



**D2.1**

# **DEP Data Management Specification and Roadmap**

Status: Final

Dissemination Level: Public



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



## Abstract

### Key words

Data Lifecycle Management, Research Infrastructures, Data Orchestration, FAIR Data Principles

This deliverable defines the data management architecture and roadmap for the RI-SCALE Data Exploitation Platform (DEP), a federated system that enables Research Infrastructures (RIs) to manage and analyse large-scale scientific data using AI. It introduces the Data Lifecycle Management (DLM) subsystem, which supports secure, policy-driven data movement across cloud and HPC environments using tools like Rucio and FTS.


Additional components cover data preparation, discovery, and popularity-based replication to optimise analysis workflows while limiting the use of storage space. Integration with federated identity systems ensures compliance with FAIR and access control requirements. Case studies from climate science, space research, biobanking, and bioimaging demonstrate real-world applications.

The deliverable also outlines a phased roadmap toward full platform deployment, establishing the DEP as a foundation for cross-domain data reuse and integration with European Data Spaces.

## Revision History

Version	Date	Description	Author/Reviewer
V 0.1	21/07/2025	First Draft	Moderator: Martin Barisits (CERN)
V 0.2	25/07/2025	Submitted for Internal Review	Moderator: Martin Barisits (CERN) Reviewer: Andrea Manzi (EGI), Carl-Fredrik Enell (EISCAT)
V 0.3	19/08/2025	Corrected version	Moderator: Martin Barisits (CERN) Reviewer: Andrea Manzi (EGI), Carl-Fredrik Enell (EISCAT), Gergely Sipos (EGI)
V 1.0	25/08/2025	Final Draft ready for Submission	Moderator: Martin Barisits (CERN)



Document Description			
D2.1 – DEP Data Management Specification and Roadmap			
Work Package Number 2			
Document Type	Deliverable		
Document Status	Final	Version	1.0
Dissemination Level	Public		
Copyright Status	 <p>This material by Parties of the RI-SCALE Consortium is licensed under a <a href="https://creativecommons.org/licenses/by/4.0/">Creative Commons Attribution 4.0 International License</a>.</p>		
Lead partner	CERN		
Document Link	<a href="https://documents.egi.eu/document/4199">https://documents.egi.eu/document/4199</a>		
DOI	<a href="https://zenodo.org/records/16993600">https://zenodo.org/records/16993600</a>		
Author(s)	<ul style="list-style-type: none"> <li>• Martin Barisits (CERN)</li> <li>• Fabrizio Antonio (CMCC)</li> <li>• Florian Goldenberg (TU Wien)</li> <li>• Robert Harb (MUG)</li> <li>• Thomas Ulich (EISCAT)</li> <li>• Teresa Zulueta-Coarasa (EMBL)</li> </ul>		
Reviewers	<ul style="list-style-type: none"> <li>• Andrea Manzi (EGI)</li> <li>• Carl-Fredrik Enell (EISCAT)</li> <li>• Gergely Sipos (EGI)</li> </ul>		
Moderated by:	<ul style="list-style-type: none"> <li>• Matteo Agati (EGI)</li> </ul>		
Approved by:	Technical Coordination Board		



Terminology / Acronyms	
Term/Acronym	Definition
Access Token	Credential used to authenticate and authorize access to protected resources in a system.
AAI	Authentication and Authorization Infrastructure – a framework for secure user identity management and access control across federated systems
AI	Artificial Intelligence
AIFS	Artificial Intelligence Forecasting System
API	Application Programming Interface
ASC	Austrian Scientific Computing
BBMRI	Biobanking and Biomolecular Resources Research Infrastructure
BIA	BioImage Archive
CEDA	Centre for Environmental Data Analysis
CERN	European Organization for Nuclear Research
CMIP	Coupled Model Intercomparison Project
CORDEX	Coordinated Regional Climate Downscaling Experiment
CSC	CSC – IT Center for Science Ltd
DD	Data Discovery
DEP	Data Exploitation Platform
DestinE	Destination Earth
DID	Rucio Data Identifier
DICOM	Digital Imaging and Communications in Medicine
DLM	Data Lifecycle Management
DOI	Digital Object Identifier
DP	Data Popularity
DPS	Data Preparation Service
DUNE	Deep Underground Neutrino Experiment



EISCAT	European Incoherent SCATter
EMPIAR	Electron Microscopy Public Image Archive
EMBL	European Molecular Biology Laboratory
ENES	European Network for Earth System modelling
ERIC	European Research Infrastructure Consortium
ESGF	Earth System Grid Federation – a distributed data infrastructure for climate modeling and Earth system science
EUCAIM	EUropean Federation for CAncer IMages
Euro-Bioimaging	European Research Infrastructure for Imaging Technologies in Biological and Biomedical Sciences
FAIR	Findable, Accessible, Interoperable, Reusable – principles guiding best practices for data management and stewardship
FTS	File Transfer Service – a high-performance system developed by CERN for secure, large-scale data transfers
GridFTP	Grid File Transfer Protocol
HPC	High-Performance Computing – computing environments with high processing power, usually parallel, used for analysis of big data volumes and to run large simulations
IARC	International Agency for Research on Cancer
JPEG	Joint Photographic Experts Group
OAuth2/OIDC	Open Authorization, a protocol for secure access delegation, is currently v2.0 with v2.1 in the draft stage. Open ID Connect, a layer for user authentication on top of OAuth.
PNG	Portable Network Graphics
REMBI	Recommended Metadata for Biological Images
REST	REpresentational State Transfer
RI	Research Infrastructure
Rucio	A data management system designed for large-scale, policy-driven data orchestration across distributed computing environments
S3	Amazon Simple Storage Service
SE	Rucio Storage Element



SKA	Square Kilometer Array
SPE	Secure Processing Environment
SRM	Storage Resource Manager
STAC	SpatioTemporal Asset Catalog – a standard for describing geospatial assets, used in data discovery services
TIFF	Tag Image File Format
TRE	Trusted Research Environment
TÜBİTAK	Türkiye Bilimsel ve Teknolojik Araştırma Kurumu
URL	Uniform Resource Locator
UUID	Universally Unique Identifier
WebDAV	Web-based Distributed Authoring and Versioning
WLCG	Worldwide LHC Computing Grid
WSI	Whole Slide Image



# Table of Contents

<b>Executive Summary.....</b>	<b>9</b>
<b>1. Introduction.....</b>	<b>10</b>
1.1. Scope and Purpose of the Deliverable.....	11
1.2. Structure of the Deliverable.....	11
1.3. Data Exploitation Platform and User Stories.....	12
<b>2. Data Lifecycle Management in the DEP.....</b>	<b>15</b>
2.1. High-level Design Considerations.....	16
2.2. Data Lifecycle Management Interactions.....	17
2.2.1. Data Lifecycle Management: DEP End-User Perspective.....	17
2.2.2. Data Lifecycle Management: DEP Model Developer Perspective.....	17
2.2.3. Data Lifecycle Management: DEP Operator Perspective.....	18
2.3. The DLM Architecture within the DEP.....	18
2.4. Case Study: ENES.....	21
2.5. Case Study: EISCAT Scheduling.....	22
2.6. Case Study: EISCAT Space Debris and Anomaly Detection.....	23
2.7. Case Study: BBMRI.....	24
2.8. Case Study: Euro-Bioimaging.....	24
<b>3. Data Orchestration Specification.....</b>	<b>26</b>
3.1. Rucio.....	26
3.2. FTS.....	27
3.3. Technology Gaps.....	28
<b>4. Data Preparation for Exploitation Specification.....</b>	<b>30</b>
<b>5. Data Discovery and Data Popularity Services Specification.....</b>	<b>32</b>
5.1. ENES Data Popularity Service.....	32
5.2. ENES Data Discovery Service.....	33
5.3. EMBL Data Discovery Service.....	34
5.4. EUCAIM Data Discovery Service.....	34
<b>6. Data Holdings Specification.....</b>	<b>36</b>
<b>7. Computing Sites Specification.....</b>	<b>38</b>
<b>8. Roadmap.....</b>	<b>40</b>
DEP Release 1 (M12, 28/02/2026).....	40
DEP Release 2 (M24, 28/02/2027).....	40
D2.2 – RIs and Data Spaces Integration Experiences (M36, 29/02/2028).....	41
<b>References.....</b>	<b>42</b>
<b>Annex I: WP2 Requirements collected in D5.1.....</b>	<b>43</b>
Functional Requirements.....	43
Non-Functional Requirements.....	52



## List of Figures

- [Figure 1: DEP Functions and co-provisioning Approach](#)
- [Figure 2: High-level Architecture Overview of the DEP](#)
- [Figure 3: DLM Internal Architecture](#)

## List of Tables

- [Table 1: Storyline of the Interaction of a DEP End-User with the DLM layer](#)
- [Table 2: Storyline of the Interaction of a DEP Operator with the DLM Layer](#)





# Executive Summary

This deliverable, D2.1 – DEP Data Management Specification and Roadmap, outlines the architectural design, key components, and implementation roadmap of the data management subsystem underpinning the RI-SCALE Data Exploitation Platform (DEP). The DEP is a scalable environment aimed at enabling Research Infrastructures (RIs) to unlock the full value of their data holdings by supporting secure, AI-driven data analysis across federated computing infrastructures.

The document presents a comprehensive view of how the DEP supports the entire research data lifecycle from ingestion and replication to processing, archiving, and re-use. A central pillar is the Data Lifecycle Management (DLM) subsystem, which orchestrates intelligent, policy-based data transfers using tools such as Rucio and FTS. This enables transparent and efficient movement of large datasets to cloud or HPC environments for advanced analytics and AI applications.

Complementing this, the deliverable describes supporting services including Data Preparation for Exploitation, which harmonises data inputs for downstream workflows; Data Discovery and Popularity Services, which prioritise and facilitate access to relevant datasets; and the integration of data holdings and computing sites, ensuring interoperability with diverse research domains.

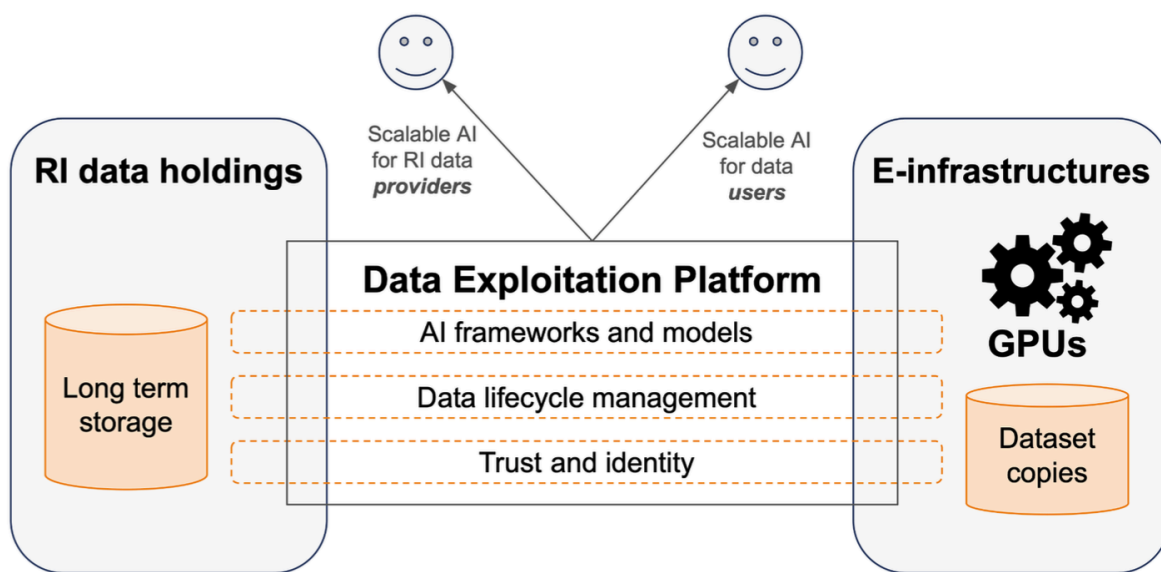
A series of case studies from partner RIs, including ENES, EISCAT, BBMRI, and Euro-BioImaging, illustrates real-world applications, highlighting the DEP's flexibility and impact across environmental and health sciences.

Finally, the roadmap defines the phased delivery of DEP components across project milestones, establishing a clear path from early prototypes to mature, production-ready services. The deliverable thus lays the groundwork for a federated, sustainable data management ecosystem that supports cross-disciplinary innovation and seamless integration with emerging European Data Spaces.



# 1. Introduction

The Data Exploitation Platform (DEP) is being designed as a scalable, open-source platform to enhance the use and value of research data generated by Research Infrastructures (RIs) and Data Spaces. It enables the replication and orchestration of large scientific datasets into cloud and HPC environments, where they can be efficiently analysed using AI frameworks and pre-trained models. The DEP integrates three core components: data lifecycle management, AI-powered computational environments, and secure identity and access management. It supports data interoperability, modular AI toolkits, and energy-efficient operations, making it a central tool for data valorisation and secondary use in environmental and health sciences.



**Figure 1:** DEP Functions and co-provisioning Approach

The RI-SCALE project aims to deliver a comprehensive Data Lifecycle Management (DLM) capability as a core component of the DEP to ensure that RIs can efficiently manage, access, and exploit their data holdings at scale. The objective is to design and implement a data orchestration layer that integrates seamlessly with various RI data repositories and Data Spaces, enabling intelligent, policy-driven data replication, caching, and staging across distributed computing environments.

The DLM functionality will support the entire lifecycle of research data from ingestion and preparation, replication to processing, archiving, and deletion, while ensuring compliance with FAIR principles and optimising the use of compute and storage resources. Key features include support for transparent data transfers, integration with HPC ingress/egress workflows, and the ability to manage data popularity and access frequency through predictive staging and smart caching mechanisms. These features will be implemented using open-source orchestration tools such as Rucio, combined with low-level transfer systems like the File Transfer Service (FTS), ensuring robust and scalable operation across diverse infrastructures.



The DEP's DLM component will also prioritise energy efficiency and resource sustainability. It will monitor and report on data usage, enabling stakeholders to make informed decisions about storage and compute allocation. Additionally, provenance tracking will be embedded to maintain data authenticity and lineage, essential for scientific reproducibility and AI model training. Interoperability with federated Authentication and Authorization Infrastructure (AAI) frameworks will ensure secure and policy-compliant access to sensitive or otherwise restricted or embargoed datasets.

Ultimately, this objective supports the broader RI-SCALE mission to empower RIs with scalable, interoperable, and sustainable infrastructures for data-intensive research and AI-driven innovation. It provides a foundation for data use and seamless integration into emerging European Data Spaces.

## 1.1. Scope and Purpose of the Deliverable

This document provides a comprehensive outline of the design and development plan for the data management components that support the broader goals of the project. Its main purpose is to define the technical specifications, architectural choices, and implementation steps required to enable efficient, secure, and scalable management of research data across distributed infrastructures.

The scope of this deliverable includes:

- Specifying how existing data transfer and orchestration technologies will be used and integrated within the project.
- Planning for alignment with access and identity management systems, including EGI Checkin (WP4), to ensure secure and policy-compliant data access.
- Outlining the approach for integrating data sources from participating infrastructures and external platforms.
- Providing an implementation roadmap that identifies key phases, dependencies, and criteria for successful integration and operation.

This deliverable sets the foundation for the technical work that follows by establishing a clear and shared understanding of the data management needs, technologies, and integration processes. It supports the broader aim of enabling data-driven research and innovation through improved access, interoperability, and usability of scientific datasets.

## 1.2. Structure of the Deliverable

This deliverable is structured to provide a comprehensive overview of the design, specification, and implementation roadmap of the Data Management subsystem within the RI-SCALE Data Exploitation Platform (DEP). It is organised as follows:



- [Section 2](#): Data Lifecycle Management in the DEP – Describes the conceptual foundation, high-level design considerations, and role-specific interactions with the Data Lifecycle Management (DLM) layer. It includes several case studies from participating Research Infrastructures (RIs) to illustrate practical applications.
- [Section 3](#): Data Orchestration Specification – Details the orchestration tools and mechanisms, primarily Rucio [Rucio] and FTS [FTS], used for managing data movement and replication. It also outlines current technology gaps and planned extensions.
- [Section 4](#): Data Preparation for Exploitation Specification – Explains the data preprocessing workflows and modular pipeline design used to ensure compatibility with AI workflows across diverse scientific domains.
- [Section 5](#): Data Discovery and Data Popularity Services Specification – Defines the services supporting dataset findability and replication prioritisation, with examples from ENES, EMBL, and EUCAIM.
- [Section 6](#): Data Holdings Specification – Focuses on onboarding and interfacing with RI data repositories, including compliance with AAI frameworks and support for integration with domain-specific tools.
- [Section 7](#): Computing Sites Specification – Covers the infrastructure requirements and integration of DEP deployments with HPC and cloud resources, including access, storage, and authorisation interfaces.
- [Section 8](#): Roadmap – Provides a phased development timeline aligned with project milestones (M12, M24, M36), identifying key deliverables and integration objectives.
- [References](#) and [Annexes](#) – Include supporting documentation, citations, and a detailed breakdown of requirements derived from deliverable [D5.1](#).

## 1.3. Data Exploitation Platform and User Stories

The DEPs will enable Research Infrastructure (RI) data holdings to expand their services with online data staging and AI/machine learning based analysis and data mining by partnering with external compute facilities where the RI data is replicated and served for user analysis.

RI challenges and limitations that DEP is designed to address are:

1. **Lack of on-site compute:** Limited compute and storage provisioning at RI data holding sites hinders data quality control, data product improvement (incl. FAIR-ification) and the widespread uptake of data by the broader user community for analysis.
2. **Large data downloads and data management:** Large datasets are cumbersome and time-consuming to download for an individual researcher; moreover, separate storage and compute systems may use different access control mechanisms.



3. **Complex software:** Installing and configuring software stacks for running environments for data science (e.g., with AI, Digital twins, Trusted Research Environments (TREs) or Secure Processing Environments (SPEs)) presents a major barrier to users.

A DEP fundamentally acts as an extension of an RI, connecting to its data holdings and offering online data analytics services. In the RI-SCALE project, the DEPs will be specifically designed for the support of AI-based analysis.

The main connections and main users of a DEP:

1. **End users** of a DEP are the main beneficiaries. They want to discover relevant datasets from an RI to perform data analysis, replicate these data from the holdings to the compute facility that operates the DEP environment, and choose a pre-configured, pre-trained AI model to analyse the data with inference runs. Typically, end users have direct access to the DEP, and they are authorised RI users who run the models on data and share the outputs with other authorised users.
2. **Model developers** create and deploy new AI models within the DEP. They either use off-the-shelf 3rd party models or develop their own models, then train the models with RI data, and share the validated models via the DEP with the end users.
3. **DEP operators** deploy, configure and operate the DEP environment within a compute centre, and ensure its proper connections to external systems. These connections include links to the data holding(s), to AI-model stores and to identity management systems that are supported by the specific DEP installation.

Based on the project objectives main goals of the DEP are:

- Replicate and manage copies of big scientific data from RI repositories and Data Spaces on high-performance and cloud compute resources.
- Facilitate the use of AI applications for scalable data analysis.
- Support real scientific use cases with big scientific data and AI applications.
- Enable seamless user access to resources and services across the entire value chain.
- Track and report resource and service consumption during the entire usage workflow.
- Increase the AI-based data exploitation and data mining capacity and resources of the RIs.

Requirements defining the project deliverable D5.1<sup>1</sup> set numerous principles for the DEP, which are linked with these goals, but also high-level activities which identify the 3 user groups described above are expected to interact with the DEP.

These activities for the End user are:

---

<sup>1</sup> Psychas, A., Spiliotopoulou, A., Tenhunen, V., & Sipos, G. (2025). RI-SCALE\_D5.1 – Data Exploitation Platform Requirements and Design Considerations (V1\_Under EC Review). Zenodo: <https://doi.org/10.5281/zenodo.15755803>



- Discover relevant datasets from a research infrastructure or data space.
- Choose a pre-configured and pre-trained AI model to analyse the data.
- Perform data analysis based on the RI data.
- Flag datasets for analysis in the processing environment.
- Discover pre-configured and pre-trained AI model to analyse the data.
- Export the results of the data analysis.

Activities for Model developers are of two types. Firstly, activities which are linked with new AI model development and secondly, activities with existing models:

Activities with new models:

- Create a new AI model;
- Deploy the new AI model for training;
- Train a new AI model with RI data;
- Validate model accuracy;
- Share the validated model in one or multiple DEP(s).

Activities with existing models:

- Select an existing model for retraining;
- Associate the existing model and the training data;
- Train an existing or 3rd party AI model with the data;
- Validate an existing model's accuracy;
- Share the validated old model in one or multiple DEP(s).

The third identified user of the DEPs is the DEP operator. This role is responsible for following activities:

- Deploy, configure and operate the DEP environment within the compute centre.
- Ensure DEP environment connection to external systems (AAI, AI model stores, data holdings, etc.).
- Ensure DEP's infrastructure availability and continuity.
- Manage DEP infrastructure incidents and service requests.
- Ensure infrastructure capacity for DEPs.
- Report DEP resource usage.



## 2. Data Lifecycle Management in the DEP

The Data Lifecycle Management (DLM) is a central pillar of the RI-SCALE project architecture, enabling Research Infrastructures (RIs) to manage, share, and exploit their data holdings in a scalable, secure, and interoperable manner. Within the context of the Data Exploitation Platforms (DEPs), DLM provides the essential capabilities needed to support the full journey of scientific data from ingestion and replication to processing, access, and archiving.

In the RI-SCALE architecture, the DLM is implemented through a **data orchestration layer** that connects RI data repositories with compute environments (cloud, HPC, EuroHPC), forming a seamless pipeline between data production and AI-enabled analysis. This subsystem ensures that data is transferred efficiently using robust technologies such as **Rucio** for orchestration and **FTS** for high-throughput transfers, while maintaining provenance, integrity, and energy-efficient operations throughout.

Key functionalities of the DLM in the project include:

- **On-demand data replication** from RI holdings to compute facilities, triggered by specific use cases or AI workflows.
- **Caching and staging mechanisms** to optimise data locality and minimise data movement costs.
- **Provenance tracking and metadata integration** to ensure data authenticity and reproducibility.
- **Policy-based data retention and decommissioning**, minimising storage use and enabling sustainable data lifecycle operations in line with FAIR and green computing principles.

Within the broader architecture, the DLM interacts with two other core subsystems:

1. **AI Frameworks and Applications** – where data replicated through the DLM feeds into training, inference, and analytics pipelines powered by community-specific and foundational AI models.
2. **Trust and Access Management** – which ensures that data movement and access respect RI-specific or other institutional, disciplinary, and sensitivity-related data access and ownership policies through federated authentication and fine-grained authorisation controls.

By enabling RIs to externalise their data to trusted DEP installations and manage it throughout its analytical lifecycle, the DLM transforms static data holdings into active assets. It supports secondary data use, cross-RIs collaboration, and innovation by lowering the technical barriers to accessing and processing large-scale, heterogeneous datasets.



## 2.1. High-level Design Considerations

The design of the DLM subsystem in the DEP is driven by the need to handle scientific data at scale: efficiently, securely, and sustainably. As a core architectural element of the DEP, the DLM must enable seamless integration between diverse Research Infrastructure (RI) data holdings and distributed compute environments while ensuring interoperability, performance, and compliance with FAIR and energy-efficiency principles.

Key high-level design considerations include:

### 1. Interoperability and Standards Compliance

The DLM must interface with a heterogeneous landscape of data repositories, file formats, and metadata schemas used across RIs and data spaces. Therefore, the design emphasizes standards-based integration using open protocols and APIs (e.g. HTTP, WebDAV, HTTP-REST) and compatibility with FAIR principles to ensure findability and reusability of data across domains.

### 2. Security and Access Control

Sensitive or restricted datasets, particularly from health sciences, require DLM to enforce federated identity and access management controls. Fine-grained authorisation policies ensure that data access aligns with disciplinary, institutional, and national policies, while supporting secure data staging across federated infrastructures.

### 3. Scalability and Extensibility

The system must handle data transfers and orchestration for datasets ranging from terabytes to petabytes. It should support horizontal scaling for ingestion, caching, and delivery across infrastructures, including national clouds, EuroHPC facilities, and commercial platforms.

### 4. Policy-Driven Automation

DLM processes should be driven by programmable policies for data replication, retention, staging, and decommissioning. This enables use-case-specific workflows (e.g. prioritised AI model training, time-bound access, or frequent reuse) while reducing the operational overhead on RI operators.

### 5. Energy Efficiency and Sustainability

Given the energy footprint of large-scale data movement, the DLM design incorporates energy-aware operations minimising data transfers, enabling smart data locality decisions, and integrating green metrics into usage accounting. These principles support RI-SCALE's broader sustainability goals.

### 6. Provenance, Auditability, and Data Integrity

End-to-end data tracking is essential for scientific reproducibility and trust. The DLM system must record provenance information, lineage, and versioning details throughout the data lifecycle.





Integrated checksum validation and audit logs ensure data integrity and traceability across the platform.

## 2.2. Data Lifecycle Management Interactions

### 2.2.1. Data Lifecycle Management: DEP End-User Perspective

**Table 1:** Storyline of the Interaction of a DEP End-User with the DLM layer

Sequence of Interactions Between a DEP End-User and the DLM Layer		
Steps	User Action	Interaction with the DLM Layer
1	The DEP End-User discovers relevant datasets from a research environment or data space.	The DEP End-User authenticates with the AAI and receives credentials, which authorises the discovery request; The DLM Layer verifies the location of the data (Is it already replicated to the processing environment)
2	The DEP End-User flags datasets for analysis in the processing environment.	The DLM layer orchestrates the replication of the data to the e-infrastructure, based on the access tokens given by the DEP's AAI.
3	The DEP End-User discovers a pre-configured and pre-trained AI model to analyse the data.	No interaction of the DLM layer needed;
4	The DEP End-User performs data analysis based on the RI data.	The processing environment stages the data, based on the access tokens given by the DEP's AAI, through the DEP's DLM layer into the processing environment;
5	The DEP End-User retrieves the exported results of the data analysis.	Result data is written to the e-infrastructures output storage, based on the access token given by the DEP AAI, through the DEP DLM layer; Subsequently, based on the policies set in the DEP, data is replicated back to the RI data stores;

### 2.2.2. Data Lifecycle Management: DEP Model Developer Perspective

Creation and sharing of models largely depends on whether they are directly accessible in an RI repository or data space, or stored on a different platform within or external to the RI (e.g. a git repository). If the former is the case, the sequence is similar to [Section 2.2.1](#); if the latter is the case, the DLM layer is not involved in handling the model.



### 2.2.3. Data Lifecycle Management: DEP Operator Perspective

**Table 2:** Storyline of the Interaction of a DEP Operator with the DLM Layer

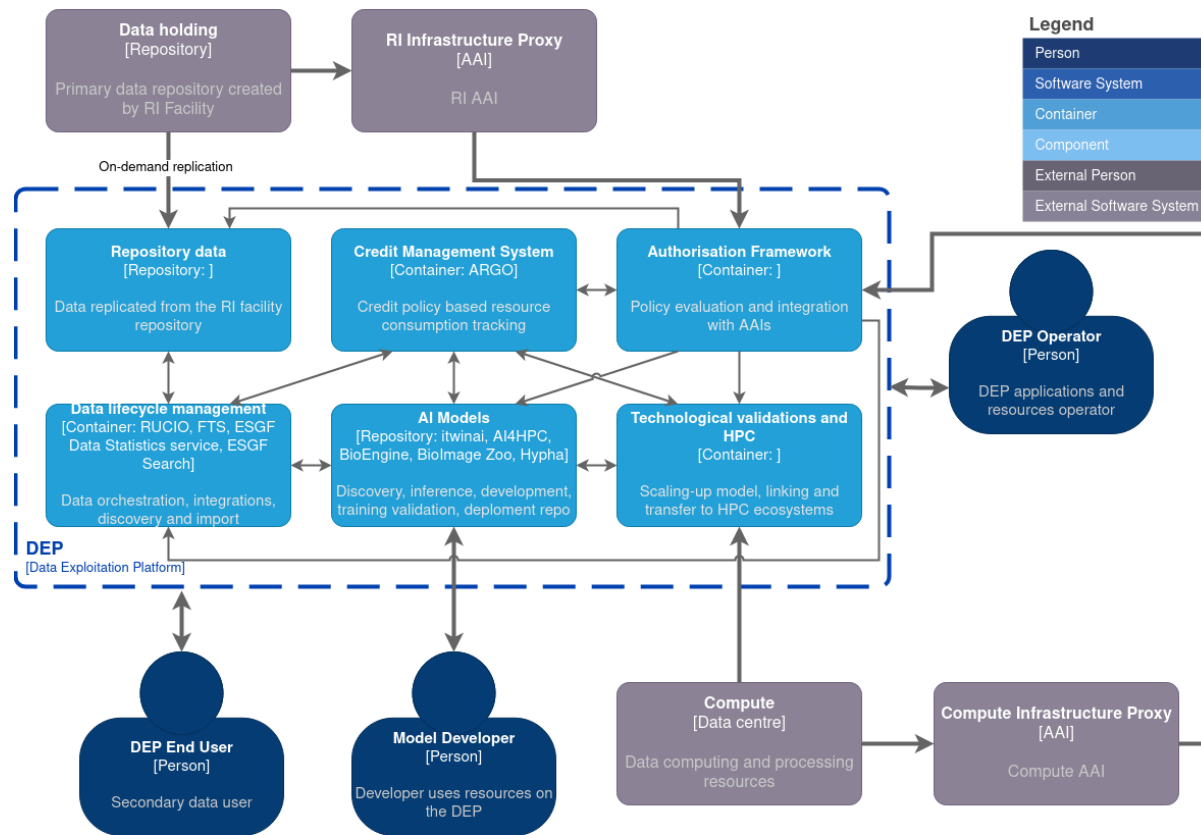
Sequence of Interactions Between a DEP Operator and the DLM Layer		
Steps	User Action	Interaction with the DLM Layer
1	The DEP Operator deploys, configures, and operates the DEP environment within the compute centre.	The DEP DLM orchestration service is deployed and configured at the compute centre.
2	The DEP Operator ensures the DEP environment's connection to external systems (AAI, AI model stores, data holdings, etc.).	The DEP DLM orchestration service is configured and connected to the relevant AAI services; The RI repositories and data holding access mechanisms (URLs, protocols, etc.) are configured in the orchestration service;
3	The DEP Operator reports DEP resource usage.	Resource usage related to the DLM, such as bytes read/written/transferred are reported;
4	The DEP Operator ensures DEP infrastructure availability and continuity.	The DLM layer reports its activity to the monitoring framework, which allows ensuring the continuity of the service;
5	The DEP Operator manages DEP infrastructure incidents and service requests.	Incident response involves analysing the logs and configurations as well as possibly updating the configuration of the DLM layer and its associated components;
6	The DEP Operator ensures infrastructure capacity for DEPs.	Operator needs to monitor and update capacity settings;

## 2.3. The DLM Architecture within the DEP

The Data Exploitation Platform (DEP) is a scalable, open-source environment that enables Research Infrastructures to replicate, manage, and process large-scale datasets within cloud and HPC systems



for AI-driven analysis. It combines data lifecycle management, AI frameworks, and secure access controls to transform static data holdings into actionable, interoperable, and FAIR-compliant research assets.



**Figure 2:** High-level Architecture Overview of the DEP

The architecture, shown in [Figure 2](#), for Work Packages 2, 3, and 4 is designed as a modular, interoperable framework that enables Research Infrastructures (RIs) to seamlessly manage, process, and exploit large-scale datasets within the Data Exploitation Platform (DEP) ecosystem. Each work package contributes a distinct layer to this architecture, ensuring a coherent and end-to-end data-to-insight pipeline.

- WP2** delivers the data lifecycle management and data orchestration capabilities that form the backbone of the DEP. This layer connects RI data holdings and external Data Spaces to distributed compute environments (cloud, HPC, EuroHPC) using technologies such as Rucio for policy-driven data orchestration and FTS for high-performance transfers. WP2 handles dataset registration, replication, staging, and metadata alignment, ensuring interoperability with various repository types and file formats. It also integrates with intelligent caching mechanisms to optimise data locality and with the credit-based resource accounting system for quota enforcement. The WP2 layer is the data pipeline enabler for AI-driven workflows in WP3, operating under the access policies defined in WP4.



- **WP3** builds on the data orchestration services from WP2 to deliver AI-ready computational environments. It integrates AI frameworks, toolkits, and pre-trained models into the DEP, enabling both reuse and training of AI models on large, multi-modal datasets. This layer supports scalable execution on heterogeneous infrastructures (CPU, GPU) using containerised environments (e.g., Kubernetes), ensuring portability and reproducibility. WP3 also handles workflow definition (e.g., Jupyter/Elyra-based interfaces), model provenance tracking, and integration of domain-specific AI tools from partner projects. It directly consumes the datasets staged by WP2, feeding them into AI pipelines for inference, training, and analytics.
- **WP4** provides the authentication, authorisation, and accounting (AAI) framework for secure and policy-compliant access to both data and computational resources across the DEP. This includes integrating federated identity systems, fine-grained policy-based authorisation, and privacy-preserving access controls for sensitive datasets (e.g., in health sciences). WP4 also manages resource accounting and credit-based allocation, tracking usage across data transfers, compute cycles, and AI services. This layer interfaces directly with WP2 to control data access and with WP3 to manage computational resource usage, ensuring consistent enforcement of institutional, national, and disciplinary regulations.

In the combined architecture, WP2 acts as the data ingress and orchestration layer, WP3 is the data exploitation and AI processing layer, and WP4 forms the security, trust, and policy backbone that spans the entire system. The modular design allows each work package to evolve independently while maintaining standardised interfaces for integration. Together, they deliver a scalable, secure, and AI-ready platform that enables RIs to unlock the full potential of their datasets through advanced analytics and cross-domain interoperability.

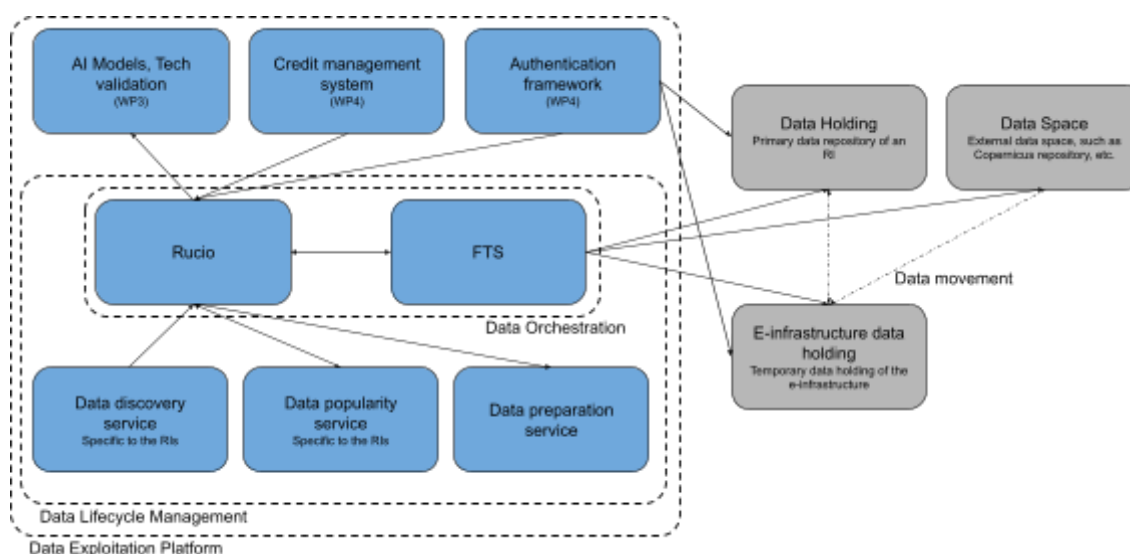


Figure 3: DLM Internal Architecture

The architecture shown in [Figure 3](#) shows the internal architecture of the DLM layer:



- When DEP users interact with the DLM layer, they will typically interact with Rucio (data orchestration), the data discovery service (of their RI), or the data preparation service.
- Rucio ([Section 3](#)) catalogs all data in the DEP, their location, and triggers, and monitors data movement between the Data Holding, Data Spaces, and the temporary data holding at an e-infrastructure.
- FTS ([Section 3](#)) executes the data movement between these data holdings.
- The data discovery service ([Section 4](#)) is an integration of the RI-specific data discovery system to the DEP. It is able to interlink data from multiple repositories with the data in the DEP, and as needed, trigger ingestion of the data to the DEP.
- The data preparation service executes user-demand data preprocessing once data has been successfully replicated to the e-infrastructure.
- Several DLM components will interact with the Authentication Framework (WP4), the Credit management system (WP4), as well as the AI Model execution (WP3)

## 2.4. Case Study: ENES

The European Network for Earth System Modelling (**ENES**) plays a central role in coordinating climate modeling efforts across Europe and delivering a distributed e-infrastructure to support the Earth System science community. Indeed, Earth System Models are becoming more and more complex and high-resolution, producing large volumes of complex, heterogeneous and multidimensional simulated data. This poses significant data management challenges in terms of data sharing, processing, analysis, visualization, preservation, curation, and archiving.

In this domain, community efforts like the Coupled Model Intercomparison Project (CMIP) represent very relevant large-scale global experiments initiatives for climate change research, which have led to the development of the Earth System Grid Federation (ESGF)[ESGF], one of the largest collaborative data efforts in Earth system science. More specifically, ESGF provides a federated data infrastructure for the Earth system modelling community involving a large set of data providers and modelling centres around the globe. By delivering the first-ever decentralized database for accessing geophysical data at dozens of federated sites, ESGF tries to address several key challenges in managing and distributing large-scale Earth system and climate data, thus providing a vital backbone for global climate research and enabling open data sharing and collaboration across scientific communities. The main challenges tackled by ESGF include:

- **Data Volume and Scalability:** ESGF stores, indexes and distributes a massive volume of climate data, including simulations, observations and reanalysis, from several Climate modeling projects across a distributed network of autonomous data nodes united by common protocols and interfaces.



- **Search and Discovery:** ESGF offers metadata cataloging, search and discovery services, allowing users to efficiently locate and access relevant datasets across the vast and globally distributed ESGF data infrastructure.
- **Monitoring and Usage Statistics:** Understanding how data is used across the federation helps optimize storage, replication, and funding. Moreover, as highlighted in Juckes et al. [JUCKES], getting a view about the most heavily utilized climate variables from the CMIP archives proves to have high utility for the evaluation and exploitation of climate simulations, thus helping spread awareness of the scope and impact of the climate variable metadata.

Overall, the integration of some of the ESGF core components into the DEP could have a significant impact on the platform's capabilities related to storing, managing, accessing, and analyzing large-scale climate and Earth system data. Further details are provided in [Section 5](#), which sets out how ESGF services can empower the DEP in terms of efficient data search and discovery, as well as intelligent data replication and caching.

## 2.5. Case Study: EISCAT Scheduling

EISCAT AB operates high-power incoherent scatter radars in Northern Fenno-Scandinavia and on Svalbard, which provide detailed information on the atmosphere and ionosphere from 70 km altitude upwards. Currently, the new tri-static, phased-array EISCAT\_3D radar is being deployed, which will be fully remotely controlled.

When researchers request radar time, they specify what kind of operations they want and the required environmental conditions. EISCAT will then prioritise the requests or decide not to run any.

The specification of radar operations consists of information such as beam pointing direction (azimuth, elevation), range extent, range resolution, timing, and, for multi-beam experiments, the schedule and order of beams to cycle through. For tri-static experiments involving the two remote receiver sites, the users also specify the altitude resolution for wind velocity vectors.

The environmental conditions for a radar experiment include space weather parameters such as solar wind density and velocity, direction of the interplanetary magnetic field, solar activity, geomagnetic activity, as well as cloudiness, and the elevation of the Sun, as well as elevation and phase of the Moon, which are important to define light, twilight, and dark conditions.

The data management challenges for this case study are:

- The definition of the radar experiment description, i.e. the radar experiment metadata, which needs to be collected into a convenient format from the EISCAT experiment database. This is straightforward, because all necessary tools are available to EISCAT and the metadata are public.



- Real-time information about space weather and atmospheric conditions needs to be defined, which includes which information is needed and where it can be obtained from.
- Once the source of the relevant data is identified, the challenge is to interface with a number of different data streams from different providers with different formats.
- Before accessing third-party data, the access policies of the data providers need to be checked for compatibility with transfer to the DEP, and discussions should be held with the originators of these data.

## 2.6. Case Study: EISCAT Space Debris and Anomaly Detection

EISCAT AB has operated incoherent scatter radars in Northern Fenno-Scandinavia since 1981, and thereby accumulated a vast archive of near-Earth space observations. Even today, new phenomena are found in the data, which earlier were disregarded or misinterpreted. Furthermore, statistical studies of the occurrence of particular phenomena are, in practice, a lot of work of manually browsing quick-look plots.

With the EISCAT use case, we want to investigate how AI methods can be used to find specific events as well as anomalies, i.e. rare events or disturbances. We hope to find phenomena in the data which have been missed in previous analyses, or which have possibly been wrongly categorised. Furthermore, we would like to see statistical information about the observations. A set of AI models developed using existing data will then be applied to future radar data to automatically flag up and in real-time phenomena and events as they happen.

The data management challenges for this case are

- Analysing existing data and preparing it to be used for training and preliminary evaluation purposes.
- Interfacing with the EISCAT data archive in a secure manner for analysing the existing data.
- Interpretation of the EISCAT experiment metadata.
- Adherence to EISCAT data security requirements.
- Incorporation of the EISCAT data embargo policies.
- Adaptation of a future stream from the new EISCAT\_3D radar for real-time analysis.
- Consolidation of data formats between historical data and new EISCAT\_3D data.



## 2.7. Case Study: BBMRI

The Biobanking and BioMolecular Resources Research Infrastructure – European Research Infrastructure Consortium (BBMRI-ERIC) is Europe’s reference infrastructure for human biospecimens and associated data. Across 23 Member and Observer countries and one international organisation (IARC), it connects more than 600 biobanks that collectively hold tens of millions of samples, clinical records, and growing volumes of digital pathology. BBMRI-ERIC provides harmonised quality-management guidelines, a public Directory for cohort discovery, and tools such as the Negotiator and Code of Conduct to ensure that access to sensitive medical data complies with European and national legislation. Its Directory already indexes well over 100 million biospecimens, offering both a user-friendly search portal and a REST API for automated queries. For the RI-SCALE use cases, the integration of the BBMRI-ERIC colorectal cancer cohort into the DEP is planned, establishing a reference implementation that can be easily extended to other cohorts for potential future research projects.

To initiate the transfer from BBMRI research infrastructure nodes to the DEP, a request is sent to each node that lists the unique slide identifiers (e.g., UUIDs) along with an optional manifest specifying which additional metadata (e.g., case IDs, patient survival rates, or image annotations) should accompany the images. The DEP data-orchestration layer then uses Rucio and FTS to replicate the requested slides, typically several hundred terabytes, into DEP storage. Once received, every file enters a data-preparation pipeline: proprietary scanner formats are converted to DICOM, metadata are harmonised, and, if requested, JPEG 2000 compression is applied to reduce storage requirements. Provenance is logged throughout, and access to both raw and processed data is controlled by the DEP policy-based AAI layer, ensuring that only authorised users can view or analyse the sensitive pathology material. A mirror workflow operates in the opposite direction for result dissemination: algorithm outputs, such as attention heat maps, predicted survival-risk scores, or synthetic WSIs, are packaged and automatically transferred back to the originating BBMRI node, where they can be examined locally and stored for re-use in future projects.

## 2.8. Case Study: Euro-Bioimaging

Euro-BioImaging ERIC is a fully distributed research infrastructure that provides open access to advanced biological and biomedical imaging technologies, along with training and data services. To support long-term data storage and FAIR sharing, Euro-BioImaging encourages its users to deposit their data in open repositories such as the BioImage Archive (BIA)<sup>2</sup> or EMPIAR<sup>3</sup>. However, understanding, categorising and providing access to large cohorts of mixed life sciences imaging data requires computational understanding of those images. We will use RI-SCALE’s DEPs to develop and train foundation models that capture domain understanding of heterogeneous image data.

<sup>2</sup> <https://www.ebi.ac.uk/bioimage-archive/>

<sup>3</sup> <https://www.ebi.ac.uk/empiar>



Large-scale foundational model inference in the DEPs will produce embeddings which support multiple downstream use cases, such as categorisation, similarity search and derived measurements.

To support this use case, the DEP should provide a robust process for orchestrating the transfer of biological images from the BIA to the DEP. Importantly, the BIA team should be able to define which images are to be transferred using, for example, UUIDs. Together with the image data, associated metadata should be transferred to the DEPs. The BIA team should be able to update the metadata or add more images to a dataset without the need to retransfer the data already staged in the DEP. To ensure that the images are ready to be used with the foundational model developed for this use case, data conversion to appropriate formats (e.g. PNG, TIFF) may be needed as part of the data orchestration process. After analysis, a mechanism that allows the model outputs (such as annotations and associated metadata) to be automatically transferred back to the BIA will be needed. It will be necessary to have a web interface that allows the BIA team to verify the successful transfer of data and metadata from the BIA to the DEP, and vice versa.



## 3. Data Orchestration Specification

Data orchestration within the Data Exploitation Platform (DEP) serves as the intelligent control layer that automates and manages the movement, access, and readiness of research data across federated infrastructures. Its primary function begins with the registration of datasets from Research Infrastructure (RI) holdings into the DEP system, where both the data and associated metadata are catalogued and made discoverable. Once registered, the orchestration engine coordinates the replication of data to the appropriate storage for the HPC infrastructure, where the processing will take place based on user demand or workflow requirements. This includes not only outward replication but also the ability to return results or processed datasets back to the original RI holdings when needed.

To ensure secure and policy-compliant access, the orchestration system interacts dynamically with Authentication and Authorization Infrastructures (AAI), developed in WP 4, validating user credentials and enforcing access rights based on sensitivity, institutional affiliation, and project-specific policies. The orchestrated datasets, along with their metadata, are delivered directly into AI-ready environments, feeding training pipelines, inference tasks, or data analytics workflows.

The data orchestration layer will be built using two existing software systems called Rucio and FTS, introduced in Sections [3.1](#) and [3.2](#). [Section 3.3](#) outlines identified technology gaps in these two software systems, which have to be addressed for the integration into the DEP.

### 3.1. Rucio

Rucio is a powerful, open-source scientific data management system developed to address the complex challenges of large-scale data handling in distributed computing environments. It originated within the ATLAS experiment at CERN's Large Hadron Collider (LHC), where the need to manage exabytes of data across an international network of storage sites led to the creation of a flexible, scalable, and automated framework. Since its release, Rucio has evolved into a general-purpose solution that supports a wide range of scientific disciplines, from high-energy physics to astronomy, climate science, and genomics.

At its core, Rucio is designed to organize, replicate, transfer, and verify vast amounts of data reliably and efficiently. It uses a policy-driven architecture to automate data placement and lifecycle management. Users interact with Rucio through command-line tools, APIs, or graphical interfaces, specifying *what* data needs to be handled and *how* it should be managed, while Rucio handles *where* and *when* these operations happen. This abstraction simplifies complex workflows and ensures data reliability across heterogeneous environments.

A central concept in Rucio is the Data Identifier (DID), which uniquely refers to individual files, datasets (collections of files), and Rucio containers (groups of datasets). These DIDs enable logical



data organization, versioning, and traceability. By applying replication rules to DIDs, users and administrators can define data placement policies – for example, ensuring that a dataset is available on two continents for redundancy or that a container is removed after a certain time period. Rucio's rule engine enforces these policies automatically by scheduling transfers, validating file integrity via checksums, and monitoring the status of each operation in real time.

Rucio supports a wide range of storage technologies and protocols, such as POSIX, S3, XRootD, GridFTP, WebDAV, HTTP, and SRM, allowing seamless integration with both traditional storage facilities and modern cloud infrastructures. It also supports various transfer tools like FTS and Globus<sup>4</sup>. This broad compatibility ensures that Rucio can be deployed in many different scientific and institutional environments, including hybrid cloud setups.

Security and scalability are built into the system. Rucio supports role-based access control and pluggable authentication, allowing integration with external identity providers (e.g., OAuth2, X.509, or Kerberos). It can scale to billions of files and handle hundreds of millions of transfer operations per month. Its modular architecture means that components such as monitoring, messaging, and analytics can be customized or extended based on the needs of a specific project.

Beyond ATLAS<sup>5</sup>, Rucio has gained traction in several large-scale collaborations, such as the CMS<sup>6</sup> experiment at CERN, the Square Kilometre Array (SKA)<sup>7</sup> in astronomy, and the DUNE<sup>8</sup> experiment in neutrino physics. Its adoption across these diverse domains demonstrates its robustness, flexibility, and ability to generalize beyond its original use case.

In summary, Rucio solves one of the most pressing challenges in modern scientific research: managing and moving massive datasets across a distributed, global computing infrastructure. It enables scientists to focus on analyzing data and producing results, rather than managing storage logistics. Through automation, scalability, and community-driven development, Rucio continues to serve as a backbone for data-intensive science.

## 3.2. FTS

The File Transfer Service (FTS) is a high-performance data movement system developed by CERN to support the large-scale, reliable, and policy-driven transfer of scientific data across distributed computing infrastructures. It plays a crucial role in the Worldwide LHC Computing Grid (WLCG), where massive amounts of data must be moved between data centers to support experiments like ATLAS, CMS, ALICE, and LHCb. FTS is designed to ensure the secure, efficient, and fault-tolerant movement of files between heterogeneous storage systems across the globe.

<sup>4</sup> <https://www.globus.org/data-transfer>

<sup>5</sup> <https://atlas.cern>

<sup>6</sup> <https://cms.cern>

<sup>7</sup> <https://www.skao.int/>

<sup>8</sup> <https://www.dunescience.org>



FTS operates as a managed transfer service, handling requests to move files between Storage Elements (SEs) using multiple supported protocols such as GridFTP, WebDAV, HTTP, SRM, and XRootD. Its architecture is asynchronous and event-driven, meaning it can scale to support millions of concurrent file transfers without requiring real-time user interaction. Users and higher-level systems (like Rucio) submit transfer requests through RESTful APIs or command-line clients, and FTS schedules, retries, and monitors the transfers automatically.

One of the key features of FTS is transfer reliability. It handles transient errors with intelligent retry mechanisms and supports features like checksum verification and third-party transfers (where data is moved directly between two storage endpoints without passing through the client). FTS also allows configurable parameters for concurrency, priority, bandwidth limits, and transfer queues, giving administrators fine control over resource usage and policy enforcement.

FTS is designed to be secure and integrates with authentication systems such as X.509 certificates and token-based access, ensuring only authorized users can initiate and manage transfers. It is also monitorable, with real-time metrics, logs, and dashboards that provide insight into transfer success rates, error patterns, and throughput across the network.

FTS is not just a simple data transfer tool - it is a middleware layer that enables automated, large-scale data logistics in distributed research infrastructures. It is typically used by higher-level data management systems such as Rucio, DIRAC, or custom workflows in different scientific communities. The flexibility and reliability of FTS have led to its adoption beyond high-energy physics, including applications in astrophysics and Earth observation, where robust data transfer across international data centers is essential.

In conclusion, FTS is a critical backbone technology that enables global data sharing and movement in data-intensive scientific research. By abstracting the complexity of heterogeneous storage systems and network conditions, FTS ensures that petabyte-scale data can be moved predictably, efficiently, and securely between institutions.

### 3.3. Technology Gaps

Rucio and FTS are mature, production-grade technologies widely used in high-energy physics for large-scale, policy-driven data management and file transfers. Critically, both Rucio and FTS already expose rich APIs that align well with the technical and operational needs of RI-SCALE, including support for multi-endpoint replication, policy enforcement, and monitoring.

However, three integration gaps have been identified that require targeted attention to fully align Rucio and FTS with the advanced requirements of the DEP architecture:

#### **Integration with AAI and Authorization Frameworks (WP4)**

While Rucio and FTS offer support for traditional identity and access mechanisms based on OIDC/oAuth, the RI-SCALE ecosystem, with its different access and security rules, requires a

broader and more flexible federated AAI model, developed under Work Package 4. This includes token-based authorization from multiple token issuers, fine-grained access control policies, and identity federation across diverse RIs and compute infrastructures. Thus, some adaptations to Rucio and FTS need to be done to integrate this slightly more complex AAI model.

#### **Alignment with the Credit-Based Access and Accounting System (WP4)**

RI-SCALE's credit system (WP4) introduces resource usage tracking and quota-based access to data, compute, and services. Currently, Rucio does have support for an internal quota/credit system; however, for Rucio to support this RI-SCALE mechanism, enhancements are needed to interface with an external credit system.

#### **Integration with Intelligent Caching Mechanisms (Task 2.3)**

Task 2.3 focuses on smart, prediction-driven caching of frequently accessed datasets across DEP instances. While Rucio supports data popularity tracking and rule-based replication, the integration with external data popularity services might require the update of some APIs.



## 4. Data Preparation for Exploitation Specification

The Data Preparation for Exploitation Service (DPS) is a user-invoked service that prepares datasets on the DEP for downstream use. After the data orchestration service has copied original RI datasets onto the platform, users launch the DPS to apply the necessary pre-processing steps. Its role is to translate the heterogeneous outputs of participating RIs, ranging from gigapixel whole-slide images (WSI) and multidimensional climate cubes to long-term radar time series of atmospheric parameters, into data objects whose formats, metadata, and directory structure meet the input requirements of the algorithms available on the DEP, while maintaining a consistent, platform-wide convention. This ensures that data ingestion is harmonised across domains and that every dataset can be used directly in workflows without further ad-hoc manipulation.

A manifest that accompanies each transfer specifies the files received and any optional actions (for example, compression or supplementary quality checks); when launching a DPS job, the user selects which of these actions to apply. The DPS then executes the prescribed transformation sequence and records completion for every item.

For every data modality processed by the RI-SCALE use cases, the DPS will provide an appropriate set of data processing operations. Because incoming datasets differ widely in file formats, metadata depth, and requirements for pre-processing, the service does not impose a single canonical representation; instead, for each data modality, the relevant processing operations are selected according to the requirements specified by the respective use cases.

The DPS will thus be implemented modularly. Each data-preparation operation is encapsulated as an independent step that can be enabled, disabled, or reordered in the processing sequence. For a given data orchestration job, the user specifies which steps to run and in what order; one use case may require an extensive transformation pipeline, while another needs only basic validation. Executed preparation steps will be logged to ensure reproducibility. This configurable architecture allows the service to adjust quickly to changing requirements while keeping the overall workflow stable and fully traceable.

To give the DPS a small baseline footprint and the ability to scale out when large transfers need to be processed, it will be containerised. Compute-intensive operations that act independently on individual files, such as image conversion, can therefore be distributed across multiple nodes in an HPC environment. The service will leverage itwinai [interTwin] to dispatch these containerised tasks in parallel, so throughput rises roughly linearly with the number of CPU or GPU cores allocated. This horizontal scaling shortens turnaround times for multi-terabyte ingests, while retaining the option to run a minimal single-node instance when workloads are light.

This consistent, low-friction ingestion is only possible when data providers know up front exactly which formats and structures the platform accepts. Therefore, modality-specific documentation will be provided that details both the input representations supported by the DPS and the canonical forms produced after preparation. For every data type integrated, this documentation will enumerate mandatory metadata fields, optional pre-conversion steps, and the directory hierarchy generated by a successful run. These reference documents supply providers with a concrete compliance checklist and give users a predictable environment in which downstream workflows can operate.



## 5. Data Discovery and Data Popularity Services Specification

Data Popularity (DP) and Data Discovery (DD) services play complementary roles in enabling efficient data exploitation within the DEP. The Data Popularity services provide insights into the most frequently accessed datasets, helping to prioritize resources and guide data handling strategies. The Data Discovery services allow searching, exploring and locating datasets exposed by research infrastructures across multiple data holdings or FAIR data repositories. Different discovery services will be integrated to serve the needs of each specific RI and initiative supported by the project: ENES, Euro-Biolmaging and the EUCAIM cancer images data space. These services might need proper adaptations and extensions to be suitably integrated with Rucio, which acts as the data orchestration layer of the DEP. Indeed, based on popularity information and discovery queries, Rucio is triggered to download and manage the movement of selected datasets from the specific repositories and data sources to the DEP, ensuring that relevant data is made available in a timely and efficient manner.

### 5.1. ENES Data Popularity Service

One of the core components of the ENES Research Infrastructure is the ESGF Data Statistics service [ESGFDataStatistics], which takes care of collecting, analyzing, and reporting a comprehensive set of data usage metrics and data archive information across the ESGF infrastructure.

ESGF consists of a federation of autonomous data nodes, distributed across several countries and united by common standards, protocols and interfaces. Data, including simulations, observations and reanalysis, is hosted at multiple sites worldwide and served through local data and metadata services. Therefore, monitoring this large distributed infrastructure has become a very challenging topic over the years.

The ESGF Data Statistics service represents a key component responsible for gathering logs from the ESGF nodes and processing them to produce a set of heterogeneous metrics both at single site and federation level, thus offering a more user-oriented perspective on the scientific experiments. Examples of metrics include the number of data downloads, the volume of data transferred, the popularity of specific datasets (including information about the most requested climate variables, experiments, models) and projects (e.g., CMIP5, CMIP6, CORDEX) and the temporal trend in data usage. All these metrics are made available through the Data Statistics User Interface<sup>9</sup>), which represents the Community Gateway and provides user-friendly access to aggregated information on how much, how frequently and how intensively the whole federation is being exploited by the end-user.

---

<sup>9</sup> <https://esgf-ui.cmcc.it>





In the context of the project, the ESGF Data Statistics service is being extended for its integration with the DEP. More specifically, it will act as a DP service able to drive the definition of suitable strategies for trusted and intelligent data replication and caching at computing facilities. In this way, the most relevant and requested datasets will be seamlessly transferred from the ENES data holdings to the DEP computing infrastructure, thus enabling data-intensive analysis and the development and delivery of AI applications targeting environmental science. The cache content will be continuously and dynamically updated - through the addition or removal of datasets - based on data usage patterns and popularity metrics derived from the ESGF Data Statistics service. This mechanism will ensure that storage resources are optimally utilized and the most relevant data remains readily accessible for computation and analysis.

## 5.2. ENES Data Discovery Service

The ENES Data Discovery Service is provided by the ESGF core system. In preparation for CMIP7, the core architecture of ESGF has been redeveloped [ESGFNextGen]. The core system manages the publication, editing, and removal of search metadata related to the model's outputs. The new system simplifies and consolidates the original design to a two-node system, one in the US and one in Europe (operated by CEDA, the Centre of Environmental Data Analysis). This aims to free data node effort from maintaining the search service to allow them to focus on the storage of data. To maintain consistency between the two nodes, a shared Kafka event stream will be used. An API service has been developed to receive publication and other events, perform authentication, authorisation, and validation checks, and post to the Kafka event stream. Similarly, this will be run in both locations. For the European service, the authentication and authorisation will use the EGI Check-in service. A second service has been developed to read from the stream and update the search catalogs at both the US and CEDA sites.

The new system has been moved to use STAC (Spatio Temporal Asset Catalog) as the default standard for data discovery across the system. STAC was originally designed by and for the Earth Observation community, but because of its modular design, it was decided that it could be adapted to meet the needs of ESGF. Existing services that interact with the original legacy ESGF search service will need to be migrated to use the new system. But once moved, these services should be easier to maintain as existing STAC packages can be used to reduce ESGF-specific code. The search API now uses the open source implementation of the STAC API specification, which was developed by the STAC community. Like the core specification, this is designed to be extensible, and we have developed to extend the API's functionality to meet the needs of ESGF. Additionally, ESGF services are now able to subscribe to events on the Kafka event stream, which will allow for more responsive services. For example, replication could be triggered by publication events at another data node.



## 5.3. EMBL Data Discovery Service

To allow DEP users to locate and access biological images suitable for transfer from the BioImage Archive (BIA), we will integrate a data catalog to index, describe, and provide search and access capabilities for diverse image datasets. The foundation of the catalog is a robust metadata schema, drawing on the two existing standards implemented on the BIA. We use the REMBI metadata standard [SARKANS] for the image metadata and the MIFA schema [ZULUETA-COARASA] for the metadata of image annotations (such as segmentation masks). These standards ensure rich and consistent descriptions of the datasets, including details on sample preparation, imaging parameters, biological context, annotation creation method, and provenance. Metadata will be stored in a flexible system, such as a document store like Elasticsearch, for fast search. Each dataset on the BIA is assigned a persistent identifier (i.e an accession number and a DOI) to ensure global traceability and interoperability. A user-facing web interface will offer structured and full-text search, and thumbnails of the images will provide a preview of the data for browsing.

## 5.4. EUCAIM Data Discovery Service

The European Cancer Imaging Federation (EUCAIM<sup>10</sup>) is a hybrid infrastructure dedicated to information related to cancer imaging. EUCAIM gathers data holders which expose imaging data through federated nodes or transfer this data to secure processing environments for their analysis.

EUCAIM defines in its user's guide<sup>11</sup> three levels of interoperability that define the data interoperability at the level of the aggregated metadata (tier 1), the searchable data fields (tier 2), and the whole hyperontology (tier 3). Data is discoverable at tier 1 and tier 2 by means of two different applications.

Tier 1 data consists of the aggregated data of the collections and follows the HealthDCAT-AP<sup>12</sup> schema. The metadata of the collections is stored on a catalogue based on Molgenis<sup>13</sup>, which exposes a FAIR Data point interface with the information and the schema<sup>14</sup> in which the information of the collections is publicly available. For example, the metadata of the dataset with the id "05674e3a-596e-40e8-97a5-21846fef19a3" can be retrieved from the URL <http://catalogue.eucaim.cancerimage.eu/Eucaim/api/rdf/Collections?id=05674e3a-596e-40e8-97a5-21846fef19a3>. The full directory is available at the top-level URL. Additionally, data can be discovered through the UI of the catalogue.

Tier 2 data relates to aggregated data that is returned by the federated search from the data that fulfils a specific criterion, based on a set of searchable items. The information returned is the number of studies and subjects for each dataset that have at least one case fulfilling the searching criteria.

<sup>10</sup> [cancerimage.eu](http://cancerimage.eu)

<sup>11</sup> <https://eucaim.gitbook.io/enduserguide/>

<sup>12</sup> <https://healthdcap.github.io/>

<sup>13</sup> [catalogue.eucaim.cancerimage.eu](http://catalogue.eucaim.cancerimage.eu)

<sup>14</sup> <https://catalogue.eucaim.cancerimage.eu/Eucaim/api/rdf/>



The data can be accessed through the UI or programmatically, using the APIs of the spot service<sup>15</sup>. The federated search is based on Samply<sup>16</sup> and works in the form of asynchronous jobs that are dispatched through all the federation nodes, which compute the searching process and return the results, which can be retrieved back through the explorer. First, a POST request must be made, specifying the address of the spot service, the providers from which we wish to obtain data, and a payload that contains: a randomly generated 128-bit unique ID that will identify the task created by the beam proxy, the required query, and the IDs of the data providers.

Secondly, a GET request must be made, specifying the ID used in the previous POST request, in order to retrieve the data requested earlier. The parameter `wait_count=4` should be included so that it waits for the response from the four data providers from which we requested data.

---

<sup>15</sup> [spot.eucaim.cancerimage.eu](http://spot.eucaim.cancerimage.eu)

<sup>16</sup> <https://github.com/samplify>



## 6. Data Holdings Specification

This task focuses on the integration of RI data holdings into the DEPs, enabling seamless access, replication, and orchestration of scientific datasets across federated compute environments. This task plays a foundational role in operationalising the DEP concept, as it ensures that data from diverse disciplines is made interoperable, discoverable, and usable within AI-powered workflows.

The goal of this task is to technically onboard the participating RI data repositories into the DEP ecosystem. This includes establishing interfaces for dataset registration, metadata harvesting, and access endpoint configuration. The task will ensure that holdings are compatible with the DEP data orchestration layer, which leverages Rucio and FTS to perform scalable, policy-driven data transfers.

Key activities in this task include:

- **Data endpoint configuration:** Defining and exposing standardised access points (e.g., HTTP, WebDAV, S3, ...) for each data holding, so that data orchestration tools can interact with them reliably and securely.
- **AAI-compliant access integration:** Ensuring that data access follows the authentication and authorisation frameworks defined in WP4, including support for federated identity and discipline-specific access restrictions (e.g., for sensitive health data or data embargos imposed by RIs).
- **Domain-specific linking/integration:** Integration of the DEP's DLM function stack in domain-specific tools of the science. This should enable users and operators to access functions of the DEP, such as the ability to search for data in the DEP, to demand replication of RI data into the DEP, or request replication of output data out of the DEP, directly within their own community tools.

This needs to be achieved by adapting/configuring the storage endpoints at the data holdings, or by deploying a data service, such as Teapot and ALISE developed in interTwin [interTwin], at the data holdings to offer this functionality.

As part of Task 2.4, special attention is given to integrating external data spaces into the DEP infrastructure, particularly the **Copernicus Climate Data Store**, **Destination Earth (DestinE)**, and the **EUCAIM Data Space Service**. These platforms represent strategic sources of high-value, cross-domain datasets relevant to the project's environmental and health science use cases. Integration efforts will focus on enabling selective ingestion from these sources into the DEP environment. For Copernicus and DestinE, this involves interfacing with APIs and catalogues to discover and retrieve climate simulations and earth observation data, which can be staged alongside RI holdings for joint analysis and AI model training. For EUCAIM, which hosts federated cancer imaging datasets, the task will address secure, AAI-compliant access mechanisms that respect data sensitivity and consent constraints, enabling clinical-grade image processing workflows within



health-oriented DEP instances. These integrations will enrich the data landscape available in the DEPs, enhance the realism and diversity of AI applications, and demonstrate the platform's interoperability with EU-level Data Space initiatives.

By the end of Task 2.4, each of the participating RIs will have one or more of their data holdings technically integrated into the DEP infrastructure. This integration is a prerequisite for enabling the scientific and technological use cases to be developed and validated in later stages of the project. Furthermore, the methods and interfaces developed in this task will be generalised for future onboarding of additional RIs, ensuring long-term scalability and sustainability of the DEP concept.



# 7. Computing Sites Specification

This task constitutes the second large interface to the Data Exploitation Platforms (DEPs), allowing the data from the RI data holdings to be moved to available computing sites in order to process them within the DEP framework. The task is essential to provide the necessary computing power along with the needed storage for the computational tasks to be solved with the established DEPs, including the use cases in the scope of this project (see WP5).

The aim is to seamlessly integrate the DEPs into existing computing infrastructure to allow the data-holding facilities to utilize the full range of computing power of the participating sites without having to leave the respective DEP. This includes data transfer to and from the computing site systems, data lifecycle management that covers the compute part, proper authorisation and access rights, as well as the computing and storage hardware itself.

Data Lifecycle Management 3.1 necessitates the deployment of storage interfaces compatible with the protocols supported by the DEPs, enabling efficient data ingress to and egress from the HPC clusters.

This will be implemented via a Data Management System that can store, manage and retrieve data over several distributed systems using common protocols and interfaces. As described in [Section 3.1](#) an already existing data orchestration solution, Rucio, is going to be used.

The actual data transfer service must facilitate optimisation of the transfer mechanisms to accommodate the large data volume requirements typical of HPC. Again, an existing solution, FTS, is capable of transferring data between different sites and/or users in a reliable and safe way as outlined in [Section 3.2 - FTS](#).

To enable access authentication and authorisation to the storage interfaces, the AAI development in WP4 is required in order to support the diverse nature of the participating RIs, which use several authentication/authorization variants that differ slightly. Standards are not always adhered to, posing possible challenges in implementation.

Storage and computing hardware will be provided by the computing sites. Sufficient storage space, computing resources and computer time for the intended scientific and technical validations must be scheduled. The specifics are up to the individual sites. Differences in handling the requests and operating the sites should be harmonized to the extent possible.

The participating sites are:

- **ASC Research Center**

Two national clusters located in Vienna, with more than 2500 CPUs and more than 1000 GPUs, are provided non-exclusively.

- **TÜBİTAK ULAKBİM**



OpenStack cloud with 20 GPUs and 1250 CPUs, plus a dedicated cluster with GPUs and 160 CPUs. More resources will be allocated to both cloud and cluster sites as the project progresses and hardware allows.

- **Masaryk University**

Managed Kubernetes cluster with 1000 CPUs, 22 GPUs and 4PB of storage, provided non-exclusive, more resources can be added on demand.

By the end of Task 2.5, all institutions holding RI data will have direct access to the necessary storage space and compute power via the DEPs, with data management features available and with data ownership and access rights observed.



## 8. Roadmap

This roadmap outlines the phased development and integration of the data management subsystems over the course of the project. It begins with defining common specifications for data registration, transfer, and metadata harmonisation across participating infrastructures. This is followed by the incremental integration of key components – such as orchestration tools, access control systems, and data holdings – into early DEP prototypes. Subsequent phases focus on performance validation, interoperability testing with external data spaces, and alignment with intelligent caching and credit-based resource management. The roadmap ensures a structured and collaborative path toward a fully functional and scalable data management layer for the DEP ecosystem.

In this section, we give an overview of key developments and milestones which should be achieved, latest, at DEP Release 1 (M12), DEP Release 2 (M24), and at the final report of WP2: D2.2 (M36).

### DEP Release 1 (M12, 28/02/2026)

- T2.1: Prototype deployment of the DEP data orchestration service (Centrally, at CERN)
- T2.2: Deliver minimum viable DPS and validate on one reference dataset.
- T2.3: Preliminary integration of the Data Popularity and Discovery services with the DEP Data Orchestration layer.
- T2.4: Integration of at least one RI repository/data holding with the data orchestration service.
- T2.5: Integration of at least one Compute Site to the data orchestration service.

### DEP Release 2 (M24, 28/02/2027)

- T2.1: Multiple deployments of the DEP data orchestration service, at least one at a e-infrastructure.
- T2.2: Expand capability to all modalities and validate with RIs under real load.
- T2.3: Data lifecycle management and related capabilities upgraded in the DEP of each RI, enabling intelligent data replication and caching capabilities on top of the RI-SCALE computing infrastructure.
- T2.4: The large majority of RI repositories and data holdings are integrated into the data orchestration service.
- T2.5: Integration of all Compute Sites.





## D2.2 – RIs and Data Spaces Integration Experiences (M36, 29/02/2028)

- T2.1: Performance optimization of the DEP data orchestration service, as well as storage access performance
- T2.2: Finalise production DPS, validated across every project use-case in end-to-end tests
- T2.3: Technical and performance validation
- T2.4: Generalisation and performance optimisation for storage site integration
- T2.5: Generalisation of the Computing Site integration to allow for additional sites in the future



# References

Reference	
No	Description/Link
<b>R1</b> (ESGF)	L. Cinquini et al., "The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data", Future Generation Computer Systems, vol. 36, pp. 400–417, 2014. DOI: 10.1016/j.future.2013.07.002
<b>R2</b> (ESGFDataStatistics)	S. Fiore, P. Nassisi, A. Nuzzo, M. Mirto, L. Cinquini, D. Williams, and G. Aloisio, "A Climate Change Community Gateway for Data Usage & Data Archive Metrics across the Earth System Grid Federation", in Proceedings of the 11th International Workshop on Science Gateways (IWSG 2019)", edited by Stankovski, V. and Gesing, S., vol. 2975 of CEUR Workshop Proceedings, p. 6, CEUR, Ljubljana, Slovenia, 12-14 June 2019. URL: <a href="https://ceur-ws.org/Vol-2975/paper5.pdf">https://ceur-ws.org/Vol-2975/paper5.pdf</a>
<b>R3</b> (ESGFNextGen)	R. Evans, D. Poulter, P. Kershaw, I. Foster, R. Ananthkrishnan, F. Hoffman, A. Radhakrishnan, S. Kinderman, S. Ames, and D. Westwood, "ESGF Next Generation and preparations for CMIP7", EGU General Assembly 2025, DOI: 10.5194/egusphere-egu25-19453
<b>R4</b> (FTS)	Long, S., Pleiter, D., Patrascioiu, M., Padrin, C., Carpena, M., More, S., & Carpio, M. (2024). Integrating FTS3 in the Fenix HPC Infrastructure, in EPJ Web of Conferences, 295, 01037 (2024). DOI: 10.1051/epjconf/202429501037.
<b>R5</b> (interTwin)	interTwin: An interdisciplinary Digital Twin Engine (DTE) for Science, a Horizon Europe project coordinated by EGI Foundation (Grant Agreement 101058386), 2022–2025
<b>R6</b>	M. Juckes, K. E. Taylor, F. Antonio, D. Brayshaw, C. Buontempo, J. Cao, P. J. Durack, M. Kawamiya, H. Kim, T. Lovato, C. Mackallah, M. Mizieliński, A. Nuzzo, M. Stockhause, D. Visioni, J. Walton, B. Turner, E. O'Rourke, and B. Dingley, "Baseline Climate Variables for Earth System Modelling", Geosci. Model Dev., 18, 2639–2663, 2025. DOI: 10.5194/gmd-18-2639-2025
<b>R7</b> (Rucio)	Barisits, M. et al. (2019). Rucio: Scientific Data Management. Computing and Software for Big Science, 3(1): 11. DOI: 10.1007/s41781-019-0026-3
<b>R8</b>	Sarkans, U., Chiu, W., Collinson, L. et al. REMBI: Recommended Metadata for Biological Images - enabling reuse of microscopy data in biology. Nat Methods 18, 1418–1422 (2021). <a href="https://doi.org/10.1038/s41592-021-01166-8">https://doi.org/10.1038/s41592-021-01166-8</a>
<b>R9</b>	Zulueta-Coarasa, Teresa, et al. "MIFA: Metadata, Incentives, Formats, and Accessibility guidelines to improve the reuse of AI datasets for bioimage analysis." arXiv preprint arXiv:2311.10443 (2023).



# Annex I: WP2 Requirements collected in D5.1

Deliverable [D5.1 "Data Exploitation Platform Requirements and Design Considerations"](#) collected several functional and non-functional requirements for the DEP. This Annex gives an analysis from a complexity and feasibility point of view for the collected requirements.

## Functional Requirements

Requirement Jira key	Requirement Source	Description	Rationale	Component Fulfil It	That Priority	D2.1 perspective
RSREQ-101	[ITU] Internal – Technical Use Case (Technical UC)	Infrastructure to operate the GROQ card in a testing/proof of concept environment	To test inference on GROQ TPU cards, they need to be run in an HPC environment	Computing Site Integration Interface (WP2-T2.5)	Must	Foreseen in work plan
RSREQ-24	[ISU] Internal – Scientific Use Case (Scientific UC)	<p>The DEP should provide a centrally managed Git server based on GitLab to support version control of training code, configuration files, and related project resources. This Git server should:</p> <ul style="list-style-type: none"> <li>• Be accessible from all computing sites of the DEP.</li> <li>• Be integrated with the DEP's user management system, ensuring that access control (e.g., project-based visibility, permissions, and roles) is centrally managed and aligned with the user's project membership and roles in the DEP.</li> </ul>	Collaborative AI model development within the DEP requires reliable, accessible, and secure version control for code and associated artifacts. A GitLab server provides users with a standard and widely adopted platform for collaborative development. The AI Model Hub needs a centralised repository to collect the code of tracked models. CI support, especially with GPU capabilities, further enhances automation and reproducibility of AI	Computing Site Integration Interface (WP2-T2.5)	Should	Further clarification is needed on which component should fulfil this



		<ul style="list-style-type: none"> <li>• Be suitable for both interactive use (via GitLab web interface) and scripted operations (via Git over SSH/HTTPS and CI/CD API integrations).</li> <li>• Support continuous integration (CI) pipelines for running automated jobs on the DEP compute infrastructure, including support for GPU-enabled CI runners to allow automated testing to be executed as part of version-controlled workflows.</li> </ul>	development. This requirement supports development in Use Cases 5 and 6, and contributes to KPI03: No. of AI frameworks/toolboxes offered within DEPs.			
RSREQ-25	[ISU] Internal – Scientific Use Case (Scientific UC)	<p>The DEP should provide a centrally managed container registry to support container-based workflows across all compute sites. The service should support Apptainer (formerly Singularity) containers. Examples of suitable solutions include Harbor or GitLab Container Registry. The registry should:</p> <ul style="list-style-type: none"> <li>• Be accessible from all DEP compute sites</li> <li>• Allow users to upload, version, and share container images associated with their projects, with access controlled via the DEP's user management system.</li> </ul>	Reliable container management is essential for reproducible, portable, and secure execution of AI workflows in the DEP. The AI Model Hub should use the container registry to execute tracked models. Furthermore, this directly supports collaborative development in use cases 5 and 6.	Computing Site Integration Interface (WP2-T2.5)	Should	Further clarification is needed on which component should fulfil this



		<ul style="list-style-type: none"> <li>Ensure Apptainer containers can be executed on all supported compute infrastructures.</li> </ul>				
RSREQ-38	[ITU] Internal – Technical Use Case (Technical UC)	The DEP needs to be able to copy data onto at least one of the EuroHPC systems that DestinE is currently using, Lumi (CSC), Leonardo (Cineca), MN5 (Bsc)	The technical use case states: Evaluate and measure the scale at which the DEP can work on a EuroHPC machine while serving a challenging AI model with large data from DestinE. Hence, the DEP needs to be able to copy data onto at least one of the EuroHPC systems that DestinE is currently using, Lumi (CSC), Leonardo (Cineca), MN5 (Bsc)	Computing Site Integration Interface (WP2-T2.5)	Must	Access to one of the EuroHPC machines is required; Further technical clarification is needed
RSREQ-39	[ITU] Internal – Technical Use Case (Technical UC)	The DEP needs to be able to spawn virtual kubelets onto at least one of the EuroHPC systems that DestinE is currently using, Lumi (CSC), Leonardo (Cineca), MN5 (Bsc).	The technical use case states: Evaluate and measure the scale at which the DEP can work on a EuroHPC machine while serving a challenging AI model with large data from DestinE. Hence, the DEP needs to be able to spawn virtual kubelets onto at least one of the EuroHPC systems that DestinE is currently using, Lumi (CSC), Leonardo (Cineca), MN5 (Bsc)	Computing Site Integration Interface (WP2-T2.5)	Must	Access to one of the EuroHPC machines is required; Further technical clarification is needed
RSREQ-89	[ITT] Internal – Technical Team	Implement data transfer mechanisms and protocols at the computing sites.	Needed to get data from data owners to the computing infrastructure	Computing Site Integration Interface (WP2-T2.5)	Must	Foreseen in the work plan



RSREQ-90	[ITT] Internal – Technical Team	Implement federated authentication/authorisation protocol at the compute sites.	Need to give data owners secure access to computing sites	Computing Site Integration Interface (WP2-T2.5)	Must	Foreseen in the work plan
RSREQ-106	[ISU] Internal – Scientific Use Case (Scientific UC)	The platform must allow users to download climate datasets through customizable queries from different data platforms, such as those within the Copernicus Climate Data Store or through the ESGF data platform.	This functionality is needed for extracting the data needed for training the ML model and for inference of downscaled data for use case WP3.3.	Data Discovery & Popularity Service (WP2-T2.3)	Must	Data download from ESGF will be supported by the ENES Data Discovery service.
RSREQ-18	[ISU] Internal – Scientific Use Case (Scientific UC)	The platform should support access to usage statistics on ESM data (i.e., CMIP) through customizable queries, allowing also for basic aggregation functions.	This functionality is needed for (i) extracting the data needed for training the ML model and (ii) for feeding the trained model with the data for inference for the use case in WP3.3. It contributes KPI#4 (No. of AI models offered within DEPs), KPI#7 (No. of AI models trained in DEP pilots), KPI#8 (No. of use cases developed for DEP validation).	Data Discovery & Popularity Service (WP2-T2.3)	Should	This will be supported by the extended version of the ENES Data Popularity service.
RSREQ-45	[ITT] Internal – Technical Team	The Data Popularity service must provide a way for extracting the most downloaded ESM data (e.g., CMIP) from the ENES research infrastructure.	This functionality is needed for properly driving the data ingestion pipeline and supporting the DEP caching mechanism. It contributes to KPI#6 (Total size of datasets used in the DEP pilots).	Data Discovery & Popularity Service (WP2-T2.3)	Must	This will be supported by the extended version of the ENES Data Popularity service.



RSREQ-100	[ITT] Internal – Technical Team	<p>The storage systems in RI-SCALE:</p> <ul style="list-style-type: none"> <li>• RI Data holdings (long-term storage).</li> <li>• Storage @ E-Infrastructure for dataset access.</li> <li>• Must support OIDC token authentication based on the work prepared in WP4.</li> </ul>	Without that support, the authentication of storage operations will be extremely difficult and fall outside the trust network built in WP4.	Data Holdings Integration Framework (WP2-T2.4)	Must	Essential technical requirement for the data holdings and e-infrastructure. Further technical investigation is needed to judge the final complexity.
RSREQ-21	[ISU] Internal – Scientific Use Case (Scientific UC)	The platform should support access to CMIP climate projections and reanalysis data.	This functionality is needed for extracting the data needed for training the ML model and for inference of downscaled data for use case WP3.3.	Data Holdings Integration Framework (WP2-T2.4)	Must	Further technical investigation is needed to fully judge the complexity.
RSREQ-42	[ITU] Internal – Technical Use Case (Technical UC)	Newly generated DestinE training data should be retrieved to the DEP to make it available for AIFS training runs on at least one of the EuroHPC systems that DestinE is currently using, Lumi (CSC), Leonardo (Cineca), or MN5 (Bsc)	There is also a performance component here, 100 Gb/s roughly between DEP and DestinE data holdings, data space. Hence, DEP should have enough storage available for the AIFS training datasets and transfer 50-100 TB in a few minutes (100 Gb/s) to at least one of the EuroHPC systems that DestinE is currently using, Lumi (CSC), Leonardo (Cineca), or MN5 (Bsc)	Data Holdings Integration Framework (WP2-T2.4)	Must	Further technical investigation is needed to fully judge the complexity.



RSREQ-22	[ISU] Internal – Scientific Use Case (Scientific UC)	<p>The DEP should provide a robust and secure process for orchestrating the transfer of Whole Slide Images (WSIs) from Research Infrastructure (RI) data holdings into the DEP. The entire orchestration should be declaratively defined and user-configurable, while abstracting away the complexity of backend transfer and storage operations. The orchestration process should:</p> <ul style="list-style-type: none"> <li>• Enable RIs to define project-scoped staging rules, specifying which WSIs are to be transferred, using unique identifiers recognised across both the RI and the DEP.</li> <li>• Verify the successful and complete transfer of datasets before making them available to the user.</li> <li>• Support the transfer of associated metadata alongside WSIs, and provide functionality to resynchronise metadata independently, allowing RIs to update or extend metadata (e.g., annotations, genetic data) without requiring the WSIs to be re-transferred.</li> </ul>	Transferring large-scale, sensitive WSI data from RI holdings into the DEP must be handled through a controlled and reliable process. Allowing RIs to define project-level dataset visibility ensures that governance and data access policies are enforced at the source. Automated transfer of WSIs will be used for the scientific use cases 5 and 6.	Data Orchestration Service (WP2-T2.1)	Should	Should already be possible with the current functionality of Rucio and FTS.
----------	------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------	--------	-----------------------------------------------------------------------------





		<ul style="list-style-type: none"> <li>Track which WSIs have already been transferred to the DEP to avoid redundant transfers when multiple projects request the same images.</li> <li>Support the definition of a data lifespan, allowing RIs or project owners to specify an expiry period after which the transferred data should be automatically deleted from the DEP.</li> <li>Provide logging capabilities for all involved stakeholders, ideally accessible via a web interface, to monitor the status of data transfers (e.g., when a transfer was initiated, completed, or failed).</li> </ul>				
RSREQ-23	[ISU] Internal – Scientific Use Case (Scientific UC)	<p>In addition to enabling the transfer of datasets from RIs into the DEP for analysis, the DEP should provide a Results Delivery Mechanism that allows research outputs - such as annotations, processed datasets, and derived results—to be automatically transferred back to the originating RI at a predefined point in time, such as the end of a project. All involved stakeholders should be able to monitor the status of these transfers,</p>	<p>Returning research outputs to the RI ensures that data products generated through DEP-based analysis are preserved, traceable, and available for future reuse. This closes the research data lifecycle and reinforces the role of RIs as long-term stewards of scientific data. It also aligns with the FAIR data principles - ensuring that outputs are</p>	Data Orchestration Service (WP2-T2.1)	Should	Current functionality of Rucio and FTS should already support this.



		e.g., through a web interface. The specific research outputs to be returned should be jointly defined at the beginning of the project or in the usage agreement between the DEP user and the corresponding RI. The Results Delivery Mechanism should then automatically initiate the transfer of outputs at a scheduled date, e.g., project end.	Findable, Accessible, Interoperable, and Reusable.			
RSREQ-91	[ISU] Internal – Scientific Use Case (Scientific UC)	EISCAT L1 and L2 (raw and spectral) data MUST be made accessible, follow and can preferably be filtered among access rules as specified in the other SUC4 Requirements. EISCAT is using EGI Checkin and Perun to manage access authN/authZ. The existing download portal (portal.eiscat.se) is not very suitable for API download, so a WebDAV or similar service to be used with Rucio/FTS SHOULD be deployed. Any used tools MUST implement data access authZ through supported protocols. The existing portal retrieves VO group membership information by OIDC from EGI Checkin- this is the preferred protocol. Analysed parameters (Level 3 data) are free for use, and can be retrieved using the Madrigal web services	Data accessibility	Data Orchestration Service (WP2-T2.1)	Must	Further technical analysis is needed to fully judge complexity; however, current functionality in Rucio/FTS should already support this.



		through a web GUI or client scripts available for Python, Matlab etc.				
RSREQ-92	[ISU] Internal – Scientific Use Case (Scientific UC)	Model outputs must be returned to the S3 buckets of Hypermeteo, so that they can be used to provide climate services.	This functionality is required in order to exploit the data produced by Scientific Use Case 1 (downscaling of climate projections).	Data Orchestration Service (WP2-T2.1)	Must	
RSREQ-33	[ISU] Internal – Scientific Use Case (Scientific UC)	<p>During the data orchestration from RIs to the DEP, datasets must be prepared in a format that ensures they are ready for use in AI workflows and data analysis. This step acts as a bridge between raw data held at RIs and downstream exploitation in the DEP.</p> <p>For instance, this preparation should:</p> <ul style="list-style-type: none"> <li>• Perform data conversion, if needed, e.g., convert WSIs to DICOM format.</li> <li>• Perform pseudo-anonymisation, if required by the RI, such as the removal of patient-identifying metadata.</li> </ul> <p>Use case providers together with RIs should provide documentation specifying the acceptable data and metadata formats for their datasets, such that they are compatible with the DEP.</p>	This requirement ensures that all data entering the DEP is consistently prepared and standardised, regardless of its origin or modality. Standardisation during data preparation is essential for enabling seamless integration with DEP tools, reproducibility of AI workflows, and interoperability across scientific use cases. It also reduces technical overhead for users and developers by ensuring that data is analysis-ready upon arrival.	Data Preparation Module (WP2-T2.2)	Should	Foreseen in the work plan



## Non-Functional Requirements

Requirement Jira key	Requirement Source	Description	Rationale	Component Fulfils It	That	Priority	D2.1 perspective
RSREQ-41	[ITU] Internal – Technical Use Case (Technical UC)	Provide enough storage space and throughput to serve data for AIFS training in a timely manner	The technical use case states: Evaluate and measure the scale at which the DEP can work on a EuroHPC machine while serving a challenging AI model with large data from DestinE. Hence, DEP should have enough storage available for the AIFS training datasets and transfer 50-100 TB in a few minutes (100 Gb/s) to at least one of the EuroHPC systems that DestinE is currently using, Lumi (CSC), Leonardo (Cineca), or MN5 (BSC)	Computing Site Integration Interface (WP2-T2.5)		Must	Access to one of the EuroHPC machines is required; Further technical clarification is needed
RSREQ-46	[ITT] Internal – Technical Team	The Data Popularity service should provide an easy-to-use interface (e.g., UI Dashboard or API-based) to support the extraction of information about the most downloaded data in an easy and quick manner.	The ESGF (Earth System Grid Federation) Data Statistics service continuously collects information about data usage from the ESGF data nodes. Therefore, providing an easy way for querying and retrieving information from the service is key to efficiently and effectively supporting the	Data Discovery & Popularity Service (WP2-T2.3)		Should	This will be the main functionality of the ENES Data Popularity service.



			DEP data lifecycle management.			
--	--	--	--------------------------------	--	--	--