D1.2

Data Management Plan

Status: Under EC Review

Dissemination Level: Public

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101188168.

Abstract	
Keywords	Data Management, FAIR, GDPR, Metadata, Reusable Data

This document presents the Data Management Plan (DMP) for the RI-SCALE project, which defines the strategy for managing the data generated, processed, and shared throughout the project's lifetime. The plan delineates the types of data handled within RI-SCALE, the standards and formats adopted, and the procedures for ensuring data quality, security, and compliance with legal and ethical requirements. It also describes how the consortium makes sure alignment with the FAIR principles, facilitating findability, accessibility, interoperability, and reusability of data whenever possible.

As the project evolves, the DMP will be regularly updated to reflect new datasets, emerging technologies, and best practices in data management.

Revision History			
Version	Date	Description	Author/Reviewer
V 0.1	08/07/2025	First Draft	Matteo Agati (EGI)
V 0.2	17/08/2025	Entering Use cases datasets information	Matteo Agati (EGI)
V 0.3	27/08/2025	Structure check and optimisation	Andrea Anzanello (EGI)
V 1.0	29/08/2025	Final Draft for Submission	Matteo Agati (EGI)

Document Description			
D1.2 - Data Management Plan			
Work Package Number 1			
Document Type	Deliverable		
Document Status	Under EC Review	Version	1.0
Dissemination Level	Public		
Copyright Status	 This material by the Parties of the RI-SCALE Consortium is licensed under a Creative Commons Attribution 4.0 International License.		
Lead partner	EGI Foundation		
Document Link	https://documents.egi.eu/document/4204		
DOI	https://zenodo.org/records/16993480		
Author(s)	• Matteo Agati (EGI)		
Reviewers	• Andrea Anzanello (EGI) • Małgorzata Krakowian (EGI)		
Moderated by:	• Matteo Agati (EGI)		
Approved by:	Activity Management Board		

Terminology / Acronyms	
Term/Acronym	Definition
API	Application Programming Interface
BBMRI	European Research Infrastructure for Biobanking and Biomolecular Resources
CMIP6	Coupled Model Intercomparison Project Phase 6
CF	Climate and Forecast
CP	Common Programmes
CRC	Colorectal Cancer
DCAT	Data Catalog Vocabulary
DEP	Data Exploitation Platform
DICOM	Digital Imaging and Communications in Medicine
DMP	Data Management Plan
DPIA	Data Protection Impact Assessment
DOI	Digital Object Identifier
EC	European Commission
ECMWF	European Centre for Medium-Range Weather Forecasts
EISCAT	EISCAT Scientific Association
EOSC	European Open Science Cloud
EUCAIM	European Federation for Cancer Images
FAIR	Findability, Accessibility, Interoperability, Reusability
FHIR	Fast Healthcare Interoperability Resources
GIF	Graphics Interchange Format
GDPR	EU General Data Protection Regulation
HDF5	Hierarchical Data Format, version 5

IP	Intellectual Property
JSON	JavaScript Object Notation
MMCI	Masaryk Memorial Cancer Institute
ML	Machine Learning
MUG	Medical University Graz
NetCDF	Network Common Data Form
OME-Zarr	For bioimaging such as microscopy, a consortium called the Open Microscopy Environment (OME) created a format called "OME-Zarr", based on Zarr with some discipline-specific extensions
OMOP	Observational Medical Outcomes Partnership format
ONNX	Open Neural Network Exchange
PID	Persistent Identifier
RI	Research Infrastructure
SP	Special Programmes
SUC	Scientific Use Case
TIFF	Tagged Image File Format
TUC	Technological Use Case
WP	Work Package

Table of Contents

Executive Summary

The deliverable D1.2 "*Data Management Plan*" defines the structure within which the RI-SCALE project will create, manage, and collect data during the project's operations. Furthermore, it will specify how the data will be used or made available for verification and re-use, as well as how the data will be curated and stored once the project is completed.

In addition, the Data Management Plan (DMP) outlines the FAIR[1] (Findable, Accessible, Interoperable, and Reusable) design of the data, agreements on data security are created, and ethical issues associated with data collection/generation are addressed.

The RI-SCALE project adheres to Horizon Europe Open Science FAIR principles and strives to make data as open and as closed as appropriate. The beneficiaries share the project's data in such a way that it is valuable to partners and their users outside the project, while ensuring that the privacy of third parties that participated in the data collection/generation is not violated.

Under these conditions, the data will be managed and released in compliance with the certifications and safeguards of the EU General Data Protection Regulation (GDPR)[2]. Every dataset is examined (in terms of sensitivity, privacy, and security) before an official decision is made on whether or not to make that specific information public.

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm
[2] EU General Data Protection Regulation (GDPR)

1. Introduction

1.1. Scope and Purpose of the Deliverable

This document presents the Data Management Plan (DMP) for the RI-SCALE project. It defines the strategy adopted by the consortium to manage the data generated, collected, and processed during the project's lifetime. The DMP highlights the procedures for making sure that data are managed responsibly, stored securely, and made as openly available as possible, in line with the Horizon Europe Open Science principles and the FAIR data guidelines (Findable, Accessible, Interoperable, Reusable).

RI-SCALE brings together a network of leading Research Infrastructures (RIs), e-infrastructures, data spaces, and technology providers to prototype and validate Data Exploitation Platforms (DEPs) capable of enabling large-scale data analytics, AI model development, and advanced data-driven services. Considering the scale and diversity of data handled by the project - varying from climate simulations and radar observations to biomedical imaging and digital pathology datasets - a strong data management strategy is essential to ensure consistent integration, sharing, and reuse.

Therefore, the purpose of this deliverable is to define:

- The types of data expected within RI-SCALE and their origins;
- How data will be collected, stored, processed, and preserved;
- The measures adopted to ensure data quality, security, and compliance with legal and ethical requirements (including GDPR);
- The strategy to ensure that project data are FAIR and to promote their responsible reuse both within and beyond the project.

As the project progresses, this DMP will be continuously updated to reflect new datasets, evolving methodologies, emerging ethical requirements, and technological advancements introduced by RI-SCALE's partners. The updates will ensure that data management practices remain fully aligned with both the project's needs and the European Commission's Open Science framework.

1.2. Structure of the Document

Section 2 provides an overview of the data generated, collected, and processed within RI-SCALE.

Section 3 outlines the application of FAIR data principles to ensure findability, accessibility, interoperability, and reusability.

Section 4 explains the allocation of resources for data management and long-term preservation.

Section 5 addresses the ethical aspects related to data handling within the project.

Section 6 describes the datasets associated with the work packages.

Section 7 details the datasets per scientific and technological use case, including their characteristics and management strategies.

Section 8 presents the conclusions and next steps for the Data Management Plan.

2. Overview of the Data in RI-SCALE

The RI-SCALE project deals with a large range of heterogeneous datasets generated and reused by four thematic Research Infrastructures - ENES, EISCAT, BBMRI, and Euro-BioImaging - as well as supporting partners providing e-infrastructure services (EGI, SLICES) and connections to European Data Spaces (Copernicus, Destination Earth, EUCAIM). These datasets support the development of the DEP prototypes and the validation of scientific and technological use cases (*WP5 – Co-design and integrated Validation*) across environmental and health domains.

2.1. Types of Data Generated and Collected

RI-SCALE produces and integrates several categories of data:

- **Environmental Data**
  - Climate simulations from the CMIP6 repository and high-resolution projections from Copernicus and Destination Earth;
  - Atmospheric observations from EISCAT radars and EISCAT_3D volumetric measurements;
  - Processed outputs derived from advanced AI models for downscaling, anomaly detection, and predictive simulations.
- **Health and Biomedical Data**
  - Digital pathology images from BBMRI-ERIC, Medical University Graz (MUG), and Masaryk Memorial Cancer Institute (MMCI);
  - Biomedical imaging datasets from Euro-BioImaging, including microscopy images, EMPIAR repositories, and multimodal imaging archives;
  - AI-generated synthetic datasets used to enable privacy-preserving research.
- **AI Models and Derived Data**
  - Foundation and domain-specific models for environmental and biomedical applications;
  - Derived datasets generated through model training, evaluation, and inference;
  - Metadata on model performance, versioning, and reproducibility.
- **Metadata and Catalogues**
  - Standardised metadata describing datasets, provenance, and data quality;
  - Service catalogues enabling the discoverability of DEPs and associated data holdings.

Data formats include NetCDF, HDF5, TIFF, DICOM, JPEG 2000, OME-Zarr, and standardised metadata formats like JSON, XML, and DCAT. For AI-related outputs, formats include ONNX, HDF5, and containerised workflows based on Docker and Kubernetes.

2.2. Data Collection Methodology

The methodologies adopted by RI-SCALE to collect the data vary depending on its type and source:

- **Direct Acquisition from RIs**: Datasets from ENES, EISCAT, BBMRI, and Euro-BioImaging are collected using their existing infrastructure and APIs, ensuring consistency with established community standards;
- **Integration with European Data Spaces**: Datasets from Copernicus, DestinE, and EUCAIM are imported via secure interfaces to ensure compatibility and interoperability;
- **New Data Generated by AI Models**: Training and validation of AI algorithms produce derived datasets, predictions, annotations, and processed outputs.

All collected data are enhanced with standardised metadata to ensure traceability, provenance, and reproducibility.

3. FAIR Data Principles

The RI-SCALE consortium is dedicated to guaranteeing that every dataset created, handled, and shared throughout the project adheres to the FAIR principles. This commitment extends to every phase of the data lifecycle, including collection, storage, curation, publication, and long-term preservation. The Data Exploitation Platforms (DEPs) play a central role in enabling these practices by providing the required infrastructure, tools, and standards for efficiently and consistently managing data across various research fields. Through this coordinated approach, RI-SCALE aims to maximise the quality, accessibility, and impact of its data, while supporting seamless integration with external infrastructures and promoting open science whenever possible.

3.1. Findability of Data/Research Outputs

Within RI-SCALE, each research infrastructure has defined clear strategies to ensure that all project outputs can be easily located and referenced by the scientific community. These approaches rely on the use of persistent identifiers, trusted repositories, and recognised community platforms, tailored to the specific needs of each domain. The following is an overview of the strategies followed by the four Research Infrastructures involved in the project:

- **ENES**: Research output data will be assigned to a DOI for future reference/citation. The research output software and AI models will be made available on GitHub and receive a DOI, e.g., through Zenodo;
- **EISCAT**: Research outputs will be published in the scientific literature with references to the data. Data will, by default, be made available through EISCAT's data servers;
- **BBMRI**: The model CRC model will be published in a scientific journal, and all data sets will be included in the BBMRI-ERIC directory with the appropriate metadata;
- **Euro-BioImaging**: All created models will be made openly and freely available through the BioImage Model Zoo, through which models receive DOIs through Zenodo, and generated image data will be published in the BioImage Archive.

3.2. Accessibility of Data/Research Outputs

As part of the commitment to FAIR data principles, the RI-SCALE research infrastructures are implementing strong strategies to enhance the accessibility of their data and outputs. The following outlines how key initiatives are making their software, datasets, and models openly available to the scientific community and beyond:

- **ENES**: Software will be openly available on GitHub with a FOSS licence. Downscaled datasets at 5km resolution, as well as higher resolution downscaled data (O(100m)) and

associated indicators, will be made available as open access data on community servers from the ENES RI participating institutions and SMEs;

- **EISCAT**: Data are either "common programmes" (CP) or "special programmes" (SP). CPs are run by EISCAT and are publicly available, while SPs are run by EISCAT Members and are embargoed for up to three years. Data from operational scheduling and new data derived from historic radar data will be available immediately after validation. New real-time data products will be available after validation unless affected by restrictions;
- **BBMRI**: The synthetic data for computational pathology will be hosted by BBMRI-ERIC as part of the EUCAIM project and will be freely available for research use via the POSIX file system;
- **Euro-BioImaging**: All created models will be made openly and freely available through the BioImage Model Zoo under a permissive licence (MIT or equivalent). Generated and derived images will use open formats (OME-Zarr) and be made available under CC0 or CC-BY licences, depending on the data from which they were derived.

3.2.1. Repositories

All documents, presentations and other materials that form an official output of the project (not just milestones and deliverables) are placed in the document repository (DocDB[3]) to provide a managed central location for all materials.

In addition, public deliverables and publications will be shared publicly via the Zenodo[4] platform to increase the discoverability of the project outputs.

All profiles, specifications, configuration files, software, workflows, and code will be deposited in GitHub[5].

Therefore, the RI-SCALE project will use DocDB, Zenodo, and GitHub as its standard and main repositories.

3.3. Accessibility of Data/Research Outputs

To promote smooth integration, reuse, and compatibility across different platforms and fields, participating research infrastructures are implementing standardised formats and metadata schemas. These initiatives support the FAIR principle of interoperability, making it easier to share, comprehend, and apply data and research outcomes among scientific communities. The following illustrates how each infrastructure is implementing interoperability to support FAIR data principles:

- **ENES**: Results from the downscaling use case will be produced in NetCDF/Zarr formats and will follow, to the best extent possible, the NetCDF-CF convention;

[3] https://documents.egi.eu/
[4] https://zenodo.org/communities/ri-scale/
[5] https://github.com/

- **EISCAT**: Typically, EISCAT uses HDF5, and all data can be made available in HDF5;
- **BBMRI**: All image data will be available in DICOM format, and metadata will follow the MIABIS standard. Input clinical data is available either in OMOP or FHIR format;
- **Euro-BioImaging**: Image metadata will use the REMBI (Recommended Metadata for Biological Images) schema and be in JSON format. Images from models or from existing data will be created in OME-Zarr format to maximise interoperability. Generated models will follow the BioImage Model Zoo RDF-based specification and controlled vocabularies.

3.4. Reusability of Data/Research Outputs

Reusability ensures that data and research outputs remain valuable beyond their initial use, enabling future studies, validation efforts, and broader scientific impact. By applying standardised licensing, enriching metadata, and archiving high-value datasets, research infrastructures are laying the groundwork for long-term accessibility and reuse.

The following examples highlight how each infrastructure is supporting reusability within the FAIR framework:

- **ENES**: Metadata in research output data will be enriched with comprehensive provenance information;
- **EISCAT**: All data produced are non-repeatable observations of nature. They are archived indefinitely, and they are of very high value to study the processes involved, even many years after the observation;
- **BBMRI**: Data and models will be available by the access rules as defined in the BBMRI-ERIC CRC cohort;
- **Euro-BioImaging**: Data (generated and derived images) and models will be licensed under CC BY 4.0. Hosting models in the BioImage Model Zoo and derived data in the BioImage Archive will maximise reusability. Curation and storage/preservation costs of research outputs are activities out of the scope, as the RI-SCALE research outputs are used for validation and piloting.

4. Allocation of Resources

Any expenses associated with the collection/production of FAIR data during the RI-SCALE activities are included in the project budget. These expenditures will be required to cover a variety of particular data processing and data management operations, ranging from data collection and documentation to storage and preservation to distribution and re-utilisation.

These operations are a component of the WP that processes the relevant data; hence, the needed effort will be part of the relevant WP.

The expenses of long-term data preservation are minimal when using the EGI Online Storage and Google Drive platforms. Using Zenodo and GitHub (both free of charge) ensures that costs for long-term preservation of the data are manageable. When applicable, a more accurate cost estimate will be provided at a later stage of the project.

5. Ethical Aspects

The RI-SCALE project deals with a wide variety of data, including potentially sensitive biomedical datasets and AI-generated outputs. The consortium follows strict ethical guidelines, EU legal frameworks, and best practices for responsible data handling.

A detailed analysis regarding the ethical measures adopted can be found in (RI-SCALE) Deliverable D1.3 - Ethics Requirements and Processes (M&), which provides a comprehensive description of the protocols and safeguards implemented by the RI-SCALE project.

6. Work Packages' Dataset

Data Summary	
Data description: Types of data	1. Project Documentation • Metrics • Risks • Procedures • Plans • Meeting agendas • Meeting participation lists • Meeting minutes • Presentations • Deliverables • Mailing list archive • External Feedback (surveys, events, etc.) • Promotional material (flyers, posters, branding materials, etc.) • Other communication & dissemination material (multimedia, video, etc.)
Data description: Origin of data	All the data will be produced and provided by project members.
Data description: Scale of data	<1GB
Standards and metadata	Plain text formats such as .docx, .txt, .rtf, .pdf, .pptx, .xml, .xls, and .html. Multimedia such as JPG/JPEG, GIF, TIFF, PNG or video formats.
Data sharing: Target groups	All project members and the EC Project Office. Communication activities will be publicly available, focusing on the target audiences of the project (including users, technology providers and infrastructure providers).

Data sharing: Scientific Impact	Scientific Publications in peer-reviewed journals, conferences & events aiming to engage stakeholders.
Data sharing: Approach to sharing	1. Shared within the consortium and the European Commission: • **Presentations**: Public presentations are made public via the Indico portal or external conference pages; • **Deliverables**: All deliverables are shared within the consortium and also with the European Commission. Public deliverables are accessible to everyone via the project website and the Zenodo portal. • **Mailing list archive**: only accessible by the mailing list members; • **Publications** will be available via the project website and the RI-SCALE community on the Zenodo Repository; • **Promotional and other audiovisual material** will be available via the project website. 2. Shared with the Project Office and management boards to support work, as well as with the European Commission. Unless otherwise stated, all content will be available under the CC BY 4.0 license and metadata under the CC0 license. Any consortium-restricted content is shared via an access-protected Confluence space.
Archiving and preservation	Once the project is finished, all the information will be preserved by EGI Foundation for at least 5 years, as well as on the EC funding portal. Publications will also be kept in the Zenodo Community.
Allocation of Resources	
Who will be responsible for data management in your WP/Task?	Work Package Leaders and Task Leaders.

How will long-term preservation be ensured?	Long-term preservation is not needed, except for the contractual 5 years after the project. A copy of all the documentation of the project is kept by the European Commission in the funding portal. Deliverables, publications and other dissemination material are shared via the Zenodo portal, which grants long-term preservation.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	To access the data shared only within the consortium, an EGI SSO account is required. Accounts and access management are the responsibility of the coordinator.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	For security and long-term preservation, RI-SCALE relies on EGI Document Repository, Zenodo and Google Drive platforms.
Other Issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	EGI Foundation will take care of the data according to the ISO 27000 standard for Information security management and GDPR.

7. Datasets per Use case

The following overview describes the Data Management Plan for the Use Cases data that will be generated within RI-SCALE. For each dataset, it describes the type of data and its origin, the related metadata standards, the approach to sharing and target groups, and the approach to archival and preservation.

Beneficiaries must responsibly manage the digital research data generated in the action ('data') in line with the FAIR principles. They should also ensure open access to research data via a trusted repository under the principle "as open as possible, as closed as necessary". The requirements for research data management apply only to data that are generated in the course of the action. Beneficiaries should also consider re-used data when developing their data management plans (DMPs), if they form part of their research and to the extent possible.

- Beneficiaries must establish a DMP, addressing important aspects of RDM.
    - Beneficiaries should maintain the DMP as a living document and update it over the course of the project whenever significant changes arise. This includes, but is not limited to: the generation of new data, changes in data access provisions or curation policies, attainment of tasks (e.g. datasets deposited in a repository, etc.), changes in relevant practices (e.g., innovation potential, the decision to file for a patent), and changes in consortium composition.
      Beneficiaries are encouraged to encode their DMP deliverables as non-restricted, public deliverables, unless there are reasons (legitimate interests or other constraints) not to do so. In the case they are made public, it is also recommended that open access is provided under a CC BY licence to allow broad re-use;
- Beneficiaries must deposit the data in a trusted repository (see explanation above) and ensure open access through the repository, as soon as possible and within the deadlines set out in the DMP.
    - Deposition of data must take place as soon as possible after data production/generation or after adequate processing and quality control have taken place, providing value and context to the data and at the latest by the end of the project. This does not entail that data must be made open, but rather that it is deposited so that metadata information is available and hence information about the data is findable. In exceptional cases in which specific constraints apply (e.g. security rules), deposition can be delayed beyond the end of the project. Data includes raw data, to the extent technically feasible, but especially if it is crucial to enable reanalysis, reproducibility and/or data reuse.

7.1. Scientific Use Cases

7.1.1. SUC1: High-resolution Downscaling of Climate Scenarios and Risk Trend Analysis in Agriculture (ENES)

WP/Task	WP3 - T3.3
Contact	Tullio De Giacomi (tullio.degiacomi@hypermeteo.com) Gianluca Ferrari (gianluca.ferrari@hypermeteo.com)
Established a DMP, addressing important aspects of RDM.	Not in place
Data Summary	
Will you re-use any existing data and what will you re-use it for?	We will re-use reanalysis datasets and climate projections to train the AI model and perform the downscaling procedure.
Will you re-use any existing data and will this generate new data?	As mentioned earlier, for the first question. We will generate new high-resolution climate projections (new data).
What types and formats of data will the project generate or re-use?	NetCDF4, JSON
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	The purpose of the data generated is to describe with higher accuracy possible impacts of climate change in the future on the agricultural sector, computing climate indicators based on high resolution and high spatial representativeness
What is the expected size of the data that you intend to generate or re-use?	TBD (~10TB)

What is the origin/provenance of the data, either generated or re-used?	The data used are available on the ESGF data platform (climate projections) and Copernicus climate data store or other platforms (i.e. API platform of CMCC) for reanalysis data.
To whom might your data be useful ('data utility'), outside your project?	Precision agriculture, climate finance and insurance, urban planning, and disaster risk management.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Research output data will be assigned to a DOI for future reference/citation. The research output software and AI models will be made available on GitHub and receive a DOI, e.g., through Zenodo.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Data will be saved in netCDF format following the CF Metadata Convention where applicable. To save the trained ML models, we will opt for pickle, yaml or equivalent formats.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? Yes
	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting? The metadata of the final datasets will be provided together with the data.
2) Making data openly accessible	

a) Repository:	***Will the data be deposited in a trusted repository?*** Downscaled datasets at 5km resolution, as well as higher resolution downscaled data (O(100m)) and associated indicators, will be made available as open-access data on community servers from the ENES RI participating institutions and SMEs.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? Not yet, but we are planning to use Zenodo.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? Zenodo can provide this solution.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?*** The data will be made openly available under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license, which permits free sharing and adaptation for non-commercial purposes, with proper attribution. For entities interested in commercial use of the datasets (e.g., agri-tech companies, insurers, or other businesses), access will be provided under a dedicated commercial license agreement to ensure fair and sustainable exploitation of the project results.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data

	should be made available as soon as possible. At present, no embargo period is foreseen for the release of the downscaled climate datasets, which will be made openly available for research and educational purposes as soon as they are validated. If required to support scientific publication or IP protection, a short embargo (up to 6 months) may be applied, but this is not currently anticipated
	Will the data be accessible through a free and standardised access protocol? Yes. The downscaled climate datasets will be distributed through trusted repositories (e.g., Zenodo), accessible via standard HTTPS download links associated with persistent identifiers (DOIs). In addition, where possible, access will be provided through community-recognised protocols, such as OGC-compliant APIs (e.g., WMS, WCS, OGC API – Coverages) ensuring full interoperability with widely used scientific software (e.g., Python, R, QGIS, Panoply)
	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? Access to the downscaled datasets will be provided in two ways: • Open access for research and education: during and after the project, datasets will be freely accessible via trusted repositories (e.g., Zenodo, ENES RI servers) under a CC BY-NC 4.0 license. Data will be associated with persistent identifiers (DOIs), ensuring long-term availability and discoverability. • Commercial access: for entities interested in commercial exploitation, Hypermeteo will provide the datasets under a dedicated commercial license. Requests for such access will be handled directly by Hypermeteo through a licensing agreement, ensuring sustainable use of the results beyond the project lifetime.
	How will the identity of the person accessing the data be ascertained?

	Open datasets on Zenodo will not require user identification. Access via the project platform will require a basic user registration (name, email, affiliation), while commercial users will be identified through the licensing agreement process
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? No. The UC1 will not generate or process personal or sensitive data; therefore, there is no need for a data access committee
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** Yes. All metadata will be openly available under CC0, following community standards (e.g. CF Metadata Conventions), and will include DOIs and access information to enable discovery and use of the datasets.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? The downscaled climate datasets will remain available for the long term through trusted repositories (e.g., Zenodo, ENES RI servers), which guarantee data preservation and accessibility for at least 10 years after project completion. Metadata will remain permanently available under CC0, linked to persistent identifiers (DOIs), ensuring that the datasets can always be discovered even if access to the underlying data is discontinued.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? Yes. Full documentation describing how to access, read, and process the datasets will be openly available to all users, ensuring transparency and facilitating reproducibility. For research and educational purposes, relevant software components (e.g., scripts, processing workflows, AI model configurations) will be made available on a private branch of

	the project's GitHub repository, with free access granted upon request, and assigned a DOI for citation. However, in line with Hypermeteo's business model, the commercial exploitation of the software tools will not be open-sourced. Commercial users will have access only to the datasets through dedicated licensing agreements, while the core software developed for commercial purposes will remain proprietary.
3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** Results from the downscaling use case will be produced in NetCDF/Zarr formats and will follow, to the best extent possible, the NetCDF-CF convention.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? We do not foresee the need to create project-specific ontologies or vocabularies, as data and metadata will follow well-established community standards (e.g., NetCDF-CF conventions, ISO 19115, INSPIRE, ESGF/ENES RI guidelines). Should the development of project-specific vocabularies (e.g., for novel climate/agricultural indicators) become unavoidable, they will be openly documented and published, with mappings provided to commonly used standards to ensure interoperability and re-use.
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? Yes, we will add references to CMIP projections, used for example
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses,***

	variable definitions, units of measurement, etc.)? We will provide comprehensive documentation together with the datasets. This will include README files describing the methodology used for downscaling, the variables included, their definitions and units of measurement, and references to the standards adopted (e.g., CF conventions). Jupyter notebooks will be provided to demonstrate how to access, process, and validate the data, as well as example workflows for calculating indicators. Where applicable, additional codebooks and metadata records will be included to ensure full reproducibility and facilitate re-use across disciplines. All documentation will be openly accessible through the same repositories hosting the datasets (e.g., Zenodo, ENES RI servers).
	Will the data produced in the project be usable by third parties, in particular after the end of the project? Yes. All datasets produced will be usable by third parties through open repositories (e.g., Zenodo, ENES RI servers) with persistent identifiers (DOIs) and metadata following community standards (CF conventions, ISO 19115). This ensures that the data remain discoverable, accessible, and reusable beyond the lifetime of the project
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? Each dataset will be assigned a DOI and accompanied by a recommended citation string including authors, year, dataset title, repository name, and DOI (e.g., Author et al., Year, Title, Repository, DOI). The license (CC BY-NC 4.0) will be clearly indicated in the metadata and documentation. Citation and license information will also be included in the README files. Monitoring of impact will be based on usage statistics (downloads, citations) provided by the repositories (e.g., Zenodo, ENES RI servers). No additional monitoring system is foreseen at this stage
	Will the provenance of the data be thoroughly documented using the appropriate standards?

	Yes. The provenance of all datasets will be thoroughly documented, including information on their origin (e.g., ESGF climate projections, Copernicus reanalysis), processing workflows (downscaling procedures, ML model training), and versioning. Metadata will follow community standards (e.g., CF conventions for NetCDF, ISO 19115/INSPIRE where applicable), and each dataset will be assigned a DOI to ensure traceability and reproducibility.
	Describe all relevant data quality assurance processes. Data quality assurance will be ensured by verifying input data consistency and compliance with community standards (e.g., NetCDF-CF conventions), and by documenting provenance and versions. The trained ML models will be validated using standard metrics (MAE, MSE, RMSE) and tested against use-case requirements. Downscaled projections and indicators will also be cross-checked against reanalysis and observational datasets to ensure plausibility and reproducibility.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects. In addition to datasets, the project will produce software components, trained AI models, and documentation. These outputs will be openly shared for research and educational purposes (via GitHub/Zenodo with DOIs), while commercial exploitation will be managed through licensing agreements. Resources (storage capacity, personnel time, repository services) have been allocated to ensure long-term preservation and proper metadata curation. Data security will be guaranteed through repository safeguards, backups, and controlled access where needed. No personal or sensitive data will be processed; therefore, no GDPR issues arise, and ethical considerations are limited to ensuring responsible use of climate projections and derived agricultural indicators.
Other research outputs	

In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or reused throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** *Yes. In addition to datasets, the project will generate digital outputs such as trained AI models, workflows, and processing scripts. These will be made available on a private branch of the project's GitHub repository, with free access granted upon request for research and educational purposes, in order to prevent commercial exploitation without authorization. Commercial-grade software developed by Hypermeteo will instead be managed under dedicated licensing agreements, while no physical research outputs are foreseen.*
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Hypermeteo's Research & Development department will handle data management within the WP under the supervision of the Project Manager and the WP leader.
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long)*** Long-term preservation will be ensured through trusted repositories (e.g., Zenodo, ENES RI servers) that guarantee availability for at least 10 years. Costs are minimal as preservation is covered by repository infrastructures, with only limited internal effort for data curation. Decisions on what to preserve will be taken by the WP leader and project Data Management Officer based on scientific relevance, reuse potential, and storage requirements
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and	The UC1 will not generate or process personal or sensitive data; therefore, no GDPR-related provisions are required. Nevertheless, data security will be ensured through trusted repositories (e.g., Zenodo, ENES RI servers) that provide secure long-term archiving and access via persistent identifiers. For internal management, datasets will be stored and backed up on Hypermeteo's secure AWS S3 infrastructure, ensuring data recovery in case of failure. Transfers between partners and repositories will be

transfer of sensitive data)?	carried out through secure protocols (HTTPS, SFTP, OPeNDAP), and access to restricted data will be controlled via user registration or licensing agreements.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Yes. All downscaled climate datasets will be deposited in trusted repositories such as Zenodo and ENES RI community servers, which ensure secure storage, long-term preservation (at least 10 years), and proper curation. Metadata will be permanently available via persistent identifiers (DOIs), guaranteeing discoverability and access even after the end of the project
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** No. The project does not generate or process personal, sensitive, or otherwise ethically sensitive data. The datasets consist exclusively of climate and environmental information, which does not raise ethical or legal concerns for sharing. Data sharing will therefore comply with FAIR principles and the Grant Agreement without restrictions.
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	No personal data will be processed or used.
If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impact-Assessment (DPIA) been performed?	No personal data will be processed.

Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	No personal data will be processed or used.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	***Please list and briefly describe them.*** Yes. In addition to complying with Horizon Europe requirements, we will follow sectoral best practices for climate and environmental data management, including standards and procedures from the Earth System Grid Federation (ESGF), Copernicus Climate Data Store (CDS), and ENES RI guidelines. At the company level, Hypermeteo will also apply its internal procedures for secure storage and backup (e.g., AWS S3), ensuring consistency and reliability in data management.

7.1.2. SUC2: Smart Detection of Anomalies in Climate Data Usage (ENES)

WP/Task	WP3 - T3.3
Contact	Fabrizio Antonio (fabrizio.antonio@cmcc.it) Donatello Elia (donatello.elia@cmcc.it)
Established a DMP, addressing important aspects of RDM.	In Place
Data Summary	
Will you re-use any existing data and what will you re-use it for?	The use case will re-use data usage information provided by the ENES Data Popularity service (aka ESGF Data Statistics) to predict and identify changes/anomalies in data usage streams across the ESGF infrastructure.
Will you re-use any existing data and will this generate new data?	See previous point.
What types and formats of data will the project generate or re-use?	The use case will re-use data stored in a tabular format in a relational database. **Output**: ML model format (e.g., PyTorch format - pth) anomaly detection information and data usage patterns
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	The use case will validate the Data Exploitation Platform through an AI application that leverages the data download information provided by the integrated data statistics service; in addition, the generated data will uncover trends in data usage, thus driving the DEP caching mechanism and enabling intelligent data replication and caching. Data will be (re-)used to develop an ML-based pipeline for anomaly detection and data trend prediction, in particular for: • Training of the ML model on past data usage information • Inference through ML models on current data download streams

What is the expected size of the data that you intend to generate or re-use?	The expected overall size of re-used data is of the TB order.
What is the origin/provenance of the data, either generated or re-used?	Re-used data comes from a data gathering and processing pipeline which is part of the ESGF infrastructure: data usage information is collected from the ESGF nodes and processed to produce a set of heterogeneous metrics both at the single site and federation level. Data is exposed by the ENES Data Popularity service.
To whom might your data be useful ('data utility'), outside your project?	ENES RI managers can benefit from this generated data, getting useful insights about infrastructure usage and potential issues linked to data transfer failures.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Trained ML models will be identified by PIDs (e.g. Zenodo DOIs) or git commits.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. There are many formats to save a trained ML model, mostly depending on the software used. We'll use existing and widely used formats such as YAML, JSON, and Pickle.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? Yes, for the trained ML models.
	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting?

	Standard formats and existing libraries will be used for metadata extraction.
2) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** *Trained ML models will be stored on recognized repositories (e.g., GitHub, Zenodo).*
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? Not yet. However, in the simplest scenario, it will be Zenodo.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? For ML models, Zenodo can provide the solution.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?*** Yes, open access will be granted to the trained ML model.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. Not applicable.
	Will the data be accessible through a free and standardised access protocol?

	Yes (e.g., via the interfaces offered by Zenodo),
	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? Not applicable.
	How will the identity of the person accessing the data be ascertained? No AuthN/AuthZ will be required.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? No.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** All metadata will be openly available.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? • Trained ML models will be available and findable for at least the project duration; • Concerning metadata, it will be managed via Zenodo, so it will remain available regardless of the data availability.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? For the trained ML models, we will adopt a widely used format, so there is no need for supplementary documentation. Nevertheless, the code for the ML model training and inference will be available on GitHub.

3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** For the ML model, we will try to use the ONNX (interoperable) format when possible.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? In case project-specific outputs are generated, they will certainly be made openly available within the community.
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? Yes, we will include references to the original data (see Data Summary section).
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** README files or Jupyter Notebooks.
	Will the data produced in the project be usable by third parties, in particular after the end of the project? Trained ML models will be openly available in the public domain.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? Citation string.
	Will the provenance of the data be thoroughly documented using the appropriate standards?

	Provenance information will be provided according to the W3C PROV family of standards.
	Describe all relevant data quality assurance processes. For the trained ML model, validation will be performed following ML best practices based on well-known metrics (e.g., MAE, MSE, RMSE) and according to the specific use case.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or reused throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** In addition to trained ML models, we are also considering three more categories of research output, like software, workflows and provenance documents. In that respect, we'll get inspired by FAIR principles according to the available best practices and guidelines (e.g. FAIR4RS).
Allocation of resources	
Who will be responsible for data management in your WP/Task?	All the partners involved in Task 3.3/Scientific Use Case #2.
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).*** Not defined yet.
Data Security	

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Not defined yet.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Not defined yet.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** *No.*
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Not applicable.
If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impa	Not applicable.

ct-Assessment (DPIA) been performed?	
Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	Not applicable.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	***Please list and briefly describe them.*** *No.*

7.1.3. SUC3: Intelligent Scheduling of Radar Observations and Experiments (EISCAT)

WP/Task	WP3
Contact	Thomas Ulich (thomas.ulich@eiscat.se)
Established a DMP, addressing important aspects of RDM.	In Progress
Data Summary	
Will you re-use any existing data and what will you re-use it for?	We will use the existing EISCAT data archive for training AI processes. Thereafter, we will use real-time data to analyse the current geophysical and meteorological situation as well as the status of the facility to evaluate the best use of the facility at the given time.
Will you re-use any existing data and will this generate new data?	The new data generated is a recommendation for how to use the facility in the next, say, 24 hours, which includes the option of not using it at all.
What types and formats of data will the project generate or re-use?	Generated data format: TBD. Used data formats: EISCAT archive of physical atmospheric parameters is available in HDF5 format, and lower-level archives are being converted to HDF5-based formats as well. Real-time meteorological, geophysical and space weather data, depending on originator (there will be many), the data selection has not yet happened.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	The project will develop a process to use data and generate new data from it. The new data is a recommendation on how to use the facility over the next, say, 24 hours.
What is the expected size of the data that	Generated data: very small, kilobytes.

you intend to generate or re-use?	Used data: TBD, see above, but significantly larger, order of 50 GB. Data used for initial training can be 100 TB or more.
What is the origin/provenance of the data, either generated or re-used?	Generated data: origin is the process developed in this project. Used data: EISCAT archives plus multiple providers, data haven't been selected yet, see above.
To whom might your data be useful ('data utility'), outside your project?	None, since it is facility-specific. However, the developed process for evaluating environmental conditions to derive a plan for operations may be applicable to other facilities with similar scheduling problems.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Not planned, but possible. It is facility-specific and of no use to outsiders.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Yes, the data will be accompanied by metadata. However, in a way, the resulting - created - data product is metadata. EISCAT archives are indexed in search portals, including the PITHIA-NRF e-Science Center (eSC). Archive data selected for training can be specified and cited as eSC Static Datasets.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? Possibly. The resulting data product will be archived so that we can review how well the process worked and also to check back regarding why a specific decision was taken and whether the generated data product - a recommendation for facility use - was taken into account by the operator or not.

	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting? *No. Generated data is for internal use only.*
2) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** *Yes, the data will be archived in EISCAT's own archive and data centre.*
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? Yes, our own data centre.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? N/A. See above.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?*** The generated data can be made publicly available; however, ownership remains with EISCAT AB. The usefulness of the data for facilities other than EISCAT is extremely limited.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data

	should be made available as soon as possible. Not applicable.
	Will the data be accessible through a free and standardised access protocol? Typically not, but it's possible. Through EISCAT's own data services.
	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? Restrictions might originate from the providers of the input data ("used data"). These have not been selected, and, therefore, this remains TBD. Access will be provided through EISCAT's own data services.
	How will the identity of the person accessing the data be ascertained? EISCAT uses registration/authorisation through EGI/Perun.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? No.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** Yes, any data provided will have metadata.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? Indefinitely.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)?

	No specific software needed, a web interface or equivalent is sufficient.
3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** N/A in this project. With respect to the EISCAT archive data, see above with respect to PITHIA-NRF, etc.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? N/A; however, PITHIA, etc., apply to the EISCAT archives used. Other data sources TBD and may use varying ontologies
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? Yes. The data product will reference the data sources used to generate it.
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** The process to generate the data will be created as part of this project, and that includes documentation. The validation procedure is documented therein. In this case, it is a human operator who agrees or does not agree with the recommendation (=generated data) from this process. A human operator will always have the last say in how the facility is used.
	Will the data produced in the project be usable by third parties, in particular after the end of the project?

	The data is a real-time application, which is facility-specific. However, the data are archived for future studies of the performance of the process (success statistics). The data are hardly useful for third parties.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? N/A, but if needed, a DOI can be assigned.
	Will the provenance of the data be thoroughly documented using the appropriate standards? Yes.
	Describe all relevant data quality assurance processes. The process will generate a suggestion on how to use the facility over the next, say, 24 hours. A human operator will either agree or disagree with the suggestion and override accordingly. It is possible that the facility will not be used if conditions are not right. After a sufficient trial period of, say, a year, we will look at the success statistics of the process.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects. Ok.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** *The data generated by SUC3 will be used to optimise the use of a large atmospheric research radar, which will produce a large amount of data (≈PB/year). These data are naturally well managed, curated, documented, and DOI'd. They are governed by the data policies of EISCAT AB, including data embargoes. These data will be used in SUC4.*

reused throughout the projects?	
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Thomas Ulich (thomas.ulich@eiscat.se)
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).*** The cost of preserving the data generated by SUC3 is marginal, since the amount is tiny. All EISCAT data are to be preserved indefinitely. Should EISCAT cease to exist, the data will be transferred to permanent national archives, e.g. PAS in Finland.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Backup at a minimum of two separate locations.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	EISCAT's own data centre.
Ethical Aspects	
Are there, or could there be, any ethics or	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).***

legal issues that can have an impact on data sharing?	*No.*
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Not applicable.
If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impact-Assessment (DPIA) been performed?	Not applicable.
Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	Not applicable.
Other issues	
Do you, or will you, make use of other national/funder/secto	TBD

rial/departmental procedures for data management? If yes, which ones?	

7.1.4. SUC4: Space Debris and Anomaly Detection (EISCAT)

WP/Task	WP3
Contact	Thomas Ulich (thomas.ulich@eiscat.se)
Established a DMP, addressing important aspects of RDM.	In Progress
Data Summary	
Will you re-use any existing data and what will you re-use it for?	We will use the existing EISCAT data archive for training purposes to find anomalies and space debris in that same data. We will later apply the process to find anomalies in new data.
Will you re-use any existing data and will this generate new data?	We will use the existing EISCAT data archive as well as future data to find anomalies and space debris. The events, phenomena, and objects we find in the radar data will be the new data product.
What types and formats of data will the project generate or re-use?	TBD. The EISCAT archive data to be used exist in HDF5 and Matlab .mat formats.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	The project will develop a process to use data and generate new data from it. The new data will form a catalogue of events, phenomena, and objects detected in the radar data.
What is the expected size of the data that you intend to generate or re-use?	Generated data: Very small, order of MBs. The generated data is a catalogue of what is found in the radar data with pointers to that data (date, time, geographic location), i.e. in some sense it is metadata. Used data: EISCAT data archive 100s of TB, future radar data is an order of PB/year.

What is the origin/provenance of the data, either generated or re-used?	All data are produced and owned by EISCAT AB.
To whom might your data be useful ('data utility'), outside your project?	The data will be useful to anyone working with the EISCAT radar facilities, exploiting EISCAT's data archive as well as conducting future radar experiments. The process developed within SUC4 will become part of the standard operational data analysis of EISCAT.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Yes.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Yes. The generated data will be information about what has been detected in the radar data, as well as date, time, geographic location, and a pointer to the original source data.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? Yes. It will be possible to browse the data for specific categories, e.g. to list space debris observations, either all or for a given interval, etc.
	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting? *Yes.*
2) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** Yes, the data will be archived in EISCAT's own archive and data centre.

	Have you explored appropriate arrangements with the identified repository where your data will be deposited? Yes, our own data centre.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? We will assign a DOI. See above.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?*** Eventually, yes, but EISCAT's embargoes might apply. TBD.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. No embargo for archive data older than three years. Otherwise, any data older than one year is available only to EISCAT's Associates and Affiliates, and data younger than one year is available only to the PI of the radar experiment and their team. The reason is to give the PI and their team time to publish their findings before others are allowed to study the data.
	Will the data be accessible through a free and standardised access protocol? Through EISCAT's data services.

	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? TBD
	How will the identity of the person accessing the data be ascertained? EISCAT uses registration/authorisation through EGI/Perun.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? Currently, this is not foreseen. However, EISCAT might instate a committee in the future if this is demanded by EISCAT's owners.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?***
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? EISCAT data are archived indefinitely.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? EISCAT's data analysis software is open source, and the documentation is online.
3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** EISCAT archives are indexed in search portals such as PITHIA-NRF eScience Centre (eSC).

	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them?
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? The data generated by SUC4 is metadata describing the actual radar data and events, phenomena, and objects observed therein. Thus, the purpose of these data is to provide pointers to the original radar data to study these events, phenomena and objects.
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** The process to generate the data will be created as part of this project and that includes documentation. The validation procedure is documented therein.
	Will the data produced in the project be usable by third parties, in particular after the end of the project? Yes. The purpose of the project is to provide a catalogue of existing and future data for researchers studying data from the archive, as well as for those running new radar experiments in the future.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? DOI
	Will the provenance of the data be thoroughly documented using the appropriate standards? Yes. All data are generated and owned by EISCAT AB.

	Describe all relevant data quality assurance processes. Resulting event identifications will be verified by a human operator. After one year or new data accumulation, a statistical study will show the error rate. This will be repeated periodically, and the process will be adapted where necessary.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects. Ok.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or reused throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** The data generated by SUC4 will be used to make the EISCAT data archive more accessible. EISCAT operates atmospheric research radars, which will produce a large amount of data (≈PB/year) for the next 30-40 years. These data are naturally well managed, curated, documented, and DOI'd. They are governed by the data policies of EISCAT AB, including data embargoes. The processes developed in SUC4 will be applied to these new data to detect their content and add to the data catalogue.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Thomas Ulich (thomas.ulich@eiscat.se)
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).*** The cost of preservation of the data generated by SUC4 is marginal, since the amount is tiny in comparison to the actual radar data. All EISCAT data are to be preserved indefinitely. Should EISCAT cease to

	exist, the data will be transferred to permanent national archives, e.g. PAS in Finland.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Backup at a minimum of two separate locations.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	EISCAT's own data centre.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** No.
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Not applicable.

If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impact-Assessment (DPIA) been performed?	Not applicable.
Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	Not applicable.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	TBD

7.1.5. SUC5: Colorectal Cancer Prediction with Explainable AI (BBMRI-ERIC)

WP/Task	WP3/T3.4
Contact	Robert Harb (robert.harb@medunigraz.at)
Established a DMP, addressing important aspects of RDM.	In Progress
Data Summary	
Will you re-use any existing data and what will you re-use it for?	We will re-use already scanned Whole-Slide Images from the BBMRI Colorectal Cancer Cohort to train algorithms for patient survival prediction.
Will you re-use any existing data and will this generate new data?	Generated data are predicted survival rates on the image level, and attention maps highlighting image areas relevant for the prediction.
What types and formats of data will the project generate or re-use?	We will use histopathological images in DICOM format.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	Algorithms will be developed for patient-survival prediction from lymph-node whole-slide images (WSIs) of colorectal cancer (CRC). Since lymph nodes are the first anatomical checkpoint in metastatic spread, their microarchitecture and immune-cell composition may contain prognostic signals that routine histopathology overlooks. Deep neural network models will be trained to regress an individual's survival from WSIs directly. By analysing the resulting attention maps and feature-attribution scores, specific microscopic structures will be identified, whose presence or absence systematically correlates with longer or shorter survival times. Such image-derived biomarkers, once validated, could refine adjuvant therapy decisions and advance the understanding of CRC progression at the biological level.

What is the expected size of the data that you intend to generate or re-use?	The training dataset size is around 200 TB.
What is the origin/provenance of the data, either generated or re-used?	The access policy of the training data can be found at: https://www.bbmri-eric.eu/services/access-policies/.
To whom might your data be useful ('data utility'), outside your project?	Medical researchers might gain additional insights into cancer disease progression by investigating which image areas are most relevant for changes in patient survival.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** The images from the BBMRI-ERIC Cohort have unique IDs..
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. The images of the training dataset are part of the BBMRI directory, where they are already discoverable through a standardised set of metadata.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? Yes.
	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting? *Yes.*
2) Making data openly accessible	

a) Repository:	***Will the data be deposited in a trusted repository?*** The predictions of the algorithms will be appended to the already existing Colorectal Cancer Cohort at BBMRI-ERIC.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? BBMRI-ERIC is a partner of the RI-Scale project. A detailed arrangement will be worked out during the project.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? Yes.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?*** The access policies of the BBMRI Colorectal Cancer cohort apply: https://www.bbmri-eric.eu/scientific-collaboration/colorectal-cancer-cohort/.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. There is no embargo.
	Will the data be accessible through a free and standardised access protocol?

	Access is managed through the BBMRI-ERIC negotiator, which provides a standardised procedure for requesting data access.
	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? Data access will be managed through BBMRI, and the licensing of the Colorectal Cancer Cohort applies.
	How will the identity of the person accessing the data be ascertained? Access to the BBMRI negotiator is managed through Life Science RI.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? Yes, data access is managed through the respective rules at BBMRI-ERIC.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** The metadata describing our predictions will be publicly available through the BBMRI directory.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? The metadata will be available through the BBMRI directory, where long-term preservation is guaranteed.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? We will provide documentation on how to load predictions from our model through example code.

3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** Prediction heatmaps will be stored in common image formats.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? We will not use uncommon ontologies.
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? No.
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** We will provide the README files and example code to load the data.
	Will the data produced in the project be usable by third parties, in particular after the end of the project? Yes, third parties will be able to request access to the data.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? Citation will be possible through a DOI.
	Will the provenance of the data be thoroughly documented using the appropriate standards? Yes.

	Describe all relevant data quality assurance processes. The quality of the data will be evaluated through performance evaluation of the algorithm on a hold-out test set.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects. Ok.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or reused throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** Besides predictions from algorithms, there will be no additional data generated.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Robert Harb (robert.harb@medunigraz.at)
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).*** As a pan-European research infrastructure with a legal mandate and secure, multi-year funding, BBMRI-ERIC retains data for the lifetime of the organisation.
Data Security	

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	The data will be hosted at the BBMRI WSI repository, where access is restricted to eligible users and secure storage is managed.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Yes.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** Yes. Deliverable D1.3 "Ethics Requirements and Processes" dedicates § 6.1.5 to Scientific Use Case 5, noting that the colorectal-cancer whole-slide images are "special-category" data and therefore trigger a GDPR-mandated Data-Protection-Impact-Assessment (DPIA) under the permanent oversight of the project's Ethics & Compliance Working Group and Ethics Advisory Board.
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Not applicable.

If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impact-Assessment (DPIA) been performed?	The images used for algorithm training are pseudonymised before being transferred to the DEP. This process includes the replacement of patient-specific identifiers with novel unlinkable IDs and the removal of image data which could leak personal information, e.g. barcodes on slides.
Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	GDPR applies because digitised histopathology slides remain special-category personal data that can potentially re-identify patients, so any processing must rely on a lawful research basis, robust pseudonymisation, and data-controller/processor safeguards. Data Governance Act 2023 applies when those slides (or the synthetic images derived from them) are shared through a data-intermediary or data-altruism arrangement, imposing transparency, neutrality, and reuse-condition rules on whoever facilitates that sharing.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	***Please list and briefly describe them.*** No.

7.1.6. SUC6: Synthetic Data for Computational Pathology (BBMRI-ERIC)

WP/Task	WP3/T3.4
Contact	Robert Harb (robert.harb@medunigraz.at)
Established a DMP, addressing important aspects of RDM.	In Progress
Data Summary	
Will you re-use any existing data and what will you re-use it for?	We will re-use already scanned Whole-Slide Images from the BBMRI Colorectal Cancer Cohort to train synthetic image generation algorithms.
Will you re-use any existing data and will this generate new data?	We will generate novel synthetic histopathology images.
What types and formats of data will the project generate or re-use?	Training data and generated data will be histopathological images in DICOM format.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	Images from the BBMRI Colorectal Cancer Cohort will be used to train algorithms for generating synthetic data. The purpose of generating synthetic images is to address privacy concerns when sharing sensitive medical images through synthetic proxies and to learn semantically meaningful representations of images through data generation.
What is the expected size of the data that you intend to generate or re-use?	The data from the BBMRI ERIC cohort is sized around 200 TB, and we plan to generate data at a similar scale.

What is the origin/provenance of the data, either generated or re-used?	The access policy of the training data can be found at: https://www.bbmri-eric.eu/services/access-policies/. A detailed access policy for the generated data will be worked out during the project.
To whom might your data be useful ('data utility'), outside your project?	Synthetic histopathological images are valuable for computer scientists developing algorithms, as they can augment small or imbalanced datasets, improving model performance and generalisation. Additionally, since synthetic data is not tied to real patients, it enables researchers to work with medically relevant images without navigating complex legal and ethical approval processes.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** The images from the BBMRI-ERIC Cohort have unique IDs. We will also assign unique identifiers to generated images.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. The Colorectal Cancer Cohort is part of the BBMRI directory, which provides an extensive interface to search and navigate datasets: https://directory.bbmri-eric.eu/ERIC/directory/. We plan to release synthetic datasets that will also be part of the directory and indexed in a similar manner.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? The BBMRI directory allows searching by keywords. We will also assign keywords to released synthetic image datasets.
	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting?

	Metadata about the released datasets will be indexed through the BBMRI directory, which can be queried automatically.
2) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** Yes, we plan to release images through the BBMRI WSI repository.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? BBMRI-ERIC is a partner of the RI-Scale project. A detailed arrangement will be worked out during the project.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? Yes.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?*** While access to the non-synthetic source WSIs is managed through BBMRI's full access-committee procedure, which entails detailed legal agreements and institutional sign-offs, the synthetic images will be covered by a separate, lighter policy whose exact mechanics will be worked out during the project. The intention is to replace lengthy contracts with a simple declaration of responsible use, preserving oversight yet making the synthetic data far more readily obtainable than the source images.

	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. There is no embargo.
	Will the data be accessible through a free and standardised access protocol? Access is managed through the BBMRI-ERIC negotiator, which provides a standardised procedure for requesting data access.
	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? Data access will be managed through BBMRI. We will work out detailed licensing during the project.
	How will the identity of the person accessing the data be ascertained? Access to the BBMRI negotiator is managed through Life Science RI.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? Approval for data access on synthetic images will be given by the respective RI that provided the training data for the generation algorithm.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** Metadata in the BBMRI directory will contain instructions on how to retrieve the image data from the BBMRI WSI repository. The metadata describing the dataset will be publicly available through the BBMRI directory.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

	The metadata will be available through the BBMRI directory, where long-term preservation is guaranteed.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? Images can be read with existing open-source software, e.g. openslide. Documentation on parsing annotations accompanying images will be provided through code examples. Annotations will be stored using standardised formats, e.g. GeoJSON, that are parseable with open-source software.
3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** Images will be delivered in DICOM, which is the industry standard for medical images.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? We will not use any uncommon or project-specific ontologies.
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? Synthetic image datasets will include a reference to the source dataset that was used for training.
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?***

	We will provide the README files and example code to load the data.
	Will the data produced in the project be usable by third parties, in particular after the end of the project? Yes, third parties will be able to request access to the data.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? There will be a DOI for datasets that can be cited.
	Will the provenance of the data be thoroughly documented using the appropriate standards? Yes.
	Describe all relevant data quality assurance processes. The quality of generated images will be evaluated through standardised metrics, e.g. FID scores.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects. Ok.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or reused throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** Besides generated images and accompanying metadata, we do not plan to release other research outputs.
Allocation of resources	

Who will be responsible for data management in your WP/Task?	Robert Harb (robert.harb@medunigraz.at)
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).*** As a pan-European research infrastructure with a legal mandate and secure, multi-year funding, BBMRI-ERIC retains data for the lifetime of the organisation.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	The data will be hosted at the BBMRI WSI repository, where access is restricted to eligible users and secure storage is managed.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Yes.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** Yes. Deliverable D1.3 *"Ethics Requirements and Processes"* devotes § 6.1.6 to Scientific Use Case 6, flagging re-identification and bias risks, mandating a GDPR-compliant DPIA and placing the work under continuous Ethics & Compliance Working Group (ECWG) and Ethics Advisory Board (EAB) supervision.

Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Not applicable.
If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impact-Assessment (DPIA) been performed?	The images used for training the synthetic image generation algorithms are pseudonymised before being transferred to the DEP. This process includes the replacement of patient-specific identifiers with novel unlinkable IDs and the removal of image data which could leak personal information, e.g. barcodes on slides.
Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	GDPR applies because digitised histopathology slides remain special-category personal data that can potentially re-identify patients, so any processing must rely on a lawful research basis, robust pseudonymisation, and data-controller/processor safeguards. Data Governance Act 2023 applies when those slides (or the synthetic images derived from them) are shared through a data-intermediary or data-altruism arrangement, imposing transparency, neutrality, and reuse-condition rules on whoever facilitates that sharing.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data	No.

management? If yes, which ones?	

7.1.7. SUC7: Foundational Models for Heterogeneous Biological Image Data (Euro-BioImaging)

WP/Task	WP3/T3.4
Contact	Teresa Zulueta-Coarasa (teresaz@ebi.ac.uk)
Established a DMP, addressing important aspects of RDM.	In Place
Data Summary	
Will you re-use any existing data and what will you re-use it for?	All data in this scientific use case will be re-used from public datasets available at the BioImage Archive (BIA). This data will be used to train foundation models that capture domain understanding of heterogeneous image data, and to fine-tune natural segmentation models.
Will you re-use any existing data and will this generate new data?	All data in this scientific use case will be re-used from public datasets available at the BioImage Archive (BIA). The foundation models used here will be able to generate embeddings which support multiple downstream use cases, such as categorisation, similarity search and derived measurements. The natural segmentation models will generate new annotations, such as segmentation masks.
What types and formats of data will the project generate or re-use?	All of the data used for the use case will be image data. Data formats in the bioimage field are numerous, and, therefore, we will use various image formats such as TIFF, OME-Zarr, or PNG.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	Data reuse in this context will be instrumental to train foundation models that will allow understanding and categorising large cohorts of mixed life sciences imaging data. This will increase the value of archived data at the BIA through better automated categorisation and search.
What is the expected size of the data that	Approximately 100TB.

you intend to generate or re-use?	
What is the origin/provenance of the data, either generated or re-used?	All data will be shared from the BIA, where users deposit it openly under permissive licenses that allow reuse.
To whom might your data be useful ('data utility'), outside your project?	Due to its heterogenous nature, the data used in this use case will be useful to a diverse cohort of life scientists. It will be also useful to AI model developers training foundational models.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Yes, each submission on the BIA is assigned a persistent identifier (i.e an accession number and a DOI) to ensure global traceability and interoperability.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. We will follow two different community developed metadata standards. We will use the REMBI metadata standard for the image metadata and the MIFA schema for the metadata of image annotations (such as segmentation masks). These standards will ensure rich and consistent descriptions of datasets, including details on sample preparation, imaging parameters, biological context, annotation creation method, and provenance.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? Yes, the BIA implementation of the REMBI metadata model includes a mandatory field of keywords.

	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting? *Yes, metadata will be shared in JSON (JavaScript Object Notation) format which is machine-readable and machine-actionable.*
2) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** All data re-used for this use case is from the BIA, a FAIR and open repository from EMBL-EBI. Any resulting annotations and derived outputs will be publicly shared on the BIA.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? Yes, this use case is led by team members of the BIA so all arrangements are covered. We can guarantee all data has, or will be assigned, assigned a unique identifier.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? Yes, all data will be assigned identifiers, and the BIA resolves these to digital objects.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?*** All data used in this use case will be publicly available under CC0 or CC-BY-4.0 licenses that allow data reuse.

	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. Not applicable.
	Will the data be accessible through a free and standardised access protocol? All data on the BIA can be freely accessed through direct download (HTTPS), FTP or Globus.
	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? Not applicable.
	How will the identity of the person accessing the data be ascertained? The BIA is a public repository open to all, and the identity of users accessing the data is not monitored.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? The BIA is a public repository open to all, and the identity of users accessing the data is not monitored.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** Yes, metadata is available to users via permissive licenses. The metadata contains information about the license, allowing users to know that all data is accessible and reusable. We also assign permanent identifiers to improve findability and accessibility.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

	All datasets (including metadata) submitted to the BIA will remain permanently accessible as part of the scientific record.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? Our metadata model includes optional fields where users can explain how the data was processed, and how the annotations were created. This includes information about what software was used. We also have a provision to include relevant links to code repositories if needed.
3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** At the BIA, we are part of the European-funded project FoundingGIDE. This project aims to bring together image data resource owners and research infrastructures from Europe, Australia and Japan to develop the basis of interoperable image data repositories. As part of FoundingGIDE, we are developing guidelines on metadata standards and ontologies, which will be published in autumn 2025. The BIA plans to adhere to those guidelines once they are published. In the meantime, the BIA uses the REMBI and MIFA community-developed metadata standards. We also use ontologies when possible, such as the NCBI organismal classification ontology (NCBITAXON) for organismal taxonomy, and the Biological Imaging Methods Ontology (FBBI). We allow our users to submit their data in any format they choose to encourage deposition. We then convert a subset of images from each dataset to the open and cloud-ready image format OME-Zarr to increase interoperability.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly

	publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? Not applicable.
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? If a new dataset is created during this project, it will contain a link to the original dataset from which it was derived. If newly created segmentation masks or other annotations are submitted as a separate submission to the BIA, we will link each annotation to the corresponding source image on the original study.
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** For any new dataset generated, the metadata following the REMBI standard includes information on image analysis. We will also add a link to the source code that will contain the README files to be able to reuse them.
	Will the data produced in the project be usable by third parties, in particular after the end of the project? Yes, any datasets produced during this project for this use case will be openly available to the public on the BIA.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? All submissions at the BIA website have a "cite" button to facilitate a citation string. Downloads are monitored and can be tracked through the project lifetime.
	Will the provenance of the data be thoroughly documented using the appropriate standards? Yes.

	Describe all relevant data quality assurance processes. All data used in this use case will be curated to ensure the quality of the annotations using the BioImage Archive's standardised internal mechanisms.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects. The data will be stored at the BIA, which is part of EMBL-EBI (European Bioinformatics Institute). EBI maintains rigorous standards for data security and ethics to ensure the responsible management of biological data. Data submitted to EMBL-EBI resources is securely stored and managed using robust infrastructure. Ethical considerations are embedded in all data processing workflows. The other output of this use case is AI models that will also be openly shared according to the FAIR principles on the BioImage Model Zoo, an open repository of AI models.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or reused throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** The other output of this use case is AI models that will also be openly shared according to the FAIR principles on the BioImage Model Zoo, an open repository of AI models.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	The BIA team will be in charge of managing the data for this use case.

How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).*** EMBL-EBI has policies that ensure long-term data preservation https://www.ebi.ac.uk/long-term-data-preservation/. Policies on data resource life cycle management and retirement, resource, staffing and infrastructure continuity, data resource backup and recovery, and distributed delivery through international consortia ensure that the data will be available for the foreseeable future.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	EMBL-EBI infrastructure is distributed in three discrete data centres in different geographical locations to guarantee data protection. Many resources make use of geo-dispersed file storage, which has automatic failover and recovery. Operating multiple instances also allows load balancing between the data centres, ensuring that services continue to be available to the public in the event of interruption of a single service instance. Geo-dispersed backup via public cloud is also an increasingly popular option. No sensitive data will be used in this use case.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Yes, at the BIA.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** No.
Will informed consent for data sharing and long-term	The only personal data shared here will be the dataset's authors' names, affiliations and emails. Authors have to complete a data

preservation be included in questionnaires dealing with personal data?	protection declaration similar to GDPR when they create an account to submit data to the BIA.
If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impact-Assessment (DPIA) been performed?	No anonymisation will be carried out. No patient information is shared with our data, and the submitters of the data agree to have their names, emails, and affiliations publicly available as part of our metadata standard. This ensures that scientists can be credited for the data they produce and deposit at the BIA.
Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	As part of EMBL, the BIA follows EMBL's data protection framework. This framework has been adapted to the needs of international scientific research, and it reflects the principles of European data protection law while remaining within the boundaries of EMBL's international legal status. More information can be found here: https://www.embl.org/info/data-protection/.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	No.

7.1.8. SUC8: Generative AI-Powered Assistant for Data Discovery and Analysis (Euro-BioImaging)

WP/Task	WP3/T3.4
Contact	Wei Ouyang (wei.ouyang@scilifelab.se)
Established a DMP, addressing important aspects of RDM.	In Progress
Data Summary	
Will you re-use any existing data and what will you re-use it for?	Yes, we will reuse the database and example image dataset hosted by the BioImage Archive. We use it to support the AI agent development to demonstrate how users can interact with the DEP through a natural language interface for image analysis.
Will you re-use any existing data and will this generate new data?	Yes, we will reuse existing data, and it will generate new conversational data for logging user interaction with the AI agents and analysis results.
What types and formats of data will the project generate or re-use?	The data we will use is mostly metadata about depositions in the Bioimage Archive; they are database records about different studies, which will be exported in JSON and imported into our vector databases for the agent to perform information retrieval. We will also use biological image data from the BioImage Archive.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	The data is reused to develop and validate an AI assistant that enables natural language access to and analysis of imaging datasets from the BioImage Archive. This supports RI-SCALE's goal of improving accessibility and reuse of RI data through AI-powered services within the DEP.
What is the expected size of the data that you intend to generate or re-use?	We expect to re-use up to 100 TB of imaging data from the BioImage Archive and generate approximately 1–5 GB of conversational logs, metadata embeddings, and analysis outputs during the validation of the AI assistant.

What is the origin/provenance of the data, either generated or re-used?	The reused data originates from the BioImage Archive, a public repository hosted by EMBL-EBI that contains biological imaging datasets and associated metadata. The generated data, including interaction logs and analysis outputs, will be derived from user interactions with the AI assistant during validation activities within the DEP environment.
To whom might your data be useful ('data utility'), outside your project?	The data will be useful to researchers in bioimage analysis, developers of AI tools for life sciences, and RI operators seeking to improve user interfaces and data accessibility. It may also benefit the broader scientific community interested in conversational AI, multimodal search, and reproducible imaging workflows.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Yes, reused datasets from the BioImage Archive already have persistent identifiers.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Yes, rich metadata will be provided, including dataset identifiers, model information, task type, analysis context, timestamps, and interaction intent where applicable. Metadata will include user query summaries, dataset IDs from the BioImage Archive, model names and versions used, analysis parameters, timestamps, and anonymized session IDs. For vector databases used by the AI assistant, metadata will also include embedding indices and source context. We will follow metadata standards adopted in the bioimaging community, such as OME (Open Microscopy Environment) for imaging data, and JSON-LD or schema.org annotations for machine-actionable metadata in AI services.

	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? Yes, keywords will be provided for searching conversations and analysis results.
	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting? Yes, we will provide metadata in JSON format and provide machine-actionable format to enable further automated harvesting.
2) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** No, we do not plan to deposit the user's conversation history or analysis results.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? No.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? Not Applicable.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** No, we do not plan to publish chat conversations of the users for privacy considerations. ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using***

	standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license? No.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. Not applicable.
	Will the data be accessible through a free and standardised access protocol? Not Applicable.
	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? Not applicable.
	How will the identity of the person accessing the data be ascertained? Not Applicable.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? No.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** No.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? Not Applicable.

	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? Not Applicable.
3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** Not Applicable.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? Not Applicable.
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? Not Applicable.
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** Not Applicable.
	Will the data produced in the project be usable by third parties, in particular after the end of the project? No.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact?

	Not Applicable.
	Will the provenance of the data be thoroughly documented using the appropriate standards? Not Applicable.
	Describe all relevant data quality assurance processes. Not Applicable.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects. Not Applicable.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or reused throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** Yes, for the software, we will follow FAIR principles.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Wei Ouyang (wei.ouyang@scilifelab.se)
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).***

	We do not have a long-term data preservation policy established; this will be decided later.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	We aim to have at least 1 copy of the data for recovery, but we are not sure where the data will be stored yet; it depends on the computing platform for deploying the cluster.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Yes, ideally, but not yet decided how.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** Yes, the conversation history will only be shared if the user provides consent, and we need to operate according to GDPR.
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Yes.

If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impact-Assessment (DPIA) been performed?	We do not have any plan to reuse/process the generated chat conversations.
Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	GDPR for personal data.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	Yes, we will manage data according to the procedures of our university (KTH Royal Institute of Technology).

7.2. Technological Use Cases

7.2.1. TUC1: Scalability on EuroHPC with Destination Earth

WP/Task	WP5/T5.3
Contact	Thomas Geenen (thomas.geenen@ecmwf.int)
Established a DMP, addressing important aspects of RDM.	In Progress
Data Summary	
Will you re-use any existing data and what will you re-use it for?	Reuse data from anemoi to run the technical use case on a EuroHPC system.
Will you re-use any existing data and will this generate new data?	The model training will produce model weights.
What types and formats of data will the project generate or re-use?	The model weights will be generated in XX format. The training data will be in ZARR format and contain meteorological data.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	The use of the data will be to be able to train the model, and the model weights are the output of a training run.
What is the expected size of the data that you intend to generate or re-use?	Input datasets are on the order of 50 TB output will be small.

What is the origin/provenance of the data, either generated or re-used?	The datasets will be pulled from anemoi and have their origin on the ECMWF MARS system that is under the control of ECMWF data provenance.
To whom might your data be useful ('data utility'), outside your project?	Anyone who is using anemoi, training AI/ML models for meteorological purposes.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** No (in the future could be requested, DOI).
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Anemoi contains a lot of relevant metadata, but can only be accessed by users with access to Anemoi. At the moment, we do not follow any standards for the metadata; they do not yet exist in our community. The metadata that is collected would be model configuration data (versions) used, model versions, etc.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? Yes, there is a metadata catalogue.
	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting? No. In the future, yes, when the data has a DOI.
2) Making data openly accessible	

a) Repository:	***Will the data be deposited in a trusted repository?*** Currently, it sits in a private ECMWF repository.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? Not Applicable.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? Yes, there are identifiers, but not yet to a DOI.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?*** The data that will be used in the context of the project will be made available.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. Not applicable.
	Will the data be accessible through a free and standardised access protocol? For the data that is used in the project, yes.

	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? Not applicable.
	How will the identity of the person accessing the data be ascertained? Anonymous access is supported.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? No.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** CC-BY-4 license will be applicable.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? We do not make any guarantees.
	Will documentation or reference about any software be needed to access or read, or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? We do not make any guarantees.
3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** Not in scope for our role in this project. https://codes.ecmwf.int/grib/param-db/.

	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? Not in scope for our role in this project. https://codes.ecmwf.int/grib/param-db/.
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? No.
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** Anemoi documentation is publicly available: https://anemoi.readthedocs.io/ https://codes.ecmwf.int/grib/param-db/
	Will the data produced in the project be usable by third parties, in particular after the end of the project? Yes, if we decide to move the training weight back into anemoi. However, it is not the scope of the project or our contribution.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? Not Applicable.
	Will the provenance of the data be thoroughly documented using the appropriate standards? Yes.

	Describe all relevant data quality assurance processes. The data follows the standard quality assurance procedures commonly used in NWP for operational data.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or reused throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** The model is open-source and documented; it could be put on Hugging Face.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	ECMWF.
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).*** Not Applicable.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as	ECMWF data security policies are applicable to the data in the ECMWF systems.

secure storage/archiving and transfer of sensitive data)?	
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Not Applicable.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** No, for the data used in the project context
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Not Applicable.
If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impact-Assessment (DPIA) been performed?	Not Applicable.

Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	Not Applicable.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	No.

7.2.2. TUC2: Advanced Image Compression

WP/Task	WP5/T5.3
Contact	Robert Harb (robert.harb@medunigraz.at)
Established a DMP, addressing important aspects of RDM.	In progress
Data Summary	
Will you re-use any existing data and what will you re-use it for?	Histopathological Whole Slide Images (WSIs) from the Scientific Use Cases 5 and 6.
Will you re-use any existing data and will this generate new data?	This use case converts WSIs from the Scientific Use Cases 5 and 6 to a different image compression format.
What types and formats of data will the project generate or re-use?	WSIs in DICOM format.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	Images are converted with advanced image compression algorithms to decrease loading and transfer times during AI model training. **This compression is done purely for internal usage in the DEP as a pre-processing step**. It is not planned to share the compressed images.
What is the expected size of the data that you intend to generate or re-use?	Around 200 TB of WSI data from the Scientific Use Cases 5 and 6.
What is the origin/provenance of	Respective Origin/provenance from the SC5 and SC6 applies.

the data, either generated or re-used?	
To whom might your data be useful ('data utility'), outside your project?	Not Applicable.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Not Applicable.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Not Applicable.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? Not Applicable.
	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting? Not Applicable.
2) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** Not Applicable.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited? Not Applicable.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

	Not Applicable.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?*** Not Applicable.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. Not applicable.
	Will the data be accessible through a free and standardised access protocol? Not Applicable.
	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? Not Applicable.
	How will the identity of the person accessing the data be ascertained? Not Applicable.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? Not Applicable.

c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** Not Applicable.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? Not Applicable.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? Not Applicable.
3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** Not Applicable.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? Not Applicable.
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? Not Applicable.
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?***

	Not Applicable.
	Will the data produced in the project be usable by third parties, in particular after the end of the project? Not Applicable.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? Not Applicable.
	Will the provenance of the data be thoroughly documented using the appropriate standards? Not Applicable.
	Describe all relevant data quality assurance processes. Not Applicable.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects. Not Applicable.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or reused throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** Not Applicable.
Allocation of resources	

Who will be responsible for data management in your WP/Task?	Robert Harb (robert.harb@medunigraz.at)
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).***
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Not Applicable.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Not Applicable.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** Not Applicable.
Will informed consent for data sharing and long-term preservation be	Not Applicable.

included in questionnaires dealing with personal data?	
If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impact-Assessment (DPIA) been performed?	Not Applicable.
Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	Not Applicable.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	No.

7.2.3. TUC3: Green Computing Improvement

WP/Task	WP5/T5.3
Contact	Bernd Saurugger (bernd.saurugger@tuwien.ac.at)
Established a DMP, addressing important aspects of RDM.	In progress
Data Summary	
Will you re-use any existing data and what will you re-use it for?	We do not plan to re-use any external datasets containing personal, biomedical, or third-party proprietary data. However, existing **benchmarking datasets** (e.g. standard AI model input data such as image or tabular datasets used for inference testing) may be reused solely for the purpose of generating runtime and energy consumption measurements across different hardware platforms. These datasets are publicly available, well-documented, and do not contain any sensitive information. Their role is limited to providing consistent and reproducible input for performance comparisons.
Will you re-use any existing data and will this generate new data?	Yes, we will re-use existing **public benchmarking datasets** to ensure comparability and reproducibility in inference performance tests. These datasets will serve as input for AI models deployed on different hardware platforms (e.g. GPU and GROQ). This process will generate **new data**, specifically structured records of runtime and energy consumption during inference tasks. These newly generated datasets will include metrics such as execution time, power usage, batch size, and system configuration, and will be used to evaluate and compare the energy efficiency and performance of different computing architectures.
What types and formats of data will the project generate or re-use?	The project will generate and re-use the following types and formats of data: **Re-used Data:** Standard, publicly available **benchmarking datasets** (e.g. image classification or tabular datasets) used as input for AI model inference.

	These datasets are typically in formats such as **CSV**, **JSON**, or image file formats (e.g. PNG, JPEG) and do not contain any personal or sensitive data. **Generated Data**: Structured measurement data from inference runs on various hardware platforms. This includes: • **Runtime metrics** (e.g. inference time, throughput); • **Energy consumption data** (e.g. power draw, energy per inference); • **Hardware and system configuration logs** (e.g. device type, batch size). These data will be stored in structured and interoperable formats such as **CSV**, **JSON**, and optionally **Parquet** for efficient analysis and future reuse.
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	The purpose of the data generation and re-use is to evaluate and compare the **performance and energy efficiency** of different hardware architectures, specifically GPU and GROQ cards, during AI model inference tasks. Re-used benchmarking datasets serve as standardized inputs to ensure consistency and reproducibility across tests. The newly generated runtime and energy consumption data directly support the project's objective of promoting **green computing** within Data Exploitation Platforms (DEPs). By systematically measuring and analyzing hardware efficiency, the project aims to provide validated guidance on selecting energy-efficient and high-performance compute architectures for scalable AI workloads in research infrastructures.
What is the expected size of the data that you intend to generate or re-use?	The expected size of the **re-used benchmarking datasets** is relatively small, typically ranging from a few megabytes (MB) to several hundred megabytes, depending on the dataset type (e.g. tabular or image-based). The **newly generated runtime and energy consumption data** will be structured and lightweight, primarily consisting of numerical logs and

	metadata. The total volume is expected to remain below **5 gigabytes** (**GB**) over the entire duration of the project. This includes logs from repeated inference runs across multiple hardware configurations, model versions, and batch sizes.
What is the origin/provenance of the data, either generated or re-used?	The **re-used data** originates from **publicly available benchmarking datasets** that are widely used in the AI community for model evaluation. These datasets are sourced from established repositories (e.g. UCI Machine Learning Repository, ImageNet subsets, or similar) and are selected for their stability, openness, and lack of sensitive content. The **generated data** originates entirely from **performance measurements** conducted within the project environment. These measurements are produced by executing AI inference workloads on different hardware platforms (GPU and GROQ) under controlled conditions.
To whom might your data be useful ('data utility'), outside your project?	• **Researchers in AI systems and hardware benchmarking** who require empirical data on inference performance and energy consumption across different architectures (e.g. GPU vs. GROQ) for comparative studies; • **Developers and operators of AI infrastructure**, who need guidance on selecting energy-efficient hardware for scalable AI workloads; • **Policy makers and sustainability analysts** interested in measuring and reducing the environmental impact of AI and HPC systems; • **Hardware designers**, who can use the performance feedback to improve or optimize future AI acceleration hardware.
FAIR Data	
1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Yes, data that is intended for long-term storage and potential public sharing will be identified by a **Persistent Identifier** (**PID**), such as a **Digital Object Identifier** (**DOI**). This applies particularly to curated subsets of the generated data (e.g. performance and energy

	benchmarks) that are published in association with deliverables, technical reports, or open datasets. The assignment of PIDs will follow the policies of the institutional repository or EOSC-compliant service used for archiving (e.g. Zenodo, TU Wien Research Data Repository), ensuring long-term accessibility, citation, and traceability.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Yes, rich metadata will be provided to ensure the discoverability, interpretability, and reusability of the data. The metadata will describe both the **technical context** of the data collection and the **content** of the datasets. **Metadata to be created includes**: • Title, authors, and institutional affiliations; • Date and location of data collection; • Description of hardware used (e.g. GPU model, GROQ card specifications); • AI model details (e.g. architecture, version, inference parameters); • Input dataset references (with source and license); • Measurement parameters (e.g. batch size, duration, sampling frequency); • Data formats and file structures; • Licensing and access conditions; • Persistent identifier (e.g. DOI). **Standards to be followed**: In the absence of a discipline-specific metadata standard for energy benchmarking in AI systems, we will adopt the **Dublin Core** standard for general metadata and align with **FAIR principles**. For technical metadata related to compute environments and performance

	measurements, we will define a structured schema using **JSON-LD** or **YAML**-based descriptors, ensuring machine readability and extensibility.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? Yes, **search keywords will be included in the metadata** to enhance discoverability and support potential reuse of the data. These keywords will be selected to reflect the technical scope, hardware platforms, and application domain of the data.
	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting? Yes, metadata will be provided in a **machine-actionable format** to support automated harvesting and indexing. When data is published through an institutional or EOSC-compliant repository (e.g. Zenodo, TU Wien Research Data Repository), the metadata will be exposed via standard protocols such as **OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)** and/or Schema.org markup. Metadata will be available in structured formats like **Dublin Core**, **DataCite**, or **JSON-LD**, depending on the repository's framework. This ensures that aggregators, search engines, and harvesting services (e.g. OpenAIRE, Google Dataset Search) can automatically retrieve and index the metadata, enhancing the data's visibility and reusability across platforms.
2) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** Yes, the data will be deposited in a **trusted repository** that ensures long-term preservation, accessibility, and compliance with FAIR principles. The preferred options include: • **Zenodo** (operated by CERN, OpenAIRE-compliant, assigns DOIs); • **TU Wien Research Data Repository** (institutional repository meeting European standards).

	Have you explored appropriate arrangements with the identified repository where your data will be deposited? Yes, appropriate arrangements have been considered for data deposition. TU Wien provides institutional support for its **Research Data Repository**, which is available to project members and aligns with FAIR and OpenAIRE standards. Internal guidelines and storage quotas have been reviewed to ensure compatibility with the expected dataset size and metadata requirements. In parallel, **Zenodo** has been evaluated as a secondary or complementary option, particularly for public datasets linked to project deliverables or publications. Zenodo's support for DOIs, integration with GitHub, and compliance with Horizon Europe open data mandates make it a suitable alternative. Final decisions on repository use will align with the data type (public or restricted), licensing terms, and the project's dissemination strategy.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? Yes, the selected repositories (**TU Wien Research Data Repository** and **Zenodo**) both ensure that each deposited dataset is assigned a **Digital Object Identifier** (**DOI**). These repositories also guarantee **resolution of the DOI to a digital object**, meaning that when the DOI is accessed (e.g. via a browser or API), it leads to a landing page containing the dataset, its metadata, and access/download options.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.***

	Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license? **Most of the data generated in the project will be made openly available** to permit the widest possible re-use, in line with the FAIR principles and the obligations set out in the Grant Agreement. The openly shared data will include **non-sensitive, structured performance and energy consumption measurements** (e.g. inference runtime, power usage, batch sizes, system configurations). These datasets will be published under a **standard open license**, such as **CC BY 4.0** (Creative Commons Attribution), allowing re-use with proper attribution. However, **certain datasets may be subject to restricted access** under the following conditions: • **Contractual or technical limitations** related to hardware vendors (e.g. GROQ), if performance data is covered by confidentiality agreements or non-disclosure clauses; • **Internal logs or configurations** that may expose security-relevant details of the infrastructure setup will be excluded or anonymized prior to release. Where data cannot be shared publicly, a **justified restriction will be documented** in the Data Management Plan, and access may be provided upon request under controlled conditions or bilateral agreement.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. Not Applicable.
	Will the data be accessible through a free and standardised access protocol?

	Yes, the data will be accessible through **free and standardized access protocols** supported by the chosen repositories. This typically includes **HTTP/HTTPS** for web access and **OAI-PMH** or **RESTful APIs** for metadata harvesting and programmatic access.
	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project? **During the project**, restricted datasets will be accessible to project partners via secured internal repositories or authenticated access portals managed by TU Wien or other hosting institutions. If external parties request access, this can be granted on a case-by-case basis under **data use agreements** (**DUAs**) or **non-disclosure agreements** (**NDAs**), where necessary. **After the project**, metadata will remain public, and access to restricted datasets may still be granted upon justified request, following a documented review and access policy.
	How will the identity of the person accessing the data be ascertained? For **openly available datasets**, no identity verification will be required—data will be accessible without authentication through public repositories. For **restricted datasets**, identity verification will be handled through **institutional authentication systems**, repository-level **user account management**, or via **request-based access workflows**. In such cases: • Users will be required to register with a verified institutional email address; • Access requests may involve the submission of a justification form and acceptance of a Data Use Agreement (DUA); • Identity may be further verified via affiliation checks or digital signatures, depending on the sensitivity of the dataset and legal obligations.

	The selected repository (e.g. TU Wien Research Data Repository) supports access control mechanisms and can log access for auditability when necessary. All identity verification and data access procedures will comply with GDPR and institutional data governance policies.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? No, a formal **Data Access Committee** (**DAC**) is not required for this project, as the data generated and used does **not contain personal**, **sensitive**, or **biomedical information**. All datasets consist of technical performance metrics (e.g. runtime, energy consumption) and do not fall under categories requiring ethical review or special access governance. However, in cases where **restricted access** may apply (e.g. due to contractual obligations with hardware vendors), **access requests will be evaluated by the responsible partner institution** (e.g. TU Wien), following clearly defined internal procedures. This ensures transparency and accountability without necessitating a formal DAC.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** Yes, all **metadata will be made openly available** and will be **licensed under a public domain dedication** (**CC0**), in full alignment with the requirements of the Grant Agreement. The metadata will include clear and structured information to enable users to locate, understand, and access the corresponding data. This includes: • Persistent identifiers (e.g. DOI); • Title and description of the dataset; • Authors and affiliations; • Licensing terms and access conditions;

	• Data collection methods and provenance; • Format and file structure; • Links or direct URLs to the data landing page or download location.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? The data will remain available and findable for **a minimum of 5 years after the end of the project**, in accordance with institutional and Horizon Europe data retention guidelines. If deposited in long-term trusted repositories such as **Zenodo** or the T**U Wien Research Data Repository**, data availability may be extended indefinitely, subject to repository policies. Even if the data is withdrawn or becomes unavailable (e.g. due to hardware dependencies or licensing expiration), the **metadata will remain publicly accessible and persistently citable**. Repositories like Zenodo and institutional platforms guarantee that metadata and persistent identifiers (e.g. DOIs) will remain resolvable and searchable, thereby ensuring long-term discoverability and traceability.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)? Yes, documentation and references to any **software tools required to access**, **read**, or **process the data** will be included alongside the datasets. This ensures transparency, reproducibility, and usability of the data. Where applicable, the following will be provided: • **Versioned references** to software or libraries used (e.g. for parsing logs, aggregating performance data); • **Documentation** on usage, parameters, and dependencies; • **Instructions or scripts** (e.g. Python, Bash) to help automate data analysis or visualization.

	If custom tools or scripts are developed during the project, they will be **released as open-source software** under a permissive license (e.g. MIT or Apache 2.0) and hosted on a public platform such as **GitHub**. These repositories will be linked directly from the dataset metadata and landing pages, ensuring users can reproduce and extend the results using the exact tools used in the project.
3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** To ensure interoperability and enable data exchange and reuse within and across disciplines, the project will follow **community-endorsed standards**, **vocabularies**, **and best practices**, as outlined below: **Data Formats**: • **CSV**, **JSON**, and optionally **Parquet** for structured numerical data (e.g. runtime, energy consumption logs); • **TXT** or **YAML** for configuration snapshots and metadata descriptors. These formats are widely used, machine-readable, and easily parsed across platforms and disciplines. **Metadata Standards and Vocabularies**: • **Dublin Core** and **DataCite** for general dataset metadata (e.g. title, authors, license, date, subject); • Schema.org/Dataset for search engine indexing and semantic discoverability; • **OpenAIRE Guidelines** to support integration with EOSC and European data infrastructures; • **CC0** licensing for metadata, as per Horizon Europe requirements. **Methodologies and Best Practices**:

	• Adoption of the **FAIR principles** (Findable, Accessible, Interoperable, Reusable); • Use of **persistent identifiers** (**DOIs**) for datasets and software; • **Version control** for datasets and associated scripts; • **Open-source tooling** for processing and visualizing the data (e.g. Python-based analysis scripts released via GitHub); • Clear **provenance information** (e.g. hardware specifications, model version, input source dataset).
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? Yes, if it becomes necessary to use **uncommon or project-specific ontologies or vocabularies** (for example, to describe detailed hardware performance metrics or inference workload characteristics not covered by existing standards), we will ensure they are: 1. **Mapped to commonly used ontologies** wherever possible, such as: • Dublin Core or DataCite for general metadata; • Schema.org for dataset discoverability; • Standard terms in HPC and AI benchmarking (e.g. SPEC, MLPerf, OpenMetrics). 2. **Openly published and documented**, preferably in **machine-readable formats** such as RDF, OWL, or JSON-LD, and deposited in open repositories (e.g. GitHub, Zenodo) with permissive licenses (e.g. CC BY or CC0). This will allow the vocabularies to be reused, refined, or extended by the broader community, and will support **semantic interoperability** within EOSC and related data infrastructures. Any project-specific extensions will be versioned, clearly scoped, and accompanied by usage examples to facilitate adoption.

	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? Yes, the data will include **qualified references** to other relevant datasets to support transparency, reproducibility, and contextual understanding. These references will be embedded in the metadata using **standard fields** (e.g. relatedIdentifier in the DataCite schema) and will include: • **References to input datasets** reused during inference benchmarking (e.g. public AI datasets used for model evaluation), with DOIs or stable URLs; • **Links to related project datasets**, such as other performance logs, system configurations, or intermediate results generated under different tasks or hardware setups; • **References to previous research datasets**, where benchmarking or performance comparisons are based on earlier studies or baseline measurements. These references will be **qualified** using standardized relationship types (e.g. isDerivedFrom, isSupplementTo, isReferencedBy), ensuring machine-readability and semantic clarity. This enables better integration with research graphs, citation tracking systems, and EOSC-wide data discovery services.
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** To ensure validation, transparency, and reusability, the project will provide comprehensive documentation for each dataset through multiple, standardized means: 1. **README files**: Each dataset will include a README file detailing: • Purpose and context of the data; • Data collection methodology (e.g. hardware used, inference setup, sampling intervals);

- Description of variables, units of measurement, and expected value ranges;
- File structure and format descriptions (e.g. column headers in CSV, JSON schema);
- Data cleaning or preprocessing steps applied;
- Versioning and update history;
- Licensing and citation instructions.

2. **Codebooks or Data Dictionaries**:

Where structured datasets contain multiple parameters (e.g. inference results, power logs), a codebook will be included listing:

- Variable names and definitions;
- Units and possible value ranges;
- Example entries for clarification.

3. **Analysis Scripts and Configuration Files**:

All custom scripts or tools used for data generation and analysis will be provided via open-source repositories (e.g. GitHub), with links in the metadata. These will include:

- Source code (e.g. Python, Bash);
- Environment configuration (e.g. requirements.txt, environment.yml);
- Instructions for reproducing key results.

4. **Metadata Records**:

Rich metadata will be embedded at the dataset-level (e.g. in DataCite, JSON-LD) and file-level (e.g. in YAML or JSON descriptors) to ensure interoperability and semantic clarity.

Will the data produced in the project be usable by third parties, in particular after the end of the project?

Yes, the data produced in the project will be designed and documented to be usable by third parties both during and after the project. This will be ensured by providing clear metadata, thorough documentation, standard data formats, and open access where possible. Data deposited in trusted repositories with persistent

	identifiers will remain accessible and findable long-term, facilitating reuse by researchers, DEP operators, AI developers, and other stakeholders beyond the project's lifetime.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? Third parties should formally cite the data using a standardized citation string that includes: the dataset title, authors or contributors, the year of publication, the repository name, and the persistent identifier (DOI) resolving to the dataset landing page. The data will be licensed under an open license (e.g., Creative Commons Attribution 4.0 International, CC BY 4.0), and the license notice will be clearly included in the metadata and documentation. Monitoring of citations and downloads will be performed through repository analytics and DOI tracking services to evaluate the dataset's impact and reuse.
	Will the provenance of the data be thoroughly documented using the appropriate standards? Yes, the provenance of the data will be thoroughly documented following appropriate standards such as the W3C PROV model and relevant community best practices. This documentation will capture the full lifecycle of the data, including its origin, generation processes, transformations, and any derived products. Provenance metadata will be embedded in machine-readable formats within dataset records to ensure transparency, reproducibility, and trustworthiness.
	Describe all relevant data quality assurance processes. • Data quality assurance will be ensured through multiple, systematic processes: • Raw data collection will use calibrated and validated hardware to guarantee accurate measurement of runtime and energy consumption.

	• Automated validation scripts will check for completeness, consistency, and correctness of recorded data, flagging missing or anomalous values for review. • Data preprocessing steps will include normalization and error correction according to predefined rules, with all transformations logged. • Version control will be applied to datasets and processing code to track changes and maintain reproducibility. • Periodic peer reviews of data and documentation will be conducted within the project team to verify adherence to quality standards. Finally, metadata completeness and compliance with standards will be verified before data deposition in trusted repositories.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects. The project will address research outputs beyond data, including software, documentation, and models. These outputs will be managed with version control, proper documentation, and open licensing where possible to facilitate reuse and transparency. Resource allocation will ensure sufficient storage, backup, and computational capacity to handle data processing and preservation needs throughout and beyond the project duration. Data security measures will include controlled access, encryption where needed, and compliance with relevant institutional and legal regulations to protect sensitive information, especially regarding proprietary performance data. Ethical aspects will be reviewed continuously to ensure compliance with applicable standards, even though the data is technical and non-personal, focusing on responsible handling, sharing, and reporting of results.
Other research outputs	

In addition to the management of data, are you also considering and planning for the management of other research outputs that may be generated or reused throughout the projects?	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** Yes, the management of other research outputs, primarily digital, such as software, workflows, protocols, and AI models, is considered and planned throughout the project. These outputs will follow FAIR principles to ensure they are findable, accessible, interoperable, and reusable. This includes using version control systems, providing rich metadata and documentation, applying open licenses where possible, and depositing relevant materials in trusted repositories. Physical outputs are not applicable in this project context.
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Bernd Saurugger (bernd.saurugger@tuwien.ac.at)
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).*** Decisions about which data to keep and for how long will be made jointly with the consortium partners, considering the costs of storage and the potential scientific and practical value of the data. Criteria will include data relevance to ongoing and future research, compliance with funder and institutional policies, and the feasibility of long-term maintenance. Cost assessments for storage and preservation will be balanced against the anticipated benefits of data reuse, with priority given to data that supports key project objectives and community needs.
Data Security	
What provisions are or will be in place for data security	Data security provisions include encrypted storage and transfer protocols to protect data confidentiality during handling and archiving.

(including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Regular backups and disaster recovery plans will be implemented to ensure data recovery in case of loss or corruption. Access to sensitive data, such as performance and energy consumption metrics, will be restricted to authorized personnel through authentication and role-based permissions. All security measures will comply with institutional and legal requirements to safeguard data integrity and confidentiality throughout the project lifecycle.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Yes, the data will be safely stored in trusted repositories that provide long-term preservation and curation, ensuring data integrity, accessibility, and compliance with best practices and standards.
Ethical Aspects	
Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** No.
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Not Applicable.
If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impa	Not Applicable.

ct-Assessment (DPIA) been performed?	
Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	This question is not directly applicable since no personal data is processed. However, general compliance with relevant EU and national regulations regarding data security and management will be ensured.
Other issues	
Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	Yes. The project will follow the data management policies and procedures of the participating institutions, including TU Wien's guidelines for research data management and relevant national and European funder requirements, such as Horizon Europe Open Science and FAIR data principles.

7.2.4. TUC4: Credit Management System

WP/Task	WP4/T4.4
Contact	Nikolaos Triantafyllis (ntriantafyl@admin.grnet.gr) Nicolas Liampotis (nliam@grnet.gr)
Established a DMP, addressing important aspects of RDM.	Not in Place
Data Summary	
Will you re-use any existing data and what will you re-use it for?	The Credit Management System (CRMS) will re-use existing data from DEP compute centers, including: • Resource Usage Data: CPU/GPU hours, storage, network transfers to track and attribute usage; • Environmental Impact Metrics: Green-index indicators, e.g., Energy consumption; • User and Project Metadata: User IDs, group affiliations, project details from IAM systems to associate usage and credits; • Policy Configurations: Pre-existing rules for credit translation and distribution. These data will be reused to track resource consumption and environmental impact metrics, allocate credits equitably, and ensure compliance with governance and sustainability goals. Potential reasons for discarding data include: Privacy/Security Concerns, non-compliance with GDPR or other regulations.
Will you re-use any existing data and will this generate new data?	The CRMS might re-use resource usage and environmental impact metrics, user/project metadata, and policy configurations. This will generate new data: • Usage Reports: Aggregated reports of metrics for auditing and infrastructure statistics.
What types and formats of data will	Both re-used and generated data; Resource Usage and Environmental Impact Metrics, User/Project Metadata, and Policy Configurations will be in structured formats (e.g., JSON).

the project generate or re-use?	
What is the purpose of the data generation or re-use and its relation to the objectives of the project?	The purpose is to enable secure, equitable, and sustainable access to computational resources, aligning with RI-SCALE's goals of enhancing data access, AI-driven analysis, and resource management.
What is the expected size of the data that you intend to generate or re-use?	The size of re-used and generated data is expected to be correlated under several factors, including the granularity of the resource usage and environmental impact metrics (e.g., daily), the number of metric types collected (e.g., CPU usage, energy consumption), the variety of resources (e.g., HPC clusters, Cloud instances) chosen for data collection, the number of user records and projects, and the defined data retention policy. Future scaling of data volume will be directly influenced by these parameters.
What is the origin/provenance of the data, either generated or re-used?	Re-used Data: • Resource Usage: DEP compute centers; • User/Project Metadata: IAM systems (e.g., Keycloak, OIDC/DID); • Policy Configurations: DEP operators/RI administrators. Generated Data: • Usage Reports: Resource Usage & Environmental Impact Tracking logical component.
To whom might your data be useful ('data utility'), outside your project?	Useful to: • RIs: For optimizing resource management/credit models; • Compute Centers: For benchmarking efficiency/sustainability; • Policy Makers/Funding Agencies: For sustainable funding policies; • Researchers/Data Scientists: For computational trend analysis; • Sustainability Researchers: For environmental footprint studies.
FAIR Data	

1) Making data findable, including provisions for metadata	***Will data be identified by a persistent identifier?*** Projects or Node entities in the CRMS may use a persistent identifier as a reference, where permitted by project policies. User data will use non-identifiable unique IDs provided by the AAI service, ensuring anonymity despite ORCID/PIDs for users.
	Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how. Metadata might be provided, and in that case, they will be designed to support findability within the DEP ecosystem and, where compliant, in external federated data ecosystems. In that case, the type of metadata will be subject to the project's agreements and WP decisions.
	Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use? This depends on the type of metadata provided and the manner in which it would be made available.
	Will metadata be offered in such a way that it can be harvested and indexed? Will you expose metadata in a machine-actionable format to enable automated harvesting? The applicability depends on the structure and semantics of the provided metadata, as well as the integration method used for its dissemination. In that case, it will be subject to the project's agreements and WP decisions.
2) Making data openly accessible	
a) Repository:	***Will the data be deposited in a trusted repository?*** CRMS data will not be deposited in a trusted repository due to concerns over sensitive data and potential privacy exposure. All data will be held only within CRMS itself.
	Have you explored appropriate arrangements with the identified repository where your data will be deposited?

	No arrangements have been explored for depositing data in a repository.
	Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object? If data is available, the repository will assign an identifier to the data and will ensure that the identifier resolves to the corresponding digital object.
b) Data:	***Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why clearly, separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects, it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*** ***Will your data be made openly available in the public domain to permit the widest re-use possible? Will your data be licensed using standard re-use licenses, in line with the obligations set out in the Grant Agreement? Under which license?*** Even if data is available, it might not be open. Legal Reasons: GDPR prohibits sharing sensitive data due to privacy risks.
	If an embargo is applied to give time to publish or seek the protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible. Embargo is subject to the project's agreements and WP decisions.
	Will the data be accessible through a free and standardised access protocol? The CRMS will provide a dedicated API interface, enabling its users to retrieve and interact with their respective data, according to their role in the project.
	If there are restrictions on use (such as licenses), how will access be provided to the data, both during and after the end of the project?

	Access to CRMS data will be restricted to privileged and associated users for the duration of the project; no access will be granted once the project concludes, unless specified otherwise in the project's agreements and WP decisions.
	How will the identity of the person accessing the data be ascertained? The identity of the person accessing the data will be ascertained through the DEP Authorisation Framework, e.g., OpenID Connect (OIDC), which provides secure, standards-based authentication and user identity verification.
	Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)? The establishment of a data access committee will depend on project-level governance decisions, the sensitivity of the data involved, GDPR compliance requirements, and the capacity of associated partners to support such a process.
c) Metadata:	***Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*** The openness and licensing of metadata will be subject to the project's strategic direction and legal evaluation, particularly with respect to GDPR and privacy risks.
	How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available? Data availability and persistence will be determined based on project-level policy decisions, which are subject to the project's agreements and WP decisions.
	Will documentation or reference about any software be needed to access or read or process the data be included? Will it be possible to include the relevant software (e.g. in open-source code)?

	Inclusion of documentation or references to software required for accessing, reading, or processing the data will be considered based on project-level policies and the partners' capacity to support such efforts. Where applicable and permissible, relevant tools and software - preferably open-source - may be included or referenced to ensure data usability and reproducibility.
3) Making data interoperable	***What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*** In case data or metadata are delivered, the CRMS will use standard formats like JSON. Community-endorsed practices, such as open API standards, will be followed to support interoperability, with metadata availability depending on partner capacity.
	In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining, or extending them? Project-specific vocabularies will be mapped to common standards where feasible, with mappings provided based on project policies, if it is needed. Non-sensitive vocabularies may be openly published for reuse, subject to metadata support capacity.
	Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)? CRMS will not include qualified references to other data.
4) Increase data re-use	***How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*** Documentation, such as guides and data descriptions, will be provided depending on project agreements and partners' efforts.

	Will the data produced in the project be usable by third parties, in particular after the end of the project? Certain data may be available to third parties, based on project policies and privacy compliance, with post-project access relying on partners' commitment to sustain data availability.
	How should third parties formally cite the data (citation string, recommended license notice, landing-page DOI)? Will you monitor citations/downloads to evaluate impact? Third parties will cite data using a project-specified format and license. Citation monitoring most probably will not occur, but it is subject to the project's agreements and WP decisions.
	Will the provenance of the data be thoroughly documented using the appropriate standards? In that case, data origins will be documented using standard formats, ensuring traceability, supported by partners' efforts and resources as defined by project agreements.
	Describe all relevant data quality assurance processes. Quality assurance will involve checks for data accuracy and compliance with privacy rules, supported by partners' efforts and resources as defined by project agreements.
	Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security, and ethical aspects. No research outputs other than data are expected.
Other research outputs	
In addition to the management of data, are you also considering and planning for the management of other research outputs that	***Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) Are those also following FAIR principles?*** No research outputs other than data.

may be generated or reused throughout the projects?	
Allocation of resources	
Who will be responsible for data management in your WP/Task?	Data management for the CRMS will be handled by project partners, as defined by project agreements, with responsibilities allocated based on roles and expertise within the Work Package. It is subject to the project's agreements and WP decisions.
How will long-term preservation be ensured?	***(Costs and potential value, who decides and how and what data will be kept and for how long).*** Long-term preservation of CRMS data and metadata will depend on project-level policies and partners' capacity to maintain storage and access infrastructure. Preservation efforts will prioritise compliance with privacy regulations and resource availability, guided by partner agreements.
Data Security	
What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?	Data security for the CRMS will include encryption for data in transit and at rest, secure access controls, and anonymisation of sensitive data to comply with privacy regulations, as guided by project agreements. Data recovery mechanisms, such as backups, will be implemented based on partner resources and commitment. Secure storage and transfer will rely on project-defined protocols, with partner efforts ensuring compliance and protection of sensitive data.
Will the data be safely stored in trusted repositories for long-term preservation and curation?	Data will be kept in CRMS only, in a secure and reliable manner.
Ethical Aspects	

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?	***Yes or No. (If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action).*** Yes. Ethical and legal issues, particularly related to privacy and GDPR compliance, may impact data sharing due to sensitive data like user/project records. These will be addressed through project agreements.
Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?	Informed consent for data sharing and long-term preservation will be obtained when personal data is involved.
If personal data are processed: what anonymisation/pseudonymisation techniques, and has a Data-Protection-Impact-Assessment (DPIA) been performed?	Personal data will be anonymised or pseudonymised to protect privacy, using techniques aligned with GDPR requirements.
Which specific EU/national laws apply (e.g. GDPR for personal data, Data Governance Act 2023, forthcoming Data Act 2025 for cloud portability & interoperability)? Describe compliance steps and responsible roles.	GDPR and potentially the Data Governance Act 2023 apply, with compliance ensured through encryption, secure tokens, and consent management, but it is subject to the project's agreements and WP decisions.
Other issues	

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	The CRMS will not adopt national, funder, sectoral, or departmental data management procedures.

8. Conclusions

The RI-SCALE Data Management Plan provides a shared framework for managing the different datasets produced and used across the project. Through the definition of common standards, formats, and procedures, the consortium makes sure that data are handled securely, remain of high quality, and comply with all legal and ethical requirements. The plan also underpins the adoption of the FAIR principles, encouraging practices that make data easier to share, reuse, and integrate within the broader European research landscape.

As RI-SCALE progresses, new datasets, technologies, and workflows will continue to emerge. For this reason, the DMP is designed as a living document that will be reviewed and updated whenever necessary to reflect the evolving needs of the project. This approach guarantees that data management remains consistent, transparent, and aligned with both the project's objectives and the Horizon Europe Open Science framework.