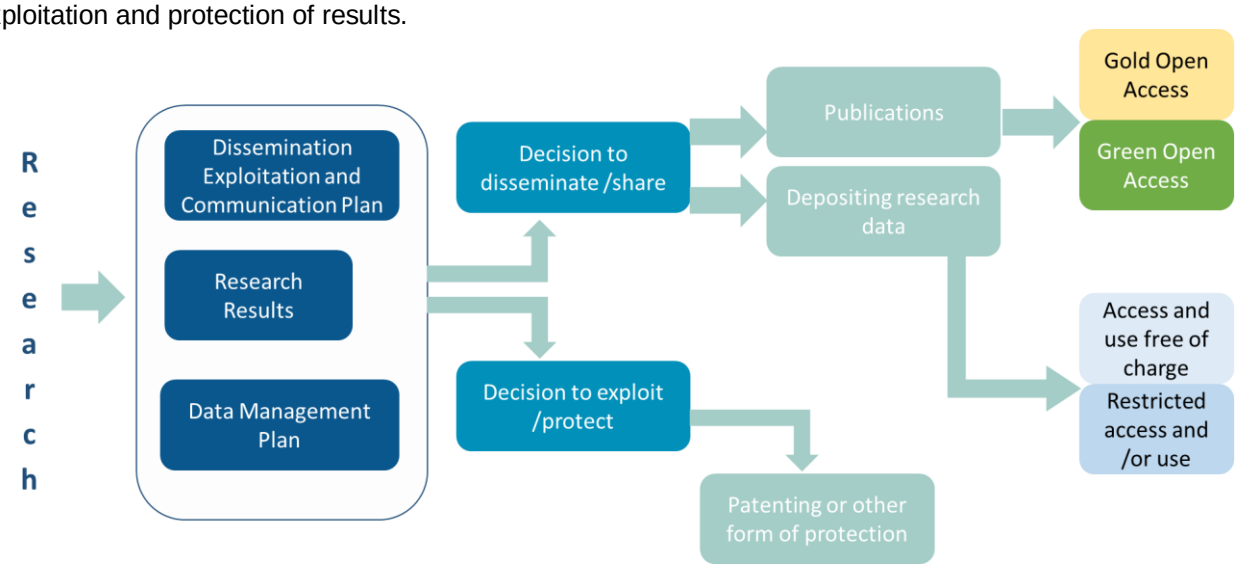

Figure 2 Open access to research data and publication decision diagram [2]

This is the first release of the DMP and will evolve reporting all significant changes made necessary during the whole duration of the ACROSS project and also all the changes required during the periodic reviews and during the reporting stages of the project.

The DMP will include the type of data, the storage, and confidentiality issues. It will include also the procedures that will be implemented for data collection, storage, access, sharing policies, protection, retention and destruction, which will be in line with EU standards as described in the Grant Agreement [RD.1] and the Consortium Agreement [RD.2].

Due to the pilots' diversity, different data standards will be used to maximize interoperability between different datasets.

A detailed description of the datasets generated will be provided, it will be based on novel metadata standards for describing datasets, such as Data Catalogue Vocabulary [3].

2.1.1 Data collection and GDPR Compliance

Within the ACROSS project, partners collect and process research data and data for general project management purposes, according to their respective internal data management procedures and in compliance with applicable regulations. Data collected for general purposes may include contact details of the partners,

their employees, consultants and subcontractors and contact details of third parties (both persons and organizations) for coordination, evaluation, communication, dissemination and exploitation activities. Research data are collected and processed in relation with the research pilots. During the project lifetime, data are kept on partners dedicated infrastructures. Data archiving, preservation, storage and access, is undertaken in accordance with the corresponding ethical standards and procedures of the partner institution where the data is captured, processed or stored.
The consortium complies with the requirements of Regulation (EU) 2016/679 and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

2.1.2 Research Data and Open Access

ACROSS is part of the H2020 Open Research Data Pilot and publication of the scientific results is chosen as a mean of dissemination. In this framework, open access is granted to publications and research data and this process is carried out in line with the Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 (see Figure 2).

2.1.2.1 *Scientific publications*

Open access is applicable to different types of scientific publication related to the research results, including its bibliographic metadata, such as:

- journal articles;
- books;
- conference proceedings, abstract and presentations;
- grey literature (informally published written material).

Grey literature also includes reports and deliverables of the projects related to the research, whose Dissemination level is marked as Public. Open access is granted as follows:

- Step 1 – Depositing a machine-readable electronic copy of a version accepted for publication in repositories for scientific publications (before or upon publication).
- Step 2 – Providing open access to the publication via the chosen repository.

For access to publications, a hybrid approach is considered (both green OA and gold OA), depending on the item and the dissemination channels that will be available:

- Green OA (self-archiving) – depositing the published article or the final peer-reviewed manuscript in repository of choice and ensure open access within at most 6 months (12 months for publications in the social sciences and humanities).
- Gold OA (open access publishing) – publishing directly in open access mode/journal.

2.1.2.2 *Data Management Policy*

ACROSS Data Management Policy is focused on the observation of FAIR (Findable, Accessible, Interoperable and Reusable) Data Management Protocols, in compliance with Open Research Data Pilot in Horizon 2020.

2.1.2.3 *Research data*

Open Access is granted also to underlying research data (data needed to validate results presented in publication) and their associated metadata, any other data (not directly attributable to the publication and raw data) and information on the tools needed to validate the data and, if possible, access to these tools (code, software, protocols etc.). Open access is granted as following.

1. Depositing the research data in a research data repository. A repository is an online database service, an archive that manages the long-term storage and preservation of digital resources and provides a catalogue for discovery and access.

2. Enabling access and usage free of charge for any user (as far as possible). The consortium will try to publish as much research data as possible, but this will be decided on a case-by-case basis, in order to be compliant with the GDPR in terms of publishing non-sensitive and personal data.

2.1.2.4 *Other project's outcomes*

As per any other outcomes of the project, they are disseminated accordingly to the Dissemination level indicated in the Description of Action and they are also subject to protection in accordance with the Consortium Agreement and in reference to Access Rights.

Project results and formal deliverables will undergo a quality control process for internal project verification and document reviews. Prior to approval by the Project Management Board and the submission to the European Commission, the deliverables will be revised at least by two different reviewers.

2.2 Data Management Policies of the Supercomputing Centres

2.2.1 IT4I

2.2.1.1 *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 IT4Iprovides access to the granted computing resources (accounted in core-hours consumed). The 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 ACROSS 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. 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" process.

2.2.1.2 *Process of granting of access permissions*

All communications 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.

2.2.1.3 *Data storage areas*

There are five types of relevant storage areas: HOME, SCRATCH, BACKUP, PRIVATE and PROJECT Data Storage. HOME, SCRATCH, BACKUP 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 and is archived on the tape library - BACKUP.

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.

2.2.1.3.1 *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 25 TB, shared among all users. Individual users are restricted by filesystem usage quotas, set to 25 GB per user. Should 25 GB prove insufficient, contacting support, the quota may be lifted upon request.

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

The filesystem is backed up, so that it can be restored in case of a catastrophic failure resulting in significant data loss. However, this backup is not intended to restore old versions of user data or to restore (accidentally) deleted files.

2.2.1.3.2 *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. Individual users are restricted by filesystem usage quotas, set to 9.3 TB per user. The purpose of this quota is to prevent runaway programs from filling the entire filesystem and deny service to other users. Should 9.3 TB prove insufficient, contact support, the quota may be lifted 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.

2.2.1.3.3 *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.

2.2.1.3.4 *BACKUP storage*

Contents of HOME storage are automatically backed up to tape library. There is a minimal period of retention, but no maximal, so we cannot guarantee time when the backups are removed from the tapes.

2.2.1.3.5 *PRIVATE BACKUP storage*

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

2.2.1.3.6 *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.

2.2.1.4 *Data access*

2.2.1.4.1 *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.

2.2.1.4.2 *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.

2.2.1.5 *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. However, the HOME system has a tape backup, and the copies are kept for indefinite time. We advise not to use HOME storage if you do not wish to keep copies of your data on tapes. PRIVATE storage will be securely deleted upon request or when the project ends.

2.2.1.6 *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[1] and Barbora[2] 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.

2.2.1.7 *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. In December 2018, IT4I was awarded ISO 27001 certification (ISO/IEC 27001:2013, ČSN ISO/IEC 27001:2014). The certificate was obtained for provision of national supercomputing infrastructure services, solution of computationally intensive problems, performance of advanced data analysis and simulations, and processing of large data sets.

2.2.2 CINECA

2.2.2.1 *HPC organisation*

The Cineca organization includes four teams for the management of the HPC facility:

- **Facility and building management**: the team manages, through an integrated approach, the buildings and services needed to support the operation of the supercomputers and to meet CINECA's business needs.
- **HPC systems management**: The team deals with the implementation, maintenance and performance optimization of HPC infrastructures. It consists of a team of 10 qualified people working in collaboration with the CINECA network team and the security team.
- **HPC user support**: the team is in charge of managing the production of HPC systems, supporting users through help desk, online documentation and installation of scientific and technical applications. In addition, it manages HPC production side services and deals with national peer-reviewed calls for access to the computing infrastructure by academic research groups.
- **CERT**: the team is in charge of IT security for both IT and HPC infrastructure, in collaboration with the systems management and user support teams.

2.2.2.2 *Users management*

2.2.2.2.1 *Access to the machine room*

In order to guarantee the physical security of the IT equipment and data in the Cineca equipment room, access is regulated and authorised only for those in charge of the work. Persons authorised by the Facility group must remain in the machine room only and exclusively to carry out tasks that cannot be carried out remotely. Furthermore, access by external personnel is subject to the presence of authorised internal personnel. All

[1] https://www.it4i.cz/en/infrastructure/karolina

[2] https://www.it4i.cz/en/infrastructure/barbora

access routes, windows and ventilation domes to the premises must remain closed at all times (except in the event of an emergency).

2.2.2.2.2 *Roles and responsibilities*

HPC system administrator

The management of supercomputers in Cineca is the responsibility of the "HPC system management" group. The Cineca system administrator (AdS) is an individual role, formalised through an act of appointment that contains an analytical list of the permitted areas of operation according to the authorisation profile assigned to the person in charge. It is the responsibility of the persons in charge of systems management to periodically check the access rights of individual AdS. The results of the audits shall be reported in the Cineca management system and it is also their task to periodically check that the actions of the AdSs correspond to the organisational, technical and security measures provided for by the regulations in force in case of personal data processing. The revocation of the AdS' assignment may occur following a transfer to a different job or due to termination of the employment relationship, and it is the direct manager's task to update the list of administrators on the Cineca management system, inserting the date of termination of the account.

User Support HPC

The User Support HPC user, both scientific and industrial, has higher permissions than a normal HPC user. Due to the nature of his role, he needs to be able to inspect the folders and files of users in read-only mode, but is not allowed access to system folders and configurations. The list of User Support users is maintained and updated by the HPC Production Manager. Cineca employment contracts include a confidentiality clause. Employees working with industrial customers sign a confidentiality agreement that is more restrictive and specific than what is normally provided for in the contract.

HPC user

A user requesting access to Cineca's supercomputing systems accepts a service contract and receives a personal (and non-transferable) account of which he/she is the owner and fully responsible. The HPC user is authorised to access the computing systems after completing the registration and access request procedure via our user database: userdb.hpc.cineca.it.

The HPC user support group verifies the identity of the individual user by means of an identification document. Once the identity and the association to a project have been verified, the user will receive in two separate emails a username and a temporary password for access to the machines.

For scientific users, the account naming policy is unencrypted, the username will be traceable to the researcher's name and surname, while for companies that require it, there is the possibility of having pseudo-anonymised accounts.

2.2.2.3 *HPC Cyber Security Policies*

The Chief Information Security Officer (CISO) reports to the General Management and defines the security and risk management policies for all Cineca services: management for universities, services for the Ministry (including competition management), HPC.

For more than 20 years, Cineca has had a Computer Emergency Response Team, currently reporting to the CISO, which operates in a preventive manner by coordinating a team of specialists from the various areas of systems management and is able to act promptly in the event of computer incidents.

Customisation of access and configuration of parts of the HPC clusters to meet specific customer requirements is possible.

The system management team's procedures for monitoring HPC systems follow the international standard UNI EN ISO 9001:2015 and for information security management the ISO 27001:2013 standard, which are integrated into internal operating procedures.

2.2.2.4 *DATA MANAGEMENT*

2.2.2.4.1 *Connection to HPC clusters*

Access to the computing infrastructure takes place by connecting via ssh protocol, to the cluster login nodes via the address:

ssh login.nome cluster.cineca.it.

The ssh protocol is a protocol that allows the exchange of information between two computers via an encrypted channel and also allows commands to be executed in remote console mode.

Access to the supercomputing machines can be through username/password authentication or through the public/private key system to avoid entering the password at each login.

2.2.2.4.2 *Cluster data handling IN/OUT*

Various protocols can be used to move data both in upload and download, all based on encrypted channels:

- SCP: this application uses an encrypted ssh connection to allow data to be copied between 2 servers, very flexible for moving small files, but not optimised for handling large amounts of data.
- RSYNC: allows local directories to be synchronised with remote ones, the exchange protocol can be an encrypted ssh channel.
- SFTP: This protocol implements ftp data transfer over an encrypted ssh channel and therefore allows secure get/put of files to remote directories.
- gridFTP: allows you to move large amounts of data efficiently between 2 gridFTP servers or between a local machine and a gridFTP server, in which case you must use the Globus online web interface and have an x.509 certificate issued by a Certificate Authority.

For more information: wiki.u-gov.it/confluence/display/SCAIUS/Data+Transfer

2.2.2.4.3 *Temporary store and transmitting Output data, Data deletion*

Filesystem and directory permissions

On all Cineca HPC systems there is the same logical storage structure, each area can have the following attributes:

- Temporary/permanent: the first has the automatic deletion of the data after a certain period, while the second is permanent and will be deleted after 6 months from the end of the project.
- User/Project: the first is an area accessible only by the single user, while the second is shared among all the members of the project
- Local/shared: the first is a local area accessible only by a single Cineca cluster, while the second is a disk area accessible by all HPC machines.

All access permissions are set by default by the system administrator through a linux ACL and only the single user who has access to his storage area can change them.

The data areas in Cineca are identified by the following environment variables and are all parallel storage areas with GPFS file system:

- $HOME (permanent, user, local): area used to contain software and small data, it has a size of 50GB and is backed up daily.
- $WORK (permanent, project, local): common area between the members of a single project and by default has a size of 1TB. It can be used to hold large data for project development and is optimised for IO of large volumes of files.
- $CINECA_SCRATCH (temporary, user, local): area used for temporary storage of files during execution of batch processes, it does not perform differently from $WORK, but is user specific. Its size is 20TB, but it is periodically cleaned to remove files not accessed in the last 40 days.

Temporary files are by default placed in the folder corresponding to the environment variable $TEMPDIR, which if redirected to a subdirectory of the areas previously mentioned makes the files accessible or not depending on the ACLs set in that area.

For more information:

wiki.u-gov.it/confluence/display/SCAIUS/UG2.5%3A+Data+storage+and+FileSystems

2.2.3 ATOS

Large datasets are used and generated by HPC thus making data management a key component of effectively using the expensive resources that underlie HPC infrastructure: indeed, the HPC community continues to generate massive amounts of file data, drawing insights, making that data useful, and protecting the data becomes a considerable effort with major implications. Having a better way to manage storage and data has multiple benefits, such as drawing insights from disparate data sources, increased researcher productivity, reduced costs, and reducing staff maintenance and administration overhead.

At a high level, a data management plan should include:

- Core architecture able to scale to petabytes/exabytes (amount of data + file count) without loss in performance;
- Visibility of files across all systems and accessibility through a (vendor agnostic) homogeneous interface;
- Intelligent archiving and recovery functionalities to free up space on primary tiers;
- Data protection with efficient policies for backup, restore and retention.

Currently, many of the above requirements are not completely solved, and other tasks (mostly related to computing and performance) take priority over understanding, using, and protecting data.

In spite of being fully aware of this critical element of HPC, ATOS does not intend to put a specific and complete data management plan in place. The main reason is that the ATOS HPC in ACROSS should be viewed as an experimental platform dedicated to the exploration of new computing technologies and paradigms and hence priority is to put more on efficiency and flexibility than on reliability and security. However, being a HPC cluster, our platform will deliver a minimal set of data management features to provide security to our partners:

1. **Access security**: To access the cluster, a user must have an authorization to connect to the network and to then to the cluster. A user group is created for each partner, and each user is associated with a user group. Direct SSH connections are known and monitored through the AGARIK firewall. Thus, users who want to connect to the ATOS platform must declare their IP address to ATOS Admin (by filling and submitting the "Request to get access " form to the Project Manager and the cluster Administrators)
2. **Data Access**: Jobs and resources are managed by Slurm that enables to request and reserve resources, to submit jobs, to query their status, and more generally to make an effective use of the cluster resources:
    - Due to the large amounts of data required by user applications, the Platform Administration Service does not provide any backup for data and/or software installed in user spaces and shared spaces.
    - Storing personal data allowing the identification of persons is forbidden: The platform is not designed with a sufficient level of security to host and protect this type of data.
3. **Resource Manager:** Batch management is based on the open source resource Slurm manager. Major enhancements in version 15.08 include:
    - Hierarchical implementation based on hardware topology using the interconnect network for all communications to improve security and availability;
    - Support of Kerberos authentication through *AUKS*module;
    - Power adaptive scheduling for applications to enhance power capping by managing unused nodes and reducing CPU frequency:
    - Energetic fairshare scheduling based on energy consumption accounting;

- Hyperthreading support to extend actual placement (socket and core) to hyperthread level.

4. **Parallel File system:** The parallel file system is based on the Intel®Enterprise Edition for Lustre (IEEL) core, providing high performance and large storage solutions for big data workloads. Its main features include :
   - Extra functionalities were added by Atos for Lustre 2.7:
   - Integration of Lustre client and router with MOFED® stack,
   - Shine centralized administration tool,
   - Monitoring with Shinken and Graphite,
   - High Availability integration based on pacemaker.

   It should be noted that in terms of data protection/isolation, no specific mechanism will be implemented to secure this feature. The only verification will be on data coherence via ECC/CRC checking.

The ATOS HPC platform developed within the framework of ACROSS will provide a minimum of data management features as described above, in compliance with the minimum requirements set out in the H2020 guidelines, but will prioritize on investigation efficiency and innovation.

2.3 Data Management Plan – Greener aero-engine modules optimization Pilot

The Greener aero-engine modules optimization pilot is constituted by two separate tasks focusing on key aeronautical components: the low-pressure turbine and combustor.

2.3.1 Data Summary

Turbine pilot:
Data will have a central importance in turbine-pilot as it is mainly focused on the development of a new Design System (DS) based on a huge database built during the project leveraging advanced CFD modelling, HPC resources and HPDA techniques. For these reasons, the data generated during the whole project will play a key role for reaching the objective of the pilot.

The articulated work-flow of the current pilot suggests a data management equally complex during the whole procedure, with several points in which data are generated and exchanged among the different tools adopted. The whole process starts with few scalars as input. From these values, a blade geometry is defined and a first set of data therefore is generated in this preliminary phase to provide CFD calculation with the geometry to be investigated.
Then URANS and LES fluid dynamics computations will be adopted to evaluate the performance of Low-pressure turbine blades. In this phase the main part of data will be generated that, in turn, will feed the input for HPDA and AI applications. Several data format will be used along the workflow coherently with the requirements of the different tools.

Data collected, as reported in the following, can be discriminated as confidential, hence not shareable, and others that will be public with the objective to promote understanding and comprehension about which type of knowledge can be extracted thanks to the advanced Computational Fluid Dynamics Analysis: a great insight to explore complex flow fields aimed at minimizing loss sources and get more efficient fluid processes.

Combustor Pilot:
The aim of the project is to optimise the U-THERM3D tool with which multi-physics and multi-scale simulations are carried out using a loosely coupled approach for predicting wall temperatures inside aeronautical combustion chambers. The tool developed within the commercial ANSYS Fluent solver allows the resolution of fluid dynamic and conductive problems.
The data produced by the simulation contains all the information related to the heat exchange that takes place inside the combustion chamber (e.g. velocity field, temperature field, species concentration produced by the chemical reaction, etc.). A great insight to explore complex flow fields in order to get more efficient fluid processes.
Results are provided in the standard ANSYS Fluent format (file.dat) reporting various quantities of interest, such as the temperature on a certain surface or the velocity on a certain plane, but also a one-dimensional profile selected at a certain station; in these cases, the exported files can be re-opened and processed with any text editor. The results obtained can be used to initialise further simulations of the same test case. For

example, the standard practice to carry out a simulation with U-THERM3D is to start from the CFD simulation with adiabatic walls and then couple it with the solid domain.
The size of the output data will be strictly dependent on the test case under analysis.

Data collected can be discriminated as confidential hence not shareable, should industrial data be involved in terms of geometries and/or boundary conditions and/or Fluent set up - and others that will be public and then open to dissemination.

2.3.2 FAIR Data

2.3.2.1 *Findable Data*

This section applies only to open Aeronautics data (with no IP restrictions) that gather input/output data or openly shareable results related to Computer Aided Engineering (CAE) simulations carried out on Aeronautical products (Combustor and Turbine).
To make these open data FAIR, two categories of metadata have been identified:

- Engineering metadata,
- Computational metadata.

From the Engineering standpoint, the following mandatory metadata should be used:
- Identifier,
- Creator,
- Title,
- Publisher (e.g. Ge Avio),
- Publication Date,
- Resource Type (e.g. CFD-simulation dataset).

To add further information about the engineering metadata, the following list of optional metadata can be used to allow an effective data traceability:

- Product category (e.g. Turbine, Combustor),
- CAE discipline (e.g. mechanical, fluid dynamic, thermal or multi-discipline),
- Used CAE solver,
- Description (free text),
- Other information (free text).

Finally, from the Computational perspective, the following metadata should be used to guarantee not only the robust identifiability of data but also their proper reproducibility stating the same computational parameters and input file:

- Server,
- Software,
- Software version,
- Job name,
- Wall time,
- No. of CPUs,
- Memory required,
- Input file,
- Output directory

2.3.2.2 *Openly Accessible Data*

Turbine pilot:

It is important to remark in advance that LES and URANS simulations' results connected to the whole turbine design system package will be assumed confidential since they will be then used by AI and HPDA techniques to generate the optimized Turbine Design System, representing one of the WP5 goals in ACROSS project. The Figure 3 below reported summarizes the entire process.

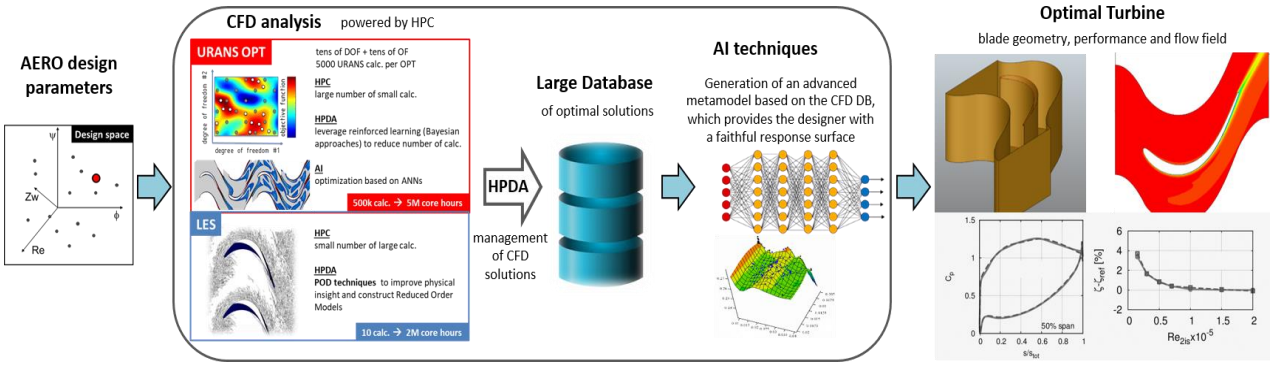

Figure 3 Turbine design System creation steps

Having said that, the LES simulations and its data reduction by means of HPDA may be of interest for the research community. Therefore, results of fluid dynamic simulations coming from an exemplary blade geometry and its advanced post processing may be made available to the community through data files reporting extracted quantities besides videos.
Concerning the different software adopted and developed within ACROSS by the turbine-Pilot team, just the HPDA routines will be open source access for further research activity at University level, given the existing collaboration between CINI-UNIGE and several Universities: University of Cambridge (UoC), University of Melbourne (UoM) and KTH.

Combustor Pilot:
Combustor test cases coming from academic research will be mainly investigated for technical reasons but also not to infringe confidentiality constraints. For openly accessible data, advanced post-processing results of Fluent simulations will be made available for dissemination through selected data files and videos.

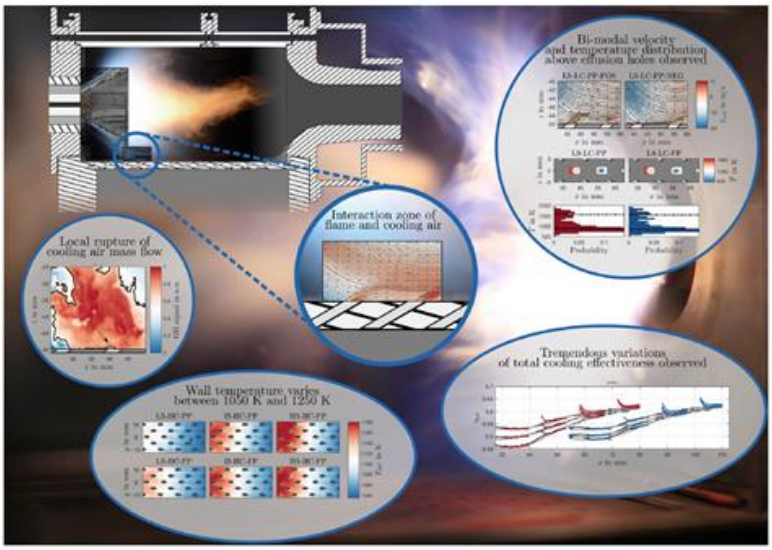

Figure 4 First test case adopted TECFLAM swirl burner

ANSYS CFD-Post software can be used but also *Paraview*, open source SW, can be used to this aim. Should Ge Avio data be involved in terms of geometries and/or boundary conditions and/or Fluent SW set-up, results data can not be disclosed and shared among partners due to IP confidentiality.

2.3.2.3 *Interoperable Data*

Interoperability refers to open data only.

Turbine pilot:

In terms of SW, just the HPDA software will be open source consisting of a series of routines that may be easily adapted between different dataset by changing the data input section. Initially, the interoperability will be granted to the ongoing collaboration with Universities. Standards adopted rely on Fortran and Matlab routines and the interoperability will be improved by releasing some samples (not specifically referring to ACROSS test case). Basic routines of HPDA software can be found available on a dedicated ACROSS platform supporting University courses as “Big data analytics for fluid machinery” the one proposed at University of Genoa. Otherwise, data produced and software used in the turbine pilot are restricted so they will not be available for other purposes.
In terms of advanced CFD results, exemplary blade data analysis will be post-processed and the relevant data will be made available for comparison and interoperability will be granted to other researchers and institutions.

Combustor Pilot:

The U-THERM3D procedure is a customisation of the ANSYS Fluent solver realised within the University of Florence. Due to this reason this pilot does not lend itself very well to inter-operability among different institutions.

2.3.2.4 *Reusable Data*

This applies only to open Aeronautics data, too.
CAE data produced as output from the computer-aided simulations that the aeronautics pilots rely on are deterministic, so the reproducibility of output data is always guaranteed under the same computational parameters and input file.
The produced CFD results open to dissemination will be generated according to standard file formats or structures allowing other researchers, institutions, organisations, countries, etc. to re-use them in order to raise the value of interconnected scientific research.

2.3.3 Allocation of resources

Costs for making data FAIR in ACROSS WP5 will be expressed in terms of PM’s as part of the Project Management task. They will be eligible as part of the Horizon 2020 grants.
Person in charge of data management will be the WP5 leader.
Long-term data preservation will not be envisaged for most of turbine and combustor generated DB‘s.

2.3.4 GDPR

Not relevant, since no personal data will be stored.

2.3.5 Data Security

For sensitive information, because of IP constraints, dataset will be only temporary available on the ACROSS data system infrastructure. After execution of numerical simulations, data will be transferred to GE Avio Digital Technology Systems. Hence, from the archiving and preservation perspectives, no long-term repository will be required.
For openly accessible data, data will be archived on ACROSS data system infrastructure and/or proprietary repository (owned by GE Avio srl or partner involved in the numerical simulation task) throughout the whole duration of ACROSS project. Repositories will be labelled through standardized metadata.

2.3.6 Other issues

No national and/or European projects and related procedures for data management are involved and used.

2.3.7 Plan of the outputs

Four categories of research outputs, specifically related to WP5 activities, have been identified as follows:

- D1: Turbine Pilot Confidential Data,

- D2: Turbine Pilot Open Data (not subject to confidentiality restrictions),
- D3: Combustor Pilot Open Data (not subject to confidentiality restrictions)
- D4: Combustor Pilot Confidential Data,

For each type of foreseen output, a table has been filled in:

Output ID	Item	Description
D1	Dataset name and reference	MFAA-TU-CO (Morfo-AA Turbine Design System Confidential Data). This confidential data-set refers to all the numerical investigations performed to support the Aeronautics Turbine Use case. Identifier - to be defined
	Dataset description	Fluid dynamics data are generated by URANS and LES computations to evaluate the performance of Low-pressure turbine blades. The data are then used, after HPDA and AI application, to elaborate the design system for aeronautical applications. Fluid-dynamics data will comprehend CFD input data (CAD geometry, mesh, code set-up, boundary conditions) and output data properly post-processed.
	Standards and metadata	The data that constitute the database are mainly: • ASCII non formatted file (overall in/output, AI database, etc.) • binary files (starccm+ solution file, ensight gold file format for Paraview post processing) • HDF5 format may be also adopted for HPDA Meta-data will not be applied since they are for internal use only.
	Data sharing	Due to IP constraints, this dataset: • Could be granted only by the ACROSS partners that will contribute to WP5 operations, and exclusively if strictly needed for the purpose of executing the CAE simulations, • Can not be used, disclosed to others, or reproduced, without the express written consent of GE Avio S.r.l.
	Is dataset accessible?	No. Only results in terms of KPI's will be accessible.
	Is dataset reusable?	No
	Dataset archiving and preservation policies (including storage and backup)	Because of IP constraints, the dataset will be temporary available at ACROSS data system infrastructure, only for the purpose of executing numerical simulations supporting the Turbine Pilot Use Case. After conclusion, data will be transferred to secure place as reported hereafter. From the archiving and preservation perspectives, no long-term repository is required to support needs and goals of this case study. The long-term storage and backup of this dataset will be implemented on the internal GE Avio Digital Technology systems.

Table 1 Turbine Pilot Confidential Data

Output ID	Item	Description
D2	Dataset name and reference	MFAA-TU-DIS (Morfo-AA Turbine design System Dissemination Open Data). This data-set includes turbine simulation, based on advanced numerical modeling, and post-processing results open to dissemination. Identifier - to be defined
	Dataset description	This data-set refers to a sample case study of turbine LES results carried out by STARCCM+. Sample case simulation results and its advanced post-processing will be accessible. Results will show the distributions of key flow quantities, under study, in the turbine flow pattern.
	Standards and metadata	The results will be provided in terms of: ➢ videos (".GIF, .AVI") and ➢ post process results (ASCII non formatted file ".TXT", ".H5"), so that they can be used as both qualitative and quantitative comparisons. The video dimensions will be up to 200MB. The following metadata will be adopted: title, Resource Type (e.g. CFD-simulation data-set, Product category (e.g. Turbine), CAE discipline (e.g. CFD-URANS , CFD-LES), Used CAE solver and Software version, Wall time, No. of CPU's, Memory required.....
	Data sharing	Concerning videos, the platform they are located in will provide capabilities for only executing the video to allow visualization, while the post-processing results will be available for download.
	Is dataset accessible?	Yes.
	Is dataset reusable?	CFD post-processed files for exemplary blade can be assumed as reusable
	Dataset archiving and preservation policies (including storage and backup)	Throughout the whole duration of ACROSS project, this dataset will be archived on ACROSS data system infrastructure and/or proprietary repository (owned by Morfo and/or GE Avio srl). The purpose will be aimed at disseminating CFD importance and impacts on aeronautical products performance optimization.

Table 2 Turbine Pilot Open Data

Output ID	Item	Description
D3	Dataset name and reference	UNIFI-CDN-DIS (UNIFI Combustor Dissemination Open Data) This dataset refers to the numerical investigations carried out through UTHERM3D solver open to dissemination. Identifier - to be defined

	Dataset description	This dataset refers to unsteady multi-physics and multi-scale UTHERM3D simulations involving academic combustor test case geometries, with not GE practice-based Fluent set-up.
	Standards and metadata	The results will be provided in terms of: ➢ videos (".GIF, .AVI") and ➢ post process results (non formatted file ".TXT", ".DAT"), so that they can be used as both qualitative and quantitative comparisons. The video dimension will be up to 100MB. Metadata not applicable.
	Data sharing	Concerning videos, the platform they are located in will provide capabilities for only executing the video to allow visualization while post-processing results will be available for download.
	Is dataset accessible?	Yes.
	Is dataset reusable?	CFD post-processed files can be considered reusable.
	Dataset archiving and preservation policies (including storage and backup)	Throughout the whole duration of ACROSS, this dataset will be archived at ACROSS data system infrastructure and/or on proprietary repository (owned by UNIFI). The purpose will be focused on disseminating CFD importance and impacts on aeronautical combustor performance optimization.

Table 3 Combustor Pilot Open Data

Output ID	Item	Description
D4	Dataset name and reference	UNIFI-CDN-CO (UNIFI Combustor confidential Data). This data-set refers to the numerical investigations carried out thru UTHERM3D solver that are considered Confidential. Identifier - to be defined
	Dataset description	This data-set refers to unsteady UTHERM3D simulations involving GE Avio srl combustor test cases. This could involve: • input data (CAD geometry, boundary conditions) • Fluent simulation set-up and meshing strategy • output results properly post-processed.
	Standards and metadata	The data are constituted by ANSYS Fluent standard files (Setup file ".CAS" and solution file ".DAT" for post processing). Metadata not applicable
	Data sharing	Due to IP constraints, this confidential data-set: • Could be granted only by the ACROSS partners that will contribute to WP5 operations, and exclusively if strictly needed for the purpose of executing the CAE simulations, • Can not be used, disclosed to others, or reproduced, without the express written consent of GE Avio S.r.l.
	Is dataset accessible?	No, only results in terms of KPI's will be accessible.
	Is dataset reusable?	No

	Dataset archiving and preservation policies (including storage and backup)	Because of IP constraints, the data-set will be temporary available at ACROSS data system infrastructure, only for the purpose of executing numerical simulations supporting the Combustor Pilot Use Case. After conclusion, data will be transferred to secure place as reported hereafter. From the archiving and preservation perspectives, no long-term repository is required to support needs and goals of this case study. The long-term storage and backup of this data-set will be implemented on the internal GE Avio Digital Technology systems.

Table 4 Combustor Pilot Confidential Data

2.4 Data Management Plan – Weather, Climate, Hydrological and Farming Pilot

2.4.1 Data Summary

Data produced by Weather, Climate, Hydrological and Farming Pilot consists of global and regional-scale maps describing physical variables of the atmosphere, soil, water streams and oceans.
Those are the key input for national/regional weather services, civil protection, water management authorities and many more stakeholders.

Global-scale meteorological and climatological simulations will produce GRIB data. GRIB file format is self-describing, compact and portable across computer architectures. The World Meteorological Organization (WMO) designed and is maintaining the GRIB standard.

In addition to GRIB, we will produce NetCDF data following the CF (Climate and Forecast) convention. NetCDF (Network Common Data Form) is a machine-independent data format that supports the creation, access, and sharing of array-oriented scientific data. It is based on HDF (Hierarchical Data Format), originally developed at the National Center for Supercomputing Applications and now managed by The HDF Group, a non-profit corporation. The conventions for CF metadata are designed to promote the processing and sharing of files created with the NetCDF.

We will use different kinds of earth observation data as input data (satellite imagery, data from meteorological ground stations, etc.). Such data will be provided by ECMWF, MPI, Deltares and Neuropublic for model calibration and model initial condition.

Since we are targeting high-resolution global-scale weather forecasts and large ensemble of climatological simulations, we are targeting several PB of output model data. The size of the output data will greatly depends on the available computing resources. Data will be useful to the overall meteorological and climatological communities that will benefit form beyond state-of-the-art sample data for model development and experimentation.

2.4.2 FAIR Data

2.4.2.1 *Findable Data*

Both GRIB and NetCDF are fully enriched with metadata. In the case of GRIB metadata are fully standardized by WMO and we strictly follow the standard. In the case of NetCDF, we rely on the most common naming convention, the CF extension.
Each ensemble model output is composed of many thousands of fields, organized in a multidimensional hypercube. We usually consider at least the following axis: latitude, longitude, (pressure) level, time, physical variable (temperature, pressure, humidity...), ensemble member, and a model run identifier.

Our domain-specific object store uses such semantic variables for indexing the data, thus user requests are based on metadata, as in the following example:

```
retrieve,
    class=rd,                // research experiment (od = operational)
```

```
expver=1234,          // experiment version identifier
date=20210708,        // date in format yyyyMMdd
time=00:00,
step=0/to/24/by/1,    // 25 time frames (the first day of forecasts)
param=t,              // temperature
domain=g,             // global data
levtype=pl,           // pressure level
levelist=925,850,800  // requested levels
```

Each dataset is clearly identified with an "experiment version" but the experiment version by itself is not enough to clearly define all data provenance. Both ECMWF and MPIM have internal procedures for defining the data provenance, but such procedures are not yet standardized.

In case of NetCDF data, as those generated by WRF regional downscaling, we rely on the CF naming convention that is widespread and widely accepted, but is not a fully binding standard.

2.4.2.2 *Openly Accessible Data*

Datasets produced in the context of WP6 will be openly available for research purposes but, for the time being, may require a licence agreement for data access.
ECMWF is an international organization whose data policies are defined by a committee composed of member states representatives.
Data access policy is moving towards full open data access, but the migration process will last a few years and will end in 2025. For the time being, a valid ECMWF account is required to accessing the data.
Global-scale meteorological data will be archived in the MARS archive and will be accessible through ECMWF web services, the high-performance mars client (requires access to ECMWF computing resources) and through polytope Rest API (https://polytope.ecmwf.int)
Climatological simulations will be archived on DKRZ archival system.

Weather forecasts regional downscaling and hydrological simulations will be executed and hosted on ACROSS computing resources. We will consider publication on Copernicus Data Store (CDS) to improve the dissemination of the ACROSS outcome.

Data are encoded in GRIB and NetCDF binary data format. The most convenient way for accessing the data is the adoption of open-source software libraries (i.e. eccodes and libnetcdf). Many applications support natively such data format.
Even the FDB domain-specific object-store is released in the public domain (https://github.com/ecmwf/fdb), but it is not required to access the data.

2.4.2.3 *Interoperable Data*

WP6 workflows produce datasets encoded in GRIB and NetCDF file format, which are the most common data format in the meteorological and climatological communities.
Such data formats are widely used in all operational and research workflows, so we expect full interoperability with the target researchers, institutions and organizations.
GRIB is a data format designed for streaming. Each atomic subset (GRIB Message) is self-describing since it is enriched with full metadata. A GRIB file is the collection of multiple GRIB Messages.
NetCDF4 is based on HDF5 which is a hierarchical data format, designed for storing complex datasets. It can easily be extended but is less suited to data streaming.

2.4.2.4 *Reusable Data*

Data produced by WP6 will be either available for research purposes or openly available to everyone. In particular, global-scale NWP, at first, are going to be accessible only for research purposes.
We do not plan to have an embargo, so data will be made available as soon as possible.
Given the size of the data, we will set up an agreement with well-known meteorological archives such as MARS or Copernicus Data Store. Such are perpetual archives, so we plan to grant data re-use well beyond the end of the ACROSS project.

2.4.3 Allocation of resources

We are committed to supporting the research community by providing full FAIR access to WP6 Data and we do not plan to charge the related costs.

ECMWF will be responsible for global-scale NWP data, MPIM will be responsible for climatological data, Deltares will be responsible for hydrological data and Neuropublic will be responsible for regional-scale NWP data and farming advisory data.
Following our commitment to permanent data archival, introduced in subsection 2.4.2, we are working on a sustainable long-term archival solution, not based on ACROSS founding.

2.4.4 GDPR

Not relevant, as no personal data are stored.

2.4.5 Data Security

We plan to adopt repositories designed for long-term (perpetual) archival based on semantic indexing of fully standardized metadata.

2.4.6 Other issues

We plan to rely on MARS archive (ECMWF institutional archive, founded by ECMWF member states), Copernicus Data Store (EU founded) or DKRZ climatological archive.

2.4.7 Plan of the outputs

Output ID	Item	Description
D1	Dataset name and reference	IFS NWP
	Dataset description	global-scale numerical weather predictions (5km resolution)
	Standards and metadata	Data encoded in GRIB format. Metadata fully compliant to WMO GRIB standard
	Data sharing	Data will be semantically indexed, retrievable trough MARS language (based on GRIB metadata)
	Is dataset accessible?	Yes. GRIB is a standard data format, widely accepted in the meteorological community. Data access will require a research licence (free of charge)
	Is dataset reusable?	Yes. Data will be made available through a perpetual archive, with full data curation and semantic indexing
	Dataset archiving and preservation policies (including storage and backup)	The archive is designed for perpetual archival, with extensive policies of archive check, tape replication, etc.

Table 5 IFS NWP Dataset

Output ID	Item	Description
D2	Dataset name and reference	ICON Climatological simulations
	Dataset description	global-scale climatological ensemble simulations
	Standards and metadata	data encoded in GRIB format. Metadata fully compliant to WMO GRIB standard
	Data sharing	Data will be semantically indexed, retrievable trough MARS language
	Is dataset accessible?	Yes. GRIB is a standard data format, widely accepted in the meteorological community. Data access will be open.
	Is dataset reusable?	Yes. Data will be made available through a perpetual archive, with full data curation and semantic indexing

	Dataset archiving and preservation policies (including storage and backup)	The archive is designed for perpetual archival, with extensive policies of archive check, tape replication, etc.

Table 6 ICON Climatological simulations dataset

Output ID	Item	Description
D3.1	Dataset name and reference	Hydro-meteorological simulation over Meuse and Rhine
	Dataset description	High-resolution hydrological simulation (1km) performed by WFLOW over Rhine and Meuse basins. Simulation forced by IFS NWP
	Standards and metadata	Data will be stored in NetCDF file format, Metadata will follow CF naming convention
	Data sharing	We plan to share the data through Copernicus Data Store
	Is dataset accessible?	Yes. NetCDF is a standard data format, and metadata compliance with CF convention is a major step towards data discovery.
	Is dataset reusable?	Yes. Data will be made available through a perpetual archive, with full data curation and semantic indexing
	Dataset archiving and preservation policies (including storage and backup)	The archive is designed for perpetual archival, with extensive policies of archive check, tape replication, etc.
Output ID	Item	Description
D3.2	Dataset name and reference	Hydro-climatological simulation over Meuse and Rhine
	Dataset description	High-resolution hydrological simulation (1km) performed by WFLOW over Rhine and Meuse basins. Simulation forced by ICON climatological runs.
	Standards and metadata	Data will be stored in NetCDF file format, Metadata will follow CF naming convention
	Data sharing	We plan to share the data through Copernicus Data Store
	Is dataset accessible?	Yes. NetCDF is a standard data format, and metadata compliance with CF convention is a major step towards data discovery.
	Is dataset reusable?	Yes. Data will be made available through a perpetual archive, with full data curation and semantic indexing
	Dataset archiving and preservation policies (including storage and backup)	The archive is designed for perpetual archival, with extensive policies of archive check, tape replication, etc.

Table 7 Hydro-climatological and Hydro-meteorological simulation over Meuse and Rhine

Output ID	Item	Description
D4	Dataset name and reference	Mesoscale NWP over Europe and Greece
	Dataset description	WRF regional downscaling of IFS NWP over Europe and Greece, with data assimilation of Neuropublic private weather stations.

Standards and metadata	Data will be stored in NetCDF file format, Metadata will follow CF naming convention
Data sharing	We plan to share the data through Copernicus Data Store
Is dataset accessible?	Yes. NetCDF is a standard data format, and metadata compliance with CF convention is a major step towards data discovery.
Is dataset reusable?	Yes. Data will be made available through a perpetual archive, with full data curation and semantic indexing
Dataset archiving and preservation policies (including storage and backup)	The archive is designed for perpetual archival, with extensive policies of archive check, tape replication, etc.

Table 8 Mesoscale NWP over Europe and Greece datasets

2.5 Data Management Plan – Energy and Carbon Sequestration Pilot

2.5.1 Data Summary

The Energy and Carbon sequestration pilot will generate large amounts of data that are simulation results, i.e. the output from running a simulator on one or more model data input cases. In addition, modified model data will be generated as a result of history matching or optimization workflows, analysis data will be generated from new analysis workflows, and computer program source code will be written during the project. The latter is special in that its value is very high, and its size comparatively small, so for source code special procedures are in place that are separate from the other data types: all source code is stored in distributed git repositories on GitHub as well as on developers' computers, giving high redundancy. The treatment of source code will not be further described in this plan.

The workflows associated with the pilot will generate data as described in the following table:

Data origin	Data type	Format	Size	Useful for whom
Simulation results	3D or 4D fields (data varying in time and space), or time series.	ECLIPSE-format restart (UNRST) and summary (UNSMRY) files, HDF5 files.	500 GB estimated for a single workflow execution (1 GB per ensemble member, typical 100 members, 5 iterations)	Application and workflow developers
Model data from history matching or optimization	Grid, property, well and schedule data defining a simulation case.	ECLIPSE format deck.	8-14 GB for a single case	Engineers/simulation users, developers
Analysis data	Time series, 3D or 4D fields.	HDF5	< 1 GB for a single case	Decision makers, users developers
Computer programs	Source code for applications and workflows	C++ and Python code	< 10 MB total	Developers, users

Table 9 Data Summary

2.5.2 FAIR Data

2.5.2.1 *Findable Data*

At this point, there are no metadata associated with any of the output data. We will consider adopting relevant metadata structures and processes from the Open Subsurface Data Universe. We may add manifest files for metadata to existing output data to be compatible with OSDU. See https://osduforum.org
Important open model data will be stored at Zenodo, as well as some data summaries and coarsened results data. See https://zenodo.org. For data relevant for CO2 storage, the CO2 datashare portal will also be used, see https://co2datashare.org.
Small (<3M cells) model data sets may be stored on github, at the opm-data repository curated by the OPM initiative. This especially applies to smaller models useful for testing and benchmarking.
Source code will be stored at GitHub, with releases registered on Zenodo to provide each release with a DOI and archived versions.

2.5.2.2 *Openly Accessible Data*

Results from running with proprietary data models will themselves be proprietary.
Any results from open data models will be shared openly, however a filtering process is required to reduce the amount of data stored to reasonable levels, that we can store in the repositories indicated in the previous section.

For the existing ECLIPSE-compatible input and output formats, open documentation and open source software exists. For new formats based on HDF5 created in this project, we will ensure full documentation is made available, as well as open source libraries for reading and writing the data.

Open data will be licensed under the Open Database License, and the Open Database Content License, as appropriate. See https://opendatacommons.org/licenses/odbl/.

2.5.2.3 *Interoperable Data*

Interoperability will be provided through using well-documented formats with provided open source libraries as described in the above section.

The OSDU metadata standards will be applied to our metadata needs.

2.5.2.4 *Reusable Data*

The licences that will be used (Open Database License and Open Database Content License) are permissive and allow reuse and re-sharing of the data by third parties.
Data will be made available at the latest whenever a scientific paper or popular science article is submitted or published in which the data are used. Also, this will be checked at project milestones and at project end.

We will apply no re-use restrictions, or time restrictions, other than those imposed by the storage repository.

Data will be retained depending on perceived need on a case to case basis. For every large experiment generating more than 10 GB data or using more than 1000 CPU-hours, the WP leader and the person performing the experiment shall discuss and decide on:

- What data shall be retained.
- What, if any, post-processing shall be done.
- Where shall the retained data be stored, for example choosing a repository.
- Update this data management plan (section 2.5.7) with critical information about the resulting data set.

2.5.3 Allocation of resources

The main cost will be in person hours curating and caretaking for the data and covered by the GA [RD.1] grant under the WP7 tasks. We will investigate if direct costs will be accrued as well.

The WP leader, or someone to which the WP leader delegates this responsibility, is responsible for data management.

Long term preservation will be considered for large experiments as discussed above.

2.5.4 GDPR

Not relevant, as no personal data are stored.

2.5.5 Data Security

For sensitive data, such as proprietary data sets, data will be encrypted when stored on file servers and similar. Decryption will happen only on the system where the data set will be run, analysed or processed. Safe storage for long term preservation is subject to practices of the chosen preservation site such as Zenodo or the CO2 datashare portal.
Source code is stored on GitHub and subject to that site's backup procedures, but as the entire git repositories are replicated by every developer working on it, in practice a backup of the full commit history is present on every developer's system.

2.5.6 Other issues

Not applicable.

2.5.7 Plan of the outputs

Output ID	Item	Description
D1	Dataset name and reference	Sleipner-refined-HM
	Dataset description	History matched high-resolution adapted version of the refined Sleipner Benchmark model, with simulation results.
	Standards and metadata	Input data: Eclipse deck input format. Output data: Eclipse binary UNRST output. OSDU metadata.
	Data sharing	To be submitted for the CO2 Datashare portal
	Is dataset accessible?	Yes
	Is dataset reusable?	Yes
	Dataset archiving and preservation policies (including storage and backup)	To be backed up by the CO2 Datashare portal.

Table 10 Sleipner-refined-HM dataset

Output ID	Item	Description
D2	Dataset name and reference	
	Dataset description	Artificial test case for the seismic cube use case, with simulation results.
	Standards and metadata	Input data: Eclipse deck input format, or possibly other format more closely related to the seismic processing data. Output data: Eclipse binary UNRST output. OSDU metadata.
	Data sharing	To be submitted for Zenodo
	Is dataset accessible?	Yes
	Is dataset reusable?	Yes

	Dataset archiving and preservation policies (including storage and backup)	To be backed up by Zenodo.

Table 11 Artificial test case dataset

3 Conclusions

The first release of the Data Management Plan is described in this deliverable and will be maintained and updated during the course of the project. The Data Management Plan is a living document that needs to be well-defined. This document provides an early first structure for a DMP, which is the key element of every data management today. Regularly updated, this document will develop into an extensive knowledge base addressing all aspects of the research data management handling and publishing related to ACROSS project.

References

[1] EC, "H2020 Programme – Guidelines on FAIR Data Management in Horizon 2020," 2016.

[2] EC, "Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020," 2017.

[3] "Data Catalog Vocabulary (DCAT) - Version 2," [Online]. Available: https://www.w3.org/TR/vocab-dcat/.