

Fostering Synergies: Joint Collaboration across DataGems, EOSC Data Commons, LUMEN and RAISE Suite

Enol Fernández on behalf of DataGEMS, EOSC Data Commons, Lumen and RAISE Suite projects

The call: HORIZON-INFRA-2024-EOSC-01-05

Innovative and customizable services for EOSC Exchange

Outcomes:

- Next generation EOSC provides researchers with the means to **easily access complete datasets and analysis platforms** and provide services that **support reproducibility**, as well as ensuring **long-term preservation and long-term availability** of these research data and tools.
- Ecosystem of **novel tools and services**, as many new **FAIR-by-design datasets** as possible, whereby researchers are able to deliver much more rapidly the outputs of each part of the research lifecycle, including data and software, with the same level of precision as they deliver publications today.
- **Enhanced services and tools** for researchers to **lower the bar and underpin the initial research planning and preparation phase** (i.e. entry phase) of the research lifecycle.



The Problem: Broken Data Discovery

Data Volume and Diversity

Huge volumes and varied data types complicate effective data discovery and management today.

Limitations of Keyword Search

Current keyword-based portals fail to grasp user intent, yielding irrelevant and superficial results.

Lack of Rich Contextual Metadata

Metadata focuses on provenance, lacking semantic and contextual details, and forcing manual dataset evaluation.

Insufficient Advanced Querying and User Guidance

Absence of cross-dataset analytics and guidance of users restricts discovery of complex, multi-source data.



The DataGEMS Solution

<https://datagems.eu/>

✓ Comprehensive Data Lifecycle

DataGEMS offers a unified architecture for:
sharing • storing • managing • discovering • reusing data efficiently.

✓ Semantic Data Profiling

Powerful profiling generates rich semantic metadata
to validate dataset relevance and interpret semantics accurately.

✓ Next-Generation Data Discovery

Supports natural language queries, complex pattern recognition,
and content-aware dataset ranking for enhanced discovery.

✓ Exploratory analytics & guidance

Guides users by recommending similar datasets,
auto-assembling data packages, and proposing exploration workflows

✓ Support for Diverse Data Types

Supports structured, semi-structured, unstructured, historical,
and real-time datasets.



Project Objectives

Enhanced Dataset Discovery

DataGEMS automates metadata creation and speeds up dataset discovery, improving research efficiency across disciplines.

Diverse Use Cases

Initial applications span **education**, **meteorology**, **lifelong learning**, and **language** data, demonstrating broad domain relevance and impact.

Sustainability and FAIR Compliance

Integration with EOSC guarantees long-term system sustainability and compliance with FAIR data principles.

Stakeholder Engagement and Scaling

Call to action includes pilot participation, EOSC onboarding, and investments to expand and scale DataGEMS adoption.



Weather

Scientists and stakeholders from sectors affected by weather and climate conditions (civil protection, agriculture, construction, tourism, etc.)

[Read more](#)



Language

CLARIN is a digital infrastructure which provides automatic access to a broad range of language data and tools

[Read more](#)



Math

The MathEd platform helps higher education students in their study of Maths.



Lifelong Learning

This use case is concerned with helping organisations find educational offerings that optimally match their workforce

Call topic:

HORIZON-INFRA-2024-EOSC-01-05

Innovative and customizable
services for EOSC Exchange

Budget: 7.000.000€

Start: 01/04/2025

Duration: 36 months

Consortium: 22 partners: **Coordinator:** EGI Foundation

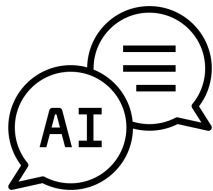
- National/international/thematic repositories
owners: CNRS, SRCE, UZH, KNAW, EODC, TUD,
SIB, FZJ, UH-Luomus, CERN
- Technology providers: SWITCH, UST/AGH, UPV,
UFR, ETH, INFN
- Research communities (brought by): UZH, CERN,
CESNET, CIRMMP, CSIC, UU, CNRS, EODC ...

Contribute to establishing
EOSC as the **European
Research Commons** by
federating repositories from
national, institutional and
thematic initiatives with a
metadata warehouse of paired
data and software for **use and
reuse of datasets** across the
compute continuum



Data Discovery and Access

AI-based natural language search tools
over a rich metadata warehouse

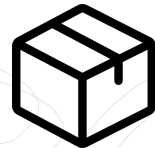


Pairing Data & Analytics

Data & Tools compatibility suggestions
with Automated Orchestration



Package for processing
datasets



8x

Dataset Growth

5492

Published Tools

6

Compute Engines

FAIR Tooling
Initiated

Requirements underway

PoC (Nov'25)

Data Repositories

4 repositories aggregated (<50k datasets):

DABAR, DANS, Onedata, SWISSUBase

Tools Discovery

Published tools count: 1

FAIR Assessment

Not planned for this phase

Compute Engines

1 engine active: Galaxy

First Release (Jun'26)

Data Repositories

8 repositories aggregated (440K+ datasets):

DABAR, DANS, DaSCH, FinBIF, HAL, SWISSUbase
Onedata indirectly onboards EODC, TopAnat, and VIP

Tools Discovery

Published tools count (mostly automatic harvesting): 5492

FAIR Assessment

Started requirement gathering

Compute Engines

Expanded Ecosystem:

Ready: Galaxy, JupyterHub, BinderHub, VIP

In progress: RRP and ScienceMesh

LUMEN OBJECTIVES AND KEY CONTRIBUTIONS

Four strategic objectives to transform how communities discover, share, and reuse research outputs

Objectives

S01

Create a shared and collaborative cross-disciplinary knowledge and **data exchange space** for the next generation EOSC

S02

Enhance the qualitative discovery of research outputs by establishing an ecosystem of **innovative features**

S03

Support the production of **FAIR-by-design research outputs**

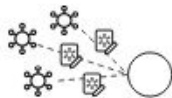
S04

Facilitate early stage of the research lifecycle by **enhancing specific tools and services**

Key contributions to EOSC



White-label discovery platforms



Data Mesh architectural framework



LUMIS: LUMEN Infrastructure for Semantics



AI-enhanced services



Progressive integration into the EOSC Federation

4 SCIENTIFIC DOMAINS AT THE HEART OF LUMEN

LUMEN is a federated collaboration of these scientific communities, leveraging their unique strengths, infrastructures, and expertise.

A fundamental discipline that provides rigorous methodologies and tools for complex problems.

Climate modelling

Cryptography

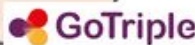
Machine learning

Optimization of supply chains

Simulation of financial markets

Mathematics

Social Sciences and Humanities (SSH)



Focused on exploring the human dimensions of societies, including culture, language, and societal organization.

Analysing social inequalities

Developing public policy frameworks

Examining ethical implications of AI

Studying cultural diversity and multilingualism

Evaluating the impact of technological adoption on societies

A branch of physical sciences that uses numerical simulations to predict molecular and atomic interactions.

Drug design and discovery

Development of advanced materials

Chemical process optimization

Understanding protein folding mechanisms

Modelling nanotechnology devices

Molecular Dynamics (MD)

Earth Systems (ES)



An integrated study of interactions between Earth's components to address global environmental challenges.

Climate change projections

Studying ocean circulation patterns

Modelling natural disasters

Monitoring biodiversity and ecosystems

Evaluating the effects of deforestation

Organised through a data mesh

THE LUMEN DATA MESH IN ACTION

Translating community needs into modular, federated components for discovery, governance, and interoperability

Interoperable Platforms

- Domain-specific discovery portals interconnected via shared data contracts and governance.

White-Label Discovery Platform

- A flexible, EOSC-ready solution adaptable to each domain, ensuring interoperability by design.

FAIR-by-Design Infrastructure

- Tools like LUMIS support the creation of semantic artefacts, aligned with EOSC standards.

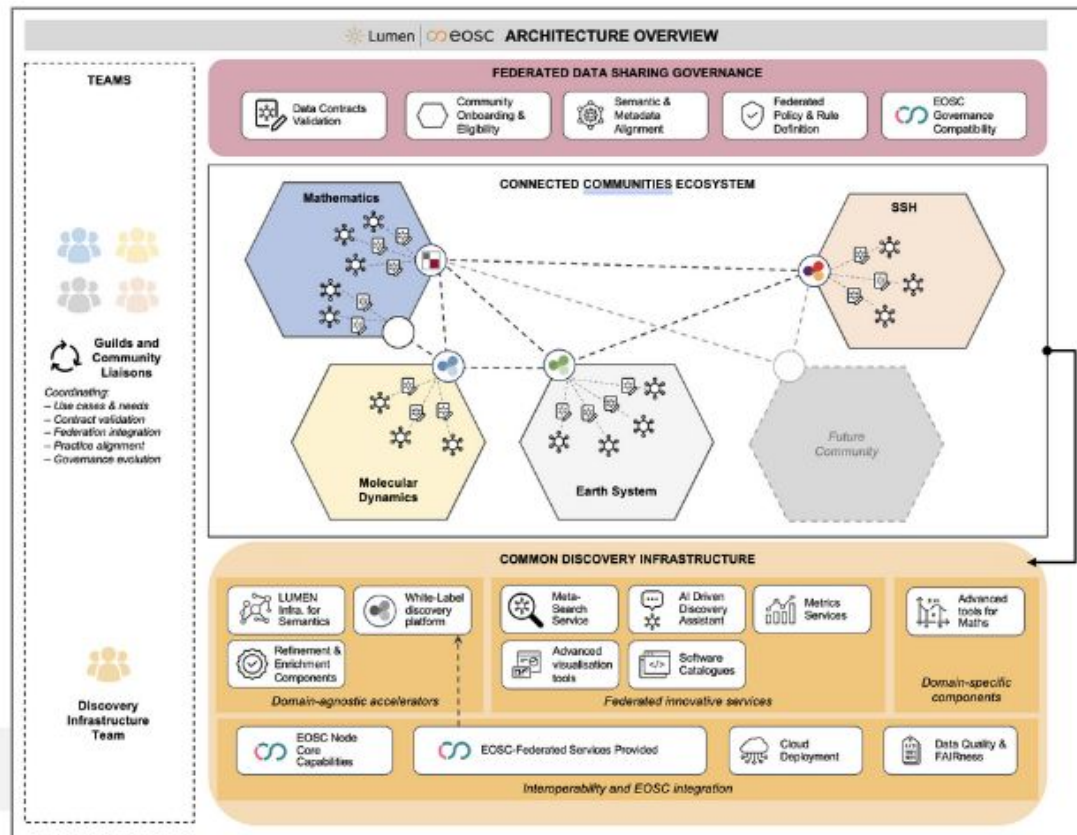
Cross-Domain Discovery

- Meta-search, AIDA, and visual tools enable exploration across datasets, disciplines, and services.

Composable & Community-Governed Services

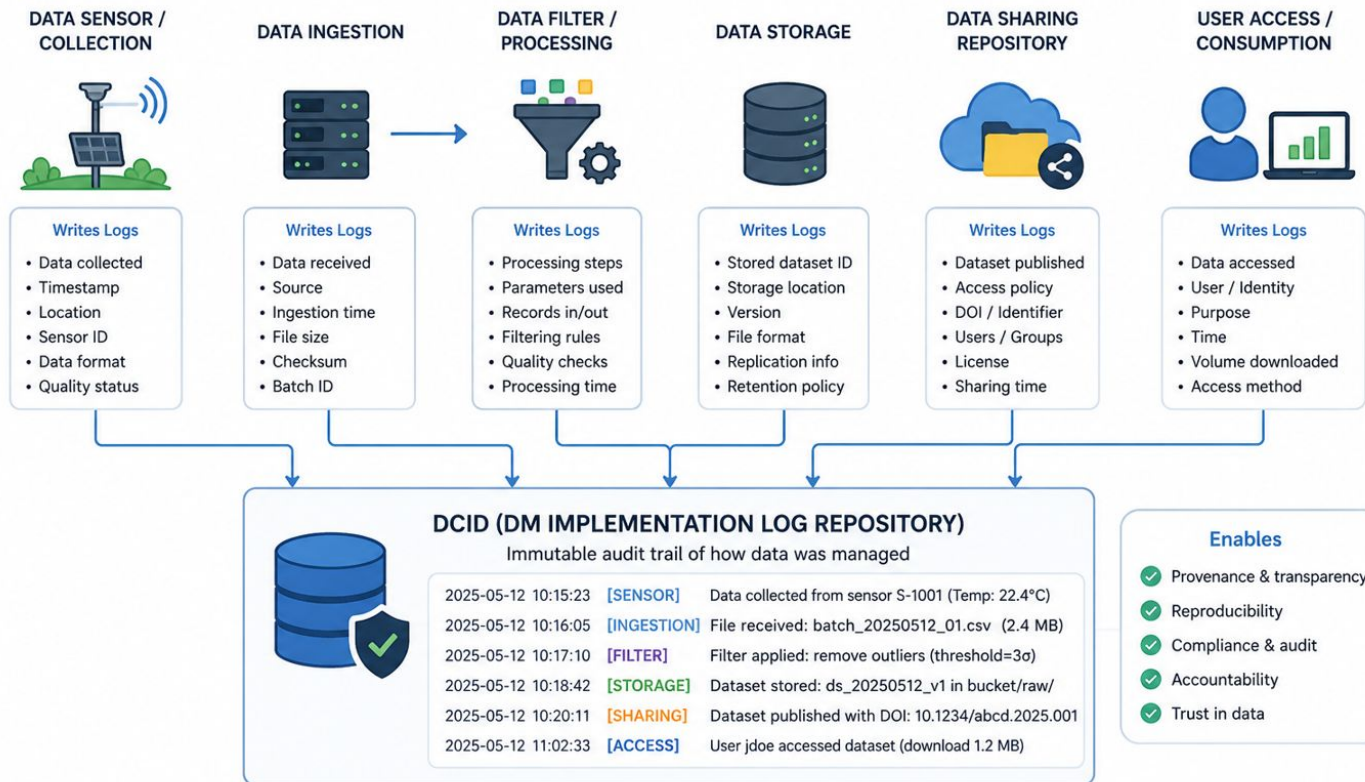
- Reusable catalogues and accelerators managed by domain guilds to foster ownership and reuse.

A modular, federated architecture rooted in community autonomy and interoperability



DM IMPLEMENTATION (DCID)

Capturing how data has been managed



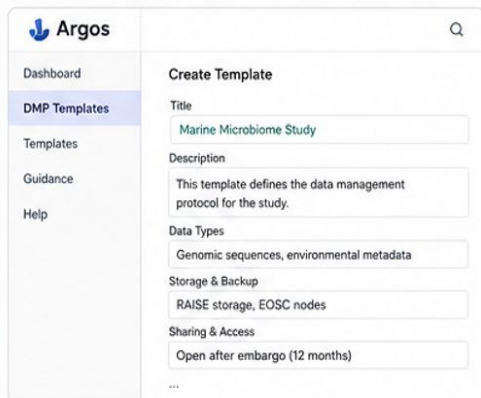


DEMO: How a User Would Do the Whole Process

Example: A researcher runs an experiment and manages data with RAISE

1 Create Generic maDMP Template (Generic DMP)

In Argos (or any maDMP system)



The screenshot shows the 'Create Template' form in the Argos system. The left sidebar has a menu with 'Dashboard', 'DMP Templates' (selected), 'Templates', 'Guidance', and 'Help'. The main form area has the following fields:

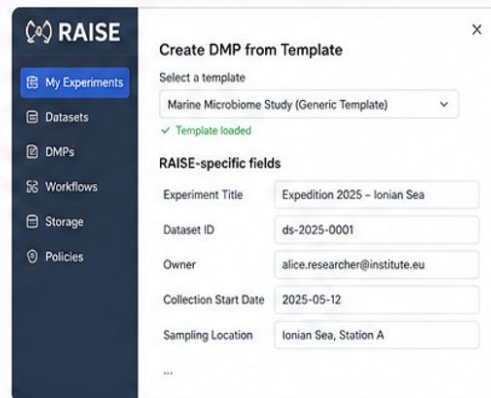
- Title:** Marine Microbiome Study
- Description:** This template defines the data management protocol for the study.
- Data Types:** Genomic sequences, environmental metadata
- Storage & Backup:** RAISE storage, EOSC nodes
- Sharing & Access:** Open after embargo (12 months)
- ...** (More options)



- I create a maDMP template (protocol) before starting my study.
- I define the general data management requirements and lifecycle.
- I save the template.

2 Create RAISE-specific maDMP (Specific DMP)

In the RAISE Portal



The screenshot shows the 'Create DMP from Template' form in the RAISE Portal. The left sidebar has a menu with 'My Experiments' (selected), 'Datasets', 'DMPs', 'Workflows', 'Storage', and 'Policies'. The main form area has the following fields:

- Select a template:** Marine Microbiome Study (Generic Template) [dropdown]
- Template loaded:** ✓ Template loaded
- RAISE-specific fields:**
 - Experiment Title:** Expedition 2025 - Ionian Sea
 - Dataset ID:** ds-2025-0001
 - Owner:** alice.researcher@institute.eu
 - Collection Start Date:** 2025-05-12
 - Sampling Location:** Ionian Sea, Station A
 - ...** (More options)



- I reuse the generic template in RAISE.
- I add RAISE-specific information about my experiment and datasets.
- I save it as my Specific maDMP.

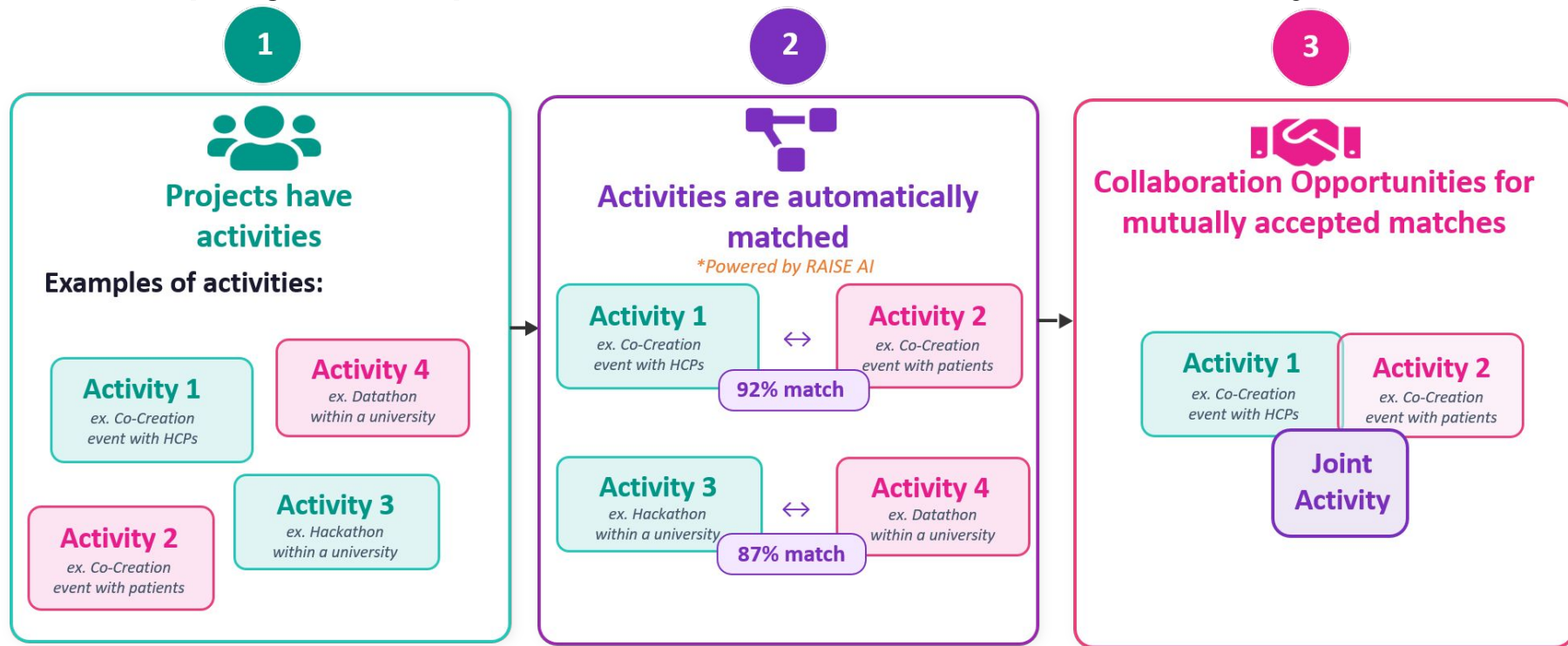
3 Provenance (DPID) (What actually happened)

Captured Automatically in RAISE



- RAISE captures what actually happened during data collection and processing.
- It includes metadata, transformations (e.g., anonymization), policy engine results, workflows, etc.
- All this forms the Provenance Record.

Cross-project requirements and activities similarity detection



<https://requirements.raise-science.eu>

Getting the MoU in place

Phase 1: Q3 2025 - Individual direct project contacts

- Initial 1 to 1 contacts from the projects, looking for potential collaboration ideas, already hinting at expanding towards others

Phase 2: Q4 2025 - Shaping up the collaboration

- Informal discussion at EOSC Symposium (Nov 2025)
- Introductory workshop for all projects (Nov 2025)
- Deciding on MoU as the right format for the collaboration

Phase 3: Q1 & Q2 2026 - Defining the MoU

- Set up of 4 interest areas, define contacts for each project, and leading person for each area
- Develop the MoU text and Joint Work Plan
- Completed signature on June 2nd

Phase 4: Q3 2026 onwards - Start the actual work

Structure of the MoU

Article 1. Parties

Article 2. Purpose and Scope

Article 3. Communications and Contact Information

Article 4. Rights and Responsibilities

Article 5. Funding

Article 6. Entry into Force, Duration and Termination

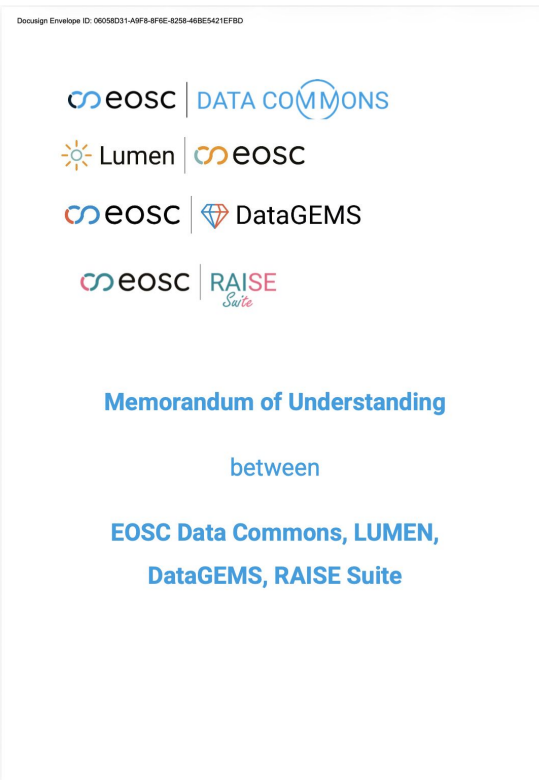
Article 7. Amendments

Article 8. Miscellaneous

Article 9. Language

Article 10. Governing Law and Dispute Resolution

Annex 1. Joint Work plan



Area 1: Sustainability of the services and relationship with the EOSC Federation

The Parties agree to

- **Share the relevant information related to the Key Exploitable Results, Outcomes and Services,**
 - including the expectation on how they can be used and adopted,
 - following the open science principle "*as open as possible, as closed as necessary*".
- **Collaborate for the future adoption and uptake of EOSC-01-05 results,**
 - by promoting them in collaboration with joint communications and dissemination activities towards the EOSC Federation and other relevant stakeholders.
- **Regarding sustainability, the parties agree to share best practices for the sustainability planning of the project results whenever they are mature**
 - (e.g. in a specific workshop),
 - including the plans for integration with the EOSC ecosystem and the potential impact.

Area 2: Communication and Events

The Parties agree to:

- **Promote key content, including press releases, news, results, best practices, and other relevant outputs**, through their respective communication channels (e.g. websites, social media, conferences)
- **Contribute to communication and outreach activities** targeting both specific stakeholders and wider audiences, including the general public.
- **Share expertise on topics of common interest**, including through joint EOSC-related activities where appropriate, and to present activities, tools, services, results, and other reusable outputs.
- Where feasible, **collaborate on joint events and training activities**, ensure timely exchanges of communication milestones, and work together to maximise project impact and visibility across shared audiences.

Area 3: Technical collaboration

The Parties agree to:

- **Organise bi-monthly cross-project technical coordination meetings** to exchange updates on architectures, implementations, and technical challenges.
- **Identify and maintain a shared list of common technical challenges**, in order to highlight cross-cutting needs and maximise synergies.
- **Align on and document common technical standards, protocols, and interfaces** relevant to EOSC integration (e.g. metadata, APIs, authentication, data access).
- **Organise technical deep-dive workshops**, including demonstrations and hands-on activities
- **Identify and map each party's contributions to the support of the phases of the data lifecycle** with the aim of finding potential integration opportunities between projects and fostering the reusability of results.

Area 4: User support and use cases

The Parties agree to:

- **Organise bi-monthly cross-team research meetings**
- **Produce a list of common research problems**, to identify cross-cutting areas, and maximise synergies
- **Define a common timetable of the different activities per project, relating to user work** (organisation/complementarity, etc.)
- **Agree on a timetable for joint participant recruitment**
- **Maintain a common evidence/knowledge repository**
- **Develop a user taxonomy of EOSC users** (methodology will be defined moving forward)
- **Creating a culture of motivation and sharing** to support early career colleagues, students, and practitioners
- **Organise data sharing presentations** with group discussions and critique sessions (during the bi-monthly meetings)

Moving forward

MoU just signed, it's time to get started with the actual activities

Area 1 & 2:

- Contacts established
- News event on the MoU signed published on every project website and EOSC Forum
- Mention of the MoU in the infra-EOSC project last communications meeting

Area 3 & 4:

- Technical and User support areas to start periodic meetings and information sharing after summer.

Questions?