Project Title	FAIR Earth Sciences & Environment services
Project Acronym	FAIR-EASE
Grant Agreement No.	101058785
Start Date of Project	01/09/2022
Duration of Project	36 Months
Project Website	fairease.eu

D3.2 – FAIR-EASE Earth Analytics Lab Services - First release

Work Package	**WP3, Earth Analytics Lab**
Lead Authors (Org)	**Bodéré Erwan (IFREMER)**
Contributing Author(s) (Org)	**Detoc Jérôme (IFREMER), Jossé Marie (CNRS), Bialota Quentin (Geomatys), Schiltzer Reiner (AWI), Keuchkerian Samuel (CNRS), Sarramia David (UCA), Marc Portier (VLIZ), Maria Luisa Chiusano (UNINA), Cox Cymon J. (CCMAR/EMBRC.PT), Boichu Marie (CNRS), Grandin Raphael (IPGP), Troupin Charles (ULiege), , Peter Thijsse (Maris), Tjerk Krijger (Maris), Pansanel Jérôme (CNRS), Giuliano langella (UNINA)**
Due Date	**27.12.2024**
Date	**19/12/2024**
Version	**Final**

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Funded by the European Union

Versioning and contribution history

Version	Date	Author	Orcid	Notes
0.1	31.10.2024	Bodéré Erwan (Ifremer)	ID	ToC
0.2	14.11.2024	Bodéré Erwan (Ifremer)	ID	Executive summary Chapters 1, 2.1
0.3	19.11.2024	Cox Cymon J. (CCMAR/EMBRC.PT)	ID	Chapter 4.5 General review
0.4	27.11.2024	Bialota Quentin (Geomatys)		Chapter 2.2.3
0.5	02.12.2024	Schiltzer Reiner (AWI)	ID	Chapters 2.2.4, 4.1
0.6	06.12.2024	Jossé Marie (CNRS) Detoc Jérôme (Ifremer)	ID ID	Chapters 2.2.2, 3.1
0.7	09.12.2024	Grandin Raphael (IPGP)	ID	Chapters 2.1, 4.3
0.8	10.12.2024	Troupin Charles (ULiege)	ID	Chapters 2.2.1, 4.1
0.9	17.12.2024	Keuchkerian Samuel (CNRS) Sarramia David (UCA)	ID ID	Chapters 2.3, 3.3
0.10	17.12.2024	Boichu Marie (CNRS)	ID	Chapter 4.3
0.11	18.12.2024	Bodéré Erwan (Ifremer)	ID	Chapters 2.3, 3.2 Conclusion General review
0.12	20.12.2024	Peter Thijsse (Maris) Tjerk Krijger (Maris) Pansanel Jérôme (CNRS)	ID ID ID	General review
0.13	02.01.2025	Marc Portier (VLIZ)	ID	General review
0.14	07.10.2025	Giuliano langella (UNINA)	ID	Chapter 4.2
0.14	07.01.2025	Maria Luisa Chiusano (UNINA)	ID	General review
1.0	13.01.2025	Bodéré Erwan (Ifremer)	ID	Final Revision
Final	16.01.2025	Bodéré Erwan (Ifremer)	ID	Final edition for submission

checked if necessary. AI tools have been used within the boundaries of existing laws, in particular regarding the respect of privacy, confidentiality and Intellectual Property Rights.

Table of Contents

Table of illustrations

Terminology

Terminology/Acronym	Description
ARCO	Analysis-Ready, Cloud-Optimised
API	Application Programmer's Interface
BGC	Bio Geo Chemical
CDSE	Copernicus Data Space Ecosystem
DCAT	Data Catalog ontology
DIVAnd	Data-Interpolating Variational Analysis in n dimensions
DoA	Description of Action
EAL	Earth Analytics Lab
EC	European Commission
EOSC	European Open Science Cloud
FAIR	Findable; Accessible; Interoperable; Reusable
GA	Grant Agreement to the project
HATEOAS	Hypermedia as the engine of application state
IAM	Identity and Access Management
IDDAS	Interdisciplinary Data Discovery and Access Services
IDE	Integrated Development Environment
JSON	JavaScript Object Notation
KER	Key Exploitable Results
KPI	Key Performance Indicator
NDVI	Normalized Difference Vegetation Index
ODV	Ocean Data View
OIDC	OpenID Connect
QCV	Qualification / Calibration / Validation
RDF	Resource Description Framework
RO	Research Object
SLC	Single Look Complex
SSO	Single Sign On
UDAL	Uniform Data Access Layer
VSO	Volcano Space Observatory
VRE	Virtual Research Environment
WOA	World Ocean Atlas

Executive Summary

The first release of the FAIR-EASE Earth Analytics Lab (EAL) service marks a significant milestone of the FAIR EASE project to advance data-driven research and innovation in Earth sciences. This milestone is the result of a process that began with the definition of detailed EAL specifications, designed to describe the functional, technical requirements of this platform. To ensure that these specifications were translated into effective solutions, in-depth technical assessments were carried out. These assessments focused on existing tools and applications, identifying opportunities for improvement in the framework of the project while ensuring alignment with the FAIR principles.

Building on this solid foundation, the project transitioned from concept to implementation, turning the envisioned EAL services into reality. The technical insights gained during the evaluation phase paved the way to the development of cutting-edge solutions, tailored to support the Earth science community's needs. The implementation phase also prioritised real-world applicability, demonstrated through the creation of working prototypes based on FAIR EASE cases that showcase the EAL's potential to facilitate advanced data analysis and interdisciplinary collaboration.

The EAL aims to facilitate Earth science analyses through up-to-date technologies while engaging the community. We hope this version is just the beginning, with future iterations designed to extend and refine the platform's capabilities in response to emerging needs and progress made.

1. Back to D3.1

In the deliverable D3.1, *Specifications of FAIR-EASE Earth Analytics Lab and implementation plan*[1], the expected capabilities to address the needs and challenges related to the discovery, access, (re-)analysis and sharing of Earth System data were described. A set of potentially relevant technical solutions was also mentioned for evaluation. An updated reminder of the context and general architecture of FAIR-EASE is provided below.

1.1. FAIR-EASE challenges and proposal

The FAIR-EASE project aims to advance Earth System science through improved data accessibility and processing, interoperability, and collaborative analysis, meeting the new and rigorous demands of modern Earth System data management. To meet these objectives, wherever possible FAIR-EASE prioritises simplicity and efficiency without compromising the openness and FAIRness of the data (Findable, Accessible, Interoperable, Reusable).

The project's strategic focus includes promoting semantic standards and web foundations, such as Resource Description Framework (RDF), Hypermedia as the engine of application state (HATEOAS) and the use of community agreed ontologies and vocabularies, which together support a cohesive framework for data and service interoperability.

FAIR-EASE promotes the use of cloud computing solutions, efficient and environmentally-friendly data access methods, specifically designed and optimised for earth system data flows. The technical solutions proposed are designed to be containerized, ensuring their portability and reusability while facilitating their deployment and integration into existing workflows.

The approach taken by FAIR-EASE underscores three main components that collectively transform data management and data utilisation:

- **IDDAS (Interdisciplinary Data Discovery and Access Services)**: This metadata catalogue leverages existing data catalogues, modeled along a tuned FAIR-EASE DCAT[2] Application Profile. By utilising RDF metadata descriptions for semantic web compatibility, the IDDAS allows users and systems alike to search, discover, and access Earth System data seamlessly. The IDDAS uses both human and machine-readable formats, thereby ensuring accessibility across various platforms within, and between, data spaces.
- **UDAL (Uniform Data Access Layer)**: The UDAL introduces an innovative client-side Data Access Layer designed to streamline functional needs by decoupling the data's "what" (functional content requirements) from its "where" and "how" (technical configurations). This decoupling enhances data accessibility while minimising technical complexities presented to the user, thereby providing a flexible, adaptable interface for users to access data regardless of technical considerations of how the data are supplied.
- **EAL (Earth Analytics Lab)**: The EAL platform provides integrated data analysis capabilities, allowing users to process, visualise, and build workflows around Earth System data in its

[1] https://doi.org/10.5281/zenodo.10069773

[2] https://www.w3.org/TR/vocab-dcat-3/

many forms. This component emphasises collaborative functionalities, enabling interdisciplinary cooperation in the analysis and utilisation of complex datasets.

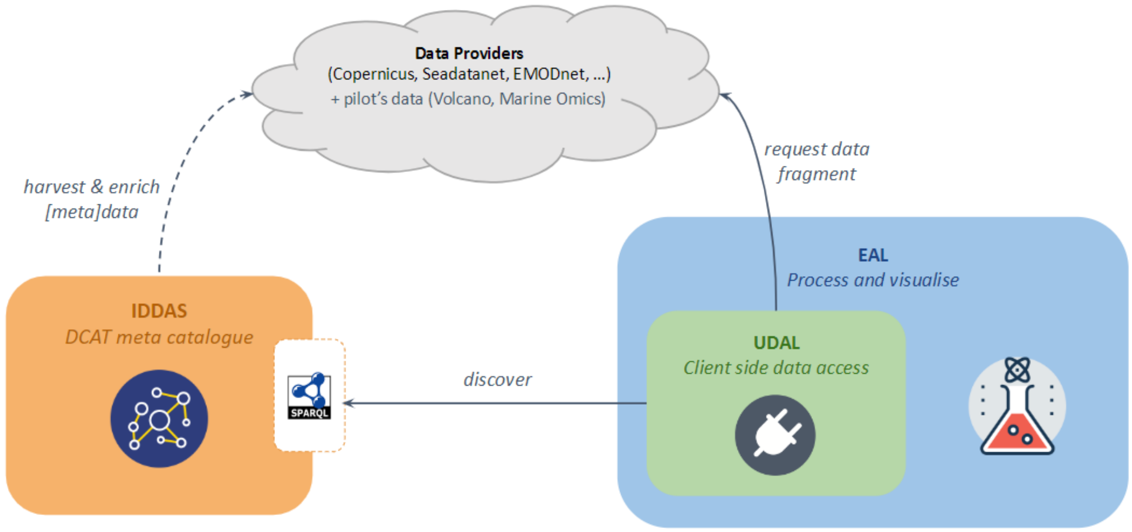


Figure 1 FAIR-EASE general architecture

Ultimately, the FAIR-EASE framework not only prioritises FAIR principles but also fosters interdisciplinarity, allowing diverse fields to benefit from enhanced Earth System data and computing resources. By moving the processing of workflows closer to the data where feasible, the project reduces data transfer and enhances sustainability. This approach reflects FAIR-EASE's commitment to a scalable, user-oriented infrastructure designed to empower Earth System science researchers and stakeholders.

1.2. EAL as guidelines

From a functional perspective, the EAL is based on a stack of building blocks that enable communities to discover, access and analyse Earth System data, including high performance storage and computing resources, geospatial data capabilities, remote processing and interactive applications that can be tailored according to the user needs.

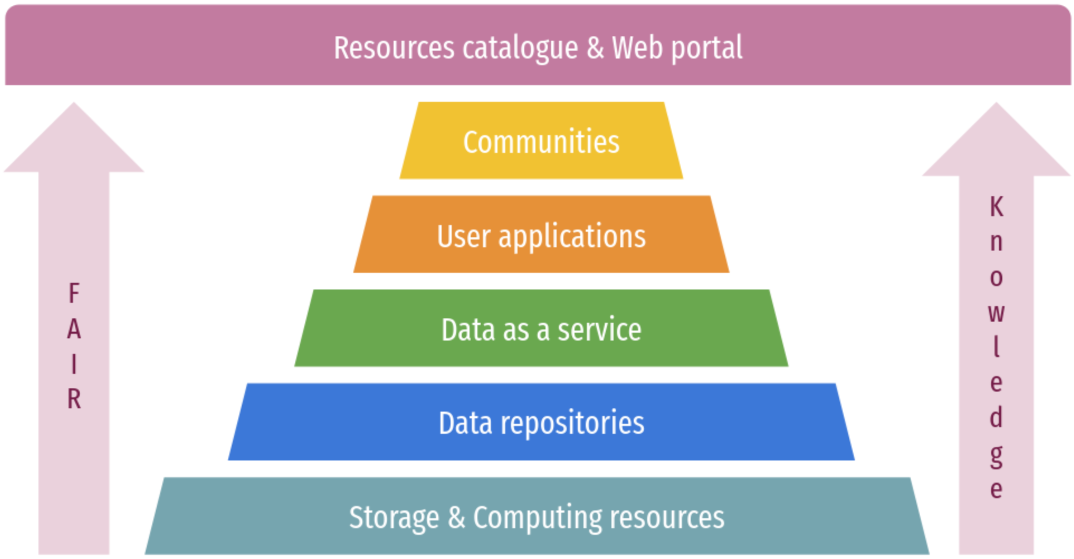


Figure 2 EAL functional architecture

Two transverse pillars form the backbone of the EAL approach: a) FAIR principles, not only for data but also applicable to all relevant research objects, such as software and workflows; b) knowledge sharing and collaborative science, whether by the teams responsible for the technical layers or by the communities using the data and software.

To ensure that users can fully leverage this ecosystem, all these assets are documented and indexed within a comprehensive resources catalogue. This catalogue serves as the central repository, offering a structured overview of the EAL offering. These resources are then presented through a user-friendly web portal designed to simplify navigation and enhance accessibility for scientists, policymakers, and the public alike.

From a technical perspective, user spaces give teams access to public or restricted resources, such as data, tools and analytics services. The link between all the services is ensured by integration through a common identity and authorisation management system.
Simple publication mechanisms should encourage knowledge sharing.

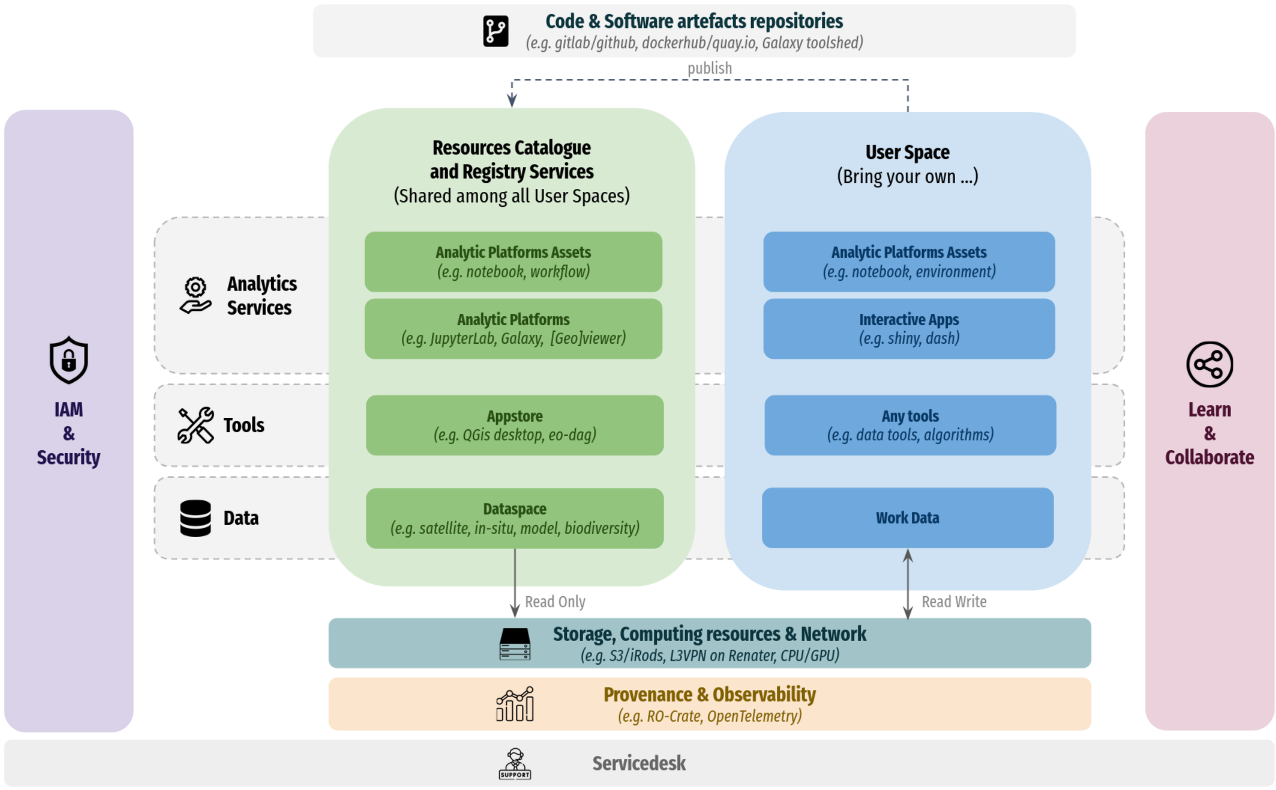


Figure 3 EAL technical architecture

The EAL has been designed to be compatible with the vision of EOSC nodes (see EOSC Federation: Architecture and Federating Capabilities[3]) as well as other architectures such as ESA's EOEPCA plus[4]. The EAL specifications can contribute to the EOSC Opportunity Area 4[5] - *User and resources Environments* - guidelines.

[3] https://doi.org/10.5281/zenodo.13939396

[4] https://eoepca.org/eoepcaplus/

[5] https://eosc.eu/opportunity-area-exp/oa4-user-and-resource-environments/

2. Technical solutions

The EAL not only focuses on technical solutions dedicated to data analysis, but shall also provide services that facilitate data discovery and access, as well as promoting collaborative science within an integrated platform. During the first two years of the FAIR-EASE project, we evaluated and used technical solutions covering these three aspects. A detailed inventory is given below. The use of these solutions by infrastructures or pilots will be presented in the following chapters.

2.1. Data discovery and access

Before presenting the tools and applications dedicated to data processing and visualisation, it is important to briefly mention here the work carried out by the partners involved in data discovery and access (i.e. WP2 and WP4), as a strong coupling between these solutions is necessary to provide an integrated platform.
Details of the work carried out by these work packages are available in deliverables D2.4, *Data Discovery and Access Services - First Release*[6], and D4.3, *Status and expectations of the FAIR-EASE data lake*[7].

2.1.1. Overview

The data services shall facilitate seamless data exploration tailored to user's specific needs, encompassing discovery, access and preview capabilities when necessary. Indeed, according to the user's profile, it can be essential to assess the relevance and complexity of a dataset.

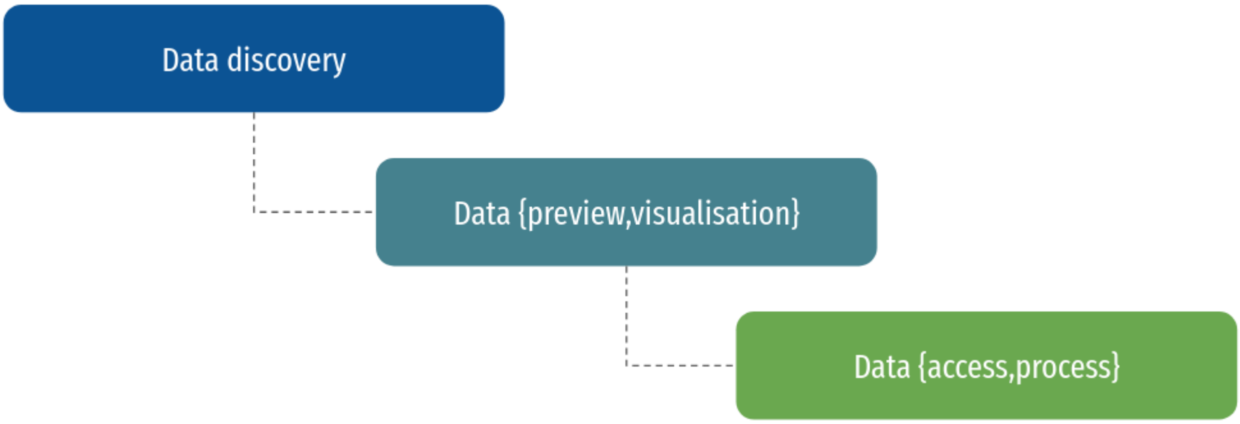


Figure 4 Diving into the data

Data discovery shall cover data sets (i.e. data collections), data files or subset corresponding to search criteria (e.g. spatial and temporal extent, taxon of interest). Data can be previewed by displaying them in a table (only a sample of data for example), chart or geoviewer. Finally, access shall allow direct access to native data files (e.g. data from an Argo float) or to a fragment of data (e.g. ocean surface temperature data over a dedicated area, the application of a Normalized Difference Vegetation Index (NDVI) filter to a satellite image, colocation between different datasets).

[6] https://doi.org/10.5281/zenodo.13165587
[7] https://doi.org/10.5281/zenodo.13933551

2.1.2. What has been done

2.1.2.1. Data discovery services

Two technologies have been evaluated to improve data discovery:

- Search based on semantic web technologies using FAIR-EASE DCAT Application Profile[8]. This allows humans or machines to perform graph-oriented searches using a SPARQL endpoint. A web UI has been developed and deployed : https://fair-ease-iddas.maris.nl.
- Search based on a data inventory using the SpatioTemporal Assets Catalog (STAC) standard. This specification for structuring metadata for geospatial data, which originated in the satellite data community, makes it possible to search for data files meeting specific criteria and benefits from a rich ecosystem, including tools such as STAC Browser, PySTAC, and STAC.jl. Additionally, the STAC and OGC standards are being aligned to harmonise their respective concepts (see OGC RFC[9]).

2.1.2.2. Data access services

Different technologies have been evaluated to improve data access:

- Cloud-optimised or Analysis Ready geospatial formats, such as Cloud-Optimised GeoTIFF (COG), Zarr and Geoparquet. The principle is to divide the data into subsets, often called *chunks*, and to organise and index them in such a way that only the chunks of interest corresponding to the search criteria are accessed. One common specification for organising data in memory is Apache Arrow. These formats, associated with an S3 file system, allow the user to access a subset of the data via the HTTP range requests capability.
- openEO API, which allows users to useEarth observation cloud back-ends in a simple and unified way. OpenEO enables a set of predefined processes to be chained together on Earth observation data. Copernicus Data Space Ecosystem (CDSE) is providing such an API. Additionally, OpenEO API and OGC standards are currently being aligned to harmonise their respective concepts (see OGC RFC[10]).
- Beacon, a high-performance data lake solution developed by Maris, and in use in the framework of SeaDataNet, EOSC-Future, and EOSC Blue Cloud 2026. Leveraging Rust's advanced optimization and parallel processing capabilities, it enables efficient storage and lightning-fast retrieval of millions of datasets, seamlessly integrating with the NetCDF ecosystem for straightforward setup and data exploration.
- Apache Iceberg, a data lakehouse solution, designed for managing large-scale analytic datasets. It supports versioned, ACID-compliant data operations, making it ideal for big data use cases. Iceberg is compatible with distributed storage systems like Hadoop and S3.

[8] https://github.com/fair-ease/asset-standards/tree/main

[9] https://www.ogc.org/requests/ogc-seeks-public-comment-on-adoption-of-stac-and-stac-api-as-community-standards/

[10] https://www.ogc.org/requests/ogc-considering-openeo-as-a-community-standard-comment-sought-on-its-adoption/

Funded by the European Union

2.1.2.3. Client side data access

Some libraries have been evaluated, such as Xarray, Rasterio, GeoPandas, Polars, DuckDB orApache Sedona, for the use of dataframes. Most of these libraries provide:

- simple query mechanisms, via minimal lines of code or SQL-based query engine ;
- efficient access to data in different formats, using principles such as parallelisation and distributed data fragment access (e.g. Mapreduce, Single Instruction/Multiple Data)
- access to data in various formats (e.g. ascii or binary) and using various protocols (e.g. posix, http, S3).

Additionally, the use of a broker to abstract the "where" and "how" of data access provides an effective solution to address the challenges posed by numerous protocols and file formats. problem. We have evaluated three solutions:

- Intake promoted by the Pangeo community, is a python library that provides a large range of reader plugins and a capability to describe the "what" and "where" and "how" in catalogue serialised in a YAML file ;
- EO-DAG, for Earth Observation Data Access Gateway, is also a python library that simplifies search and access to a subset. A Jupyterlab extension is provided and can transform search results to code cells into the active Python notebook to further process/visualize the dataset.
- UDAL, proposed by FAIR-EASE, is an innovative specification to solve this issue. Two main concepts, which can operate independently, form the core of UDAL:
  - Named query (i.e. an URI) that contains the information about the "what " the user wants. A catalogue of demands (i.e. named queries) can act as a registry to share common needs and virtual datasets ;
  - UDAL Broker (i.e. plugin based) that requests the backend service corresponding to the "where" and "how".

2.1.3. What remains to be done

We have seen independently that these different solutions are of interest to users of Earth System data. The interconnection between these different solutions remains a priority in order to provide a coherent stack that makes it easy to dive into the data and provide pilots with the keys to building their demonstrators.

For example, the coupling of IDDAS and UDAL would be useful for some pilots, as would the use of a data lakehouse solution with Analysis-Ready and Cloud-Optimised formats, or the use of Beacon. For the latter, Beacon will be released under an open-source licence at the start of 2025.
Concerning the CDSE, here are some points that should be studied and discussed in order to retrieve satellite data fragments:

- Performance in the perspective of scaling up the service
- Authentication, which still relies on individual credentials for CDSE S3-access
- Quotas for the OpenEO usage

2.2. Analytics services

The EAL shall provide the tools and frameworks necessary for analysing, visualising, and interpreting data. WP3 sought to promote existing solutions by extending their initial scope, where possible, to the use of data from all compartments of the Earth system.

2.2.1. Data Sciences IDE

2.2.1.1. Overview

An integrated development environment (IDE) for data sciences provides a set of tools for developing and executing code, potentially in several programming languages.

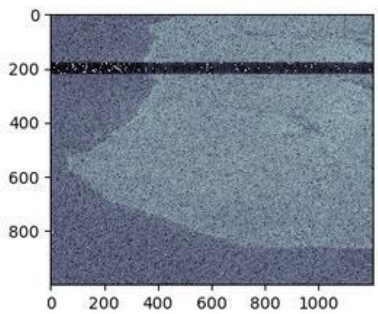
2.2.1.2. What has been done

Two data sciences IDE have been used:

- JupyterLab is a web-based interactive computing environment built upon the core functionality of Jupyter Notebook while introducing new features and enhancements. It provides a comprehensive, customisable and extensive environment. It also offers an intuitive and flexible workspace that empowers users to explore, analyse, and communicate data effectively.

Read subset of TIFF file with Rasterio

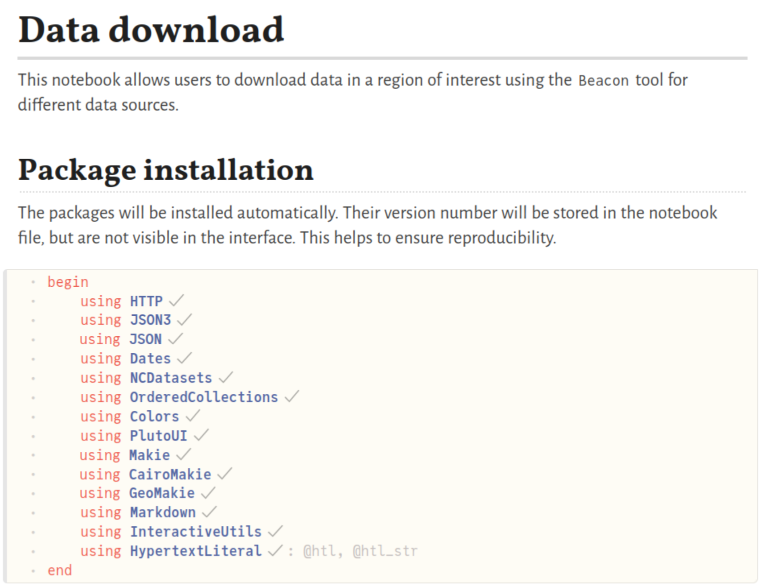
```
[18]: bucket_dir = "Sentinel-1/SAR/SLC/2022/09/28/S1A_IW_SLC__1SDV_20220928T014736_20220928T014804_045198_0566F4_31E8.SAFE/measurement/"
filename = 's1a-iw2-slc-vv-20220928t014737-20220928t014802-045198-0566f4-005.tiff'

start_time = time.time()
with rio.Env(
    AWSSession(
        session, endpoint_url="s3.dataspace.copernicus.eu"
    ),
    AWS_VIRTUAL_HOSTING=False,
    ):
    with rio.open('s3://' + bucket_name + "/" + os.path.join(bucket_dir + filename), "r", driver='GTiff') as src:
        #print(src)
        profile = src.profile
        win = Window(950, 1300, 1200, 1000)
        arr = src.read(1, window=win)

        plt.figure(figsize=(6, 3))
        plt.imshow(np.log10(np.abs(arr+1)), interpolation="none", cmap="bone")
        plt.show()
elapsed_time = time.time() - start_time
print("Wall time: %s s" % elapsed_time)
```




```
Wall time: 15.414582014083862 s
```

Figure 5 Using CDSE S3 bucket from a Jupyter Notebook (subsetting via HTTP Range)

- Pluto.jl is an interactive and reactive notebook environment for the Julia programming language. Pluto.jl focuses on providing a notebook experience that encourages exploratory coding, interactivity, and a responsive workflow. It aims to make data analysis and scientific computing more accessible and enjoyable. Furthermore, it is designed to ensure that the workflows are reproducible by providing a list of the package versions and dependencies within the notebook file.



Data download

This notebook allows users to download data in a region of interest using the `Beacon` tool for different data sources.

Package installation

The packages will be installed automatically. Their version number will be stored in the notebook file, but are not visible in the interface. This helps to ensure reproducibility.

```
begin
    using HTTP ✓
    using JSON3 ✓
    using JSON ✓
    using Dates ✓
    using NCDatasets ✓
    using OrderedCollections ✓
    using Colors ✓
    using PlutoUI ✓
    using Makie ✓
    using CairoMakie ✓
    using GeoMakie ✓
    using Markdown ✓
    using InteractiveUtils ✓
    using HypertextLiteral ✓ : @htl, @htl_str
end
```

Figure 6 Pluto.jl notebook

2.2.1.3. What remains to be done

No particular requests have been made by the pilots, either to share environments or to be able to spawn an environment by providing the URL of a git repository for example (e.g. BinderHub, repo2docker).

2.2.2. Galaxy

2.2.2.1. Overview

Galaxy was first created about 20 years ago for Bioinformatics, nowadays it spreads across domains, notably, thanks to FAIR-EASE, it can now host Earth System sciences. Galaxy provides a comprehensive environment where sharing, processing, and analyzing data is both accessible and efficient. Its intuitive interface lowers technical barriers, making advanced analytical tools accessible without requiring IT skills.

At its core, Galaxy fosters collaboration through its open-access framework. You can build your environment from a controlled public repository called ToolShed, which acts as an application store. This store already offers thousands of tools to access, view and process multiple data formats from different domains. This common repository encourages, firstly, best practices in the design of a software and or code and, secondly, the reuse and improvement of existing tools.

This collaborative work extends to shared histories and workflows, which provide an integrated way to document and exchange insights, ensuring all team members to stay aligned without requiring complex setups. Galaxy ensures that every piece of data used and every tool employed during an analysis is recorded in a history, making each analysis traceable and reproducible. This is essential for validating and advancing research findings. Galaxy natively offers a comprehensive suite for accessing scientific data. Users can upload data, whether large or small, from local storage, FTP, HTTP, or S3 servers. Each dataset is carefully tracked through appropriate metadata, making it easy to share and reuse. The platform supports visualisation and common processing for diverse data formats. As with all features in Galaxy, these are simple to use and can evolve.

For resource-intensive tasks, Galaxy integrates the ability to connect to various types of high-performance computing infrastructures. Moreover, the distributed execution of workflows on remote systems is supported. Galaxy allows users to delegate tasks to specialized infrastructures providing data services directly from their web browser. Currently, there are over a dozen nodes offering data access and computation services with the ability to fetch data directly from private storage solutions, such as Amazon S3, further improving workflows capabilities by minimizing data transfer steps. This delegation of services is secure and authorized through the support of OpenID Connect.

With OpenID Connect, users can securely log in and manage permissions for their data, workflows, and analyses. Galaxy offers fine-grained and flexible access controls for every element it handles or produces, which are essential for multidisciplinary and international collaborations.
Galaxy is designed to support any discipline that uses workflows to analyze data (see figure 5). Its adaptability allows users to tailor their workflows to specific needs, whether it involves genomics, climate modeling, oceanography, or even social sciences. Numerous high-quality training materials covering most of Galaxy's capabilities, along with a large and active community, further enhance this adaptability, equipping users with the skills needed to use Galaxy effectively, regardless of their field.

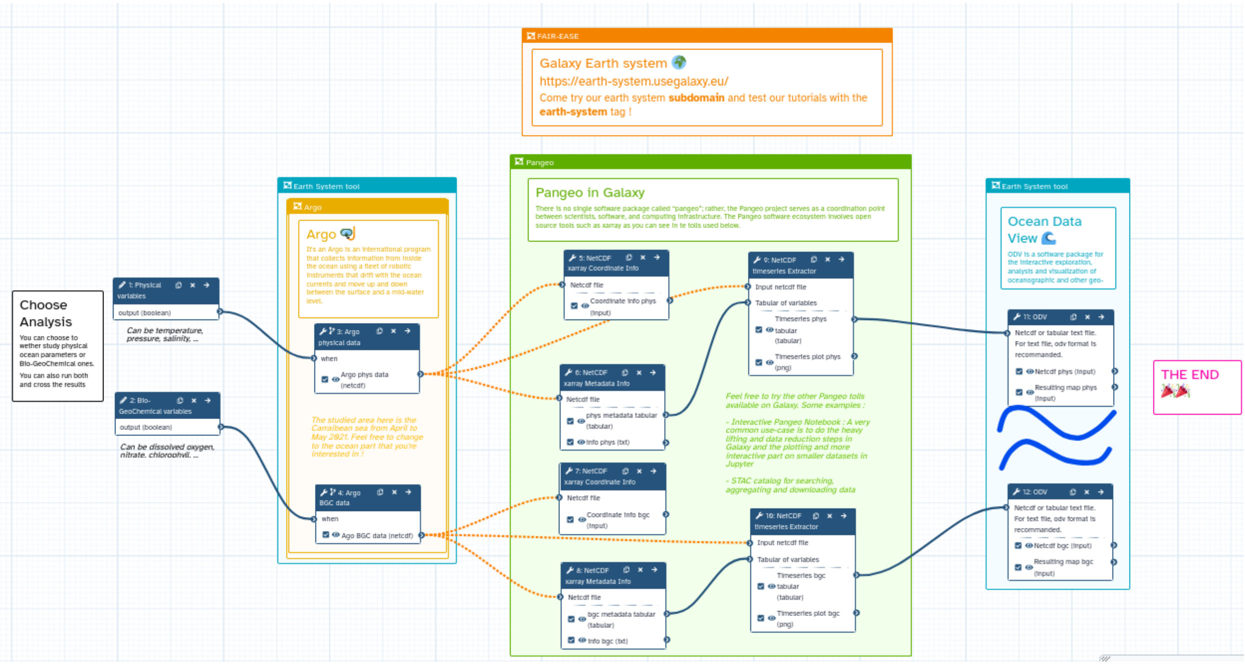


Figure 7 Workflow handling Argo data on the Galaxy workflow editor

Upstream, Galaxy is connected to containerized software image repositories. It can execute Docker images or Conda packages. Each tool available in Galaxy combines such an image with a uniformly designed graphical interface, allowing users to configure the execution of each tool. Downstream, users can directly publish their data analysis results on a trust repository (e.g. Easy Data, Pangea or Seanoe) and their workflows in RO-Crate format on WorkflowHub.

In conclusion, Galaxy represents a powerful solution for researchers navigating the complexities of modern data-driven science. By combining a collaborative framework with robust tools for data management, analysis, and dissemination, Galaxy empowers teams to achieve greater efficiency, transparency, and impact in their work (see pyramid below).

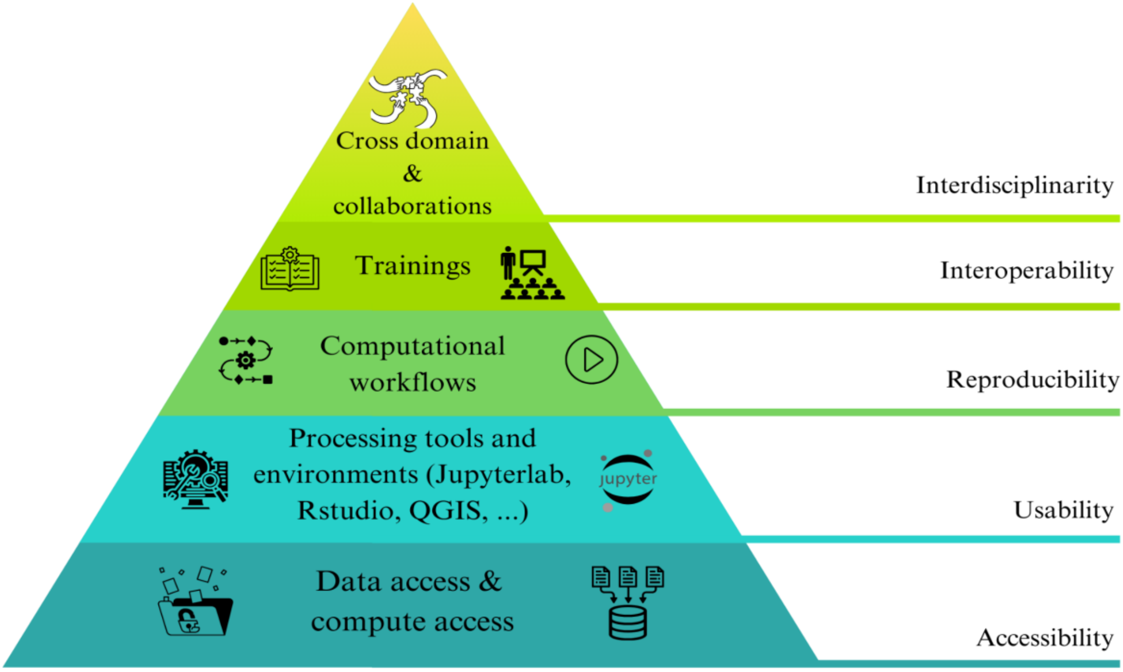


Figure 8 Galaxy maturity pyramid

2.2.2.2. What has been done

FAIR-EASE has added numerous tools useful for geosciences to the Galaxy ToolShed. Here are a few representative examples:

- QGIS: A powerful open-source Geographic Information System (GIS) for viewing, editing, and analyzing geospatial data. QGIS supports a variety of formats and functionalities for creating maps and conducting spatial analyses.
- Argo Data access: provides access to the global network of oceanographic floats that measure temperature, salinity, and other parameters. The tool allows researchers to retrieve and analyze data from ARGO floats, essential for studying ocean dynamics and climate.
- Copernicus Marine Data Store (MDS): facilitates access to Copernicus Marine Environment Monitoring Service (CMEMS) datasets. Researchers can download marine data products, such as ocean currents, temperature, and biogeochemical indicators, for analysis.
- Copernicus Data Space Ecosystem: a tool for accessing and processing Earth observation data from the Copernicus program, particularly Sentinel satellite data. It supports analysis of land, marine, and atmospheric observations for various applications.
- Ocean data view (ODV) : plots geo-referenced and other arrays from netCDF, HDF, GRIB, and other datasets. For now, Galaxy is supporting only netCDF and text files as odv text file as input.
- STAC catalog : for searching, aggregating and downloading data. It's a Spatio-Temporal Asset Catalog (STAC) browser for static catalogs.

Alongside those implementations updates of existing tools were made like Panoply, a geospatial visualization tool used for exploring and plotting data from netCDF, HDF, and GRIB files. The update enhances its compatibility and usability within the Galaxy environment, providing improved functionality for climate and atmospheric data visualization.

To document the usage of each added tool, FAIR-EASE has specifically developed workflows accompanied by training materials, which are published and accessible to anyone on Galaxy's training platform, the Galaxy Training Network. Here are a few representative examples:

- From NDVI data with OpenEO to time series visualisation with Holoviews
- Visualise volcano activities with Sentinel 5 data
- Use the Pangeo ecosystem to analyse ARGO network ocean data

Galaxy offers interactive tools. These tools, such as JupyterLab or QGIS, allow researchers to explore, visualize, and process data interactively at any stage of an analysis performed in Galaxy without leaving the platform, ensuring that all steps of the analysis remain reproducible and traceable. Previously, adding new interactive tools to Galaxy was a complex task. By integrating best practices for containerization and resource management, FAIR-EASE has contributed to making these tools more easily implemented (see https://github.com/usegalaxy-eu/docker-qgis).

Galaxy allows workflows and analyses to be exported using the RO-Crate standard. FAIR-EASE has enhanced this functionality by enriching the standardized content with new metadata, tools, parameters, as well as the input and output data used in an analysis (https://github.com/galaxyproject/galaxy/pull/18820).

2.2.2.3. What remains to be done

The FAIR-EASE project aims to extend the interoperability layers of Galaxy. To achieve this, FAIR-EASE is developing the capability to invoke a Galaxy workflow using the standard protocols of the Open Geospatial Consortium (OGC). The goal is to enable the remote execution of Galaxy workflows on platforms supporting the OGC standard, which is widely used in Earth sciences, thereby facilitating their integration into distributed research infrastructures while ensuring that these workflows comply with FAIR principles.

2.2.3. Examind

2.2.3.1. Overview

Examind community is the open-source platform / server of the Examind ecosystem developed by Geomatys. This map server has a wide range of functions available, manages multiple data formats (e.g. NetCDF, GeoTIFF, etc.), multiple OGC standards (e.g. WCS, WMS, OGC API), and has a number of ready-to-use processes for various uses in the geospatial world. Examind also offers several ways of managing and structuring data. The server can connect to an existing data source, hold the data locally, or generate new data via different processes (e.g. WPS, OGC API Process, openEO), all via different paradigms, such as data cube structuring.

2.2.3.2. What has been done

As part of FAIR-EASE, 3 developments were carried out in response to various pilot use cases.

2.2.3.2.1. Datacube - Temporal Data Aggregation

Work has been carried out on datacube storage of soil data in Italy. This first iteration of Examind's datacube aggregates several raster data files (geotiff or cog type) stored locally or remotely via S3, by specifying the temporal axis in a configuration file (each raster file corresponds to a date, Examind aggregates these files in a 3-dimensional datacube : latitude, longitude, time). It is then possible to expose and query this temporal datacube via various OGC standards such as WMS and WCS, using tools such as QGIS.

Work was subsequently carried out in this datacube to aggregate NetCDF-type data with x dimensions, in order to achieve an n-dimensional datacube.

Figure 9 Aggregation of Soil Data (Italy) by Examind and displayed in QGIS

2.2.3.2.2. OGC WPS / API:processes

The second development on Examind as part of FAIR-EASE was the implementation of a connection with Galaxy, in order to be able to run Galaxy processes via Examind Community's WPS and OGC API Processes interfaces. The aim here was to use the standards developed in Examind to access, run and retrieve the results of scientific workflows hosted on Galaxy. Thanks to these developments, it is now possible to access these workflows from APIs standardised by the OGC, opening up this data and processing to a wider audience.

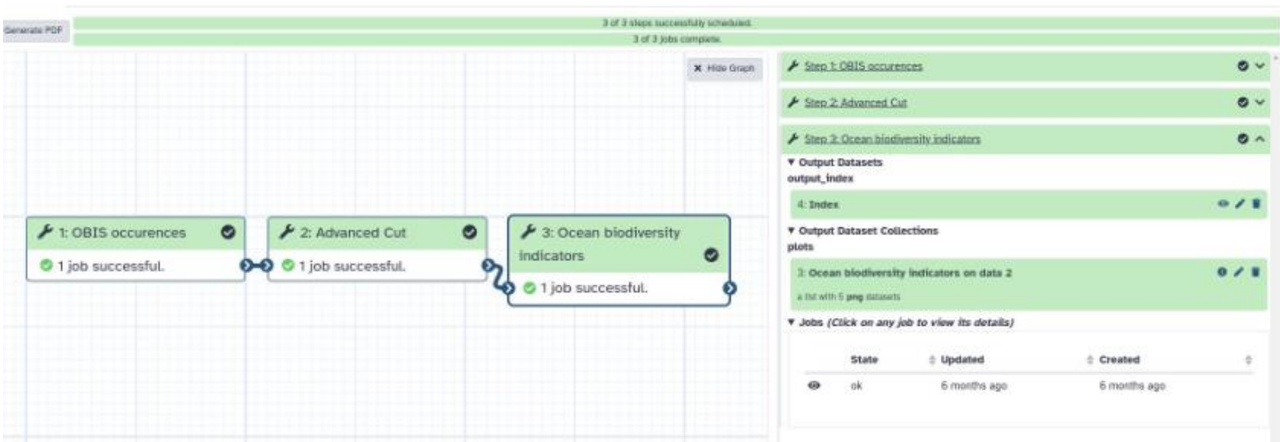


Figure 10 "Marine Omics Visualization" workflow in Galaxy

```
xecute xmlns:wps="http://www.opengis.net/wps/2.0" xmlns:ows="http://www.opengis.net/ows/2.0"
lns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
rvice="WPS" version="2.0.0"
i:schemaLocation="http://www.opengis.net/wps/2.0 http://schemas.opengis.net/wps/2.0/wps.xsd">
ws:Identifier>urn:exa:wps:examind::marine:omics</ows:Identifier>
ps:Input>
  <ows:Identifier>urn:exa:wps:examind::marine:omics:input:new_history_name</ows:Identifier>
  <wps:Data>
      <wps:LiteralData>Marine Omics visualisation</wps:LiteralData>
  </wps:Data>
wps:Input>
ps:Input>
  <ows:Identifier>urn:exa:wps:examind::marine:omics:input:use_cached_job</ows:Identifier>
  <wps:Data>
      <wps:LiteralData>false</wps:LiteralData>
  </wps:Data>
wps:Input>
ps:Input>
  <ows:Identifier>urn:exa:wps:examind::marine:omics:input:inputs</ows:Identifier>
  <wps:Data>
      <wps:LiteralData>{}</wps:LiteralData>
  </wps:Data>
wps:Input>
ps:Input>
  <ows:Identifier>urn:exa:wps:examind::marine:omics:input:parameters</ows:Identifier>
  <wps:Data>
      <wps:LiteralData>{ "0":{ "species":null, "taxon":null, "lat_min":"40.5", "lat_max":"44",
      "long_min":"6.8", "long_max":"9.0" }, "1":{ "complement":"", "delimiter":"",
      "cut_type_options|cut_element":"-f", "cut_type_options|list":["1","7","8","44","78"] },
      "2":{ "colnames":true, "separator":"t", "longitude":"3", "latitude":"2", "species":"4",
      "records":"5", "type":"0", "resolution":"9" } }</wps:LiteralData>
  </wps:Data>
wps:Input>
ps:Input>
  <ows:Identifier>urn:exa:wps:examind::marine:omics:input:parameters_normalized</ows:Identifier>
  <wps:Data>
      <wps:LiteralData>{}</wps:LiteralData>
  </wps:Data>
wps:Input>
ps:Input>
  <ows:Identifier>urn:exa:wps:examind::marine:omics:input:batch</ows:Identifier>
  <wps:Data>
      <wps:LiteralData>true</wps:LiteralData>
  </wps:Data>
wps:Input>
ps:ResponseForm>
  <wps:ResponseDocument status="true" storeExecuteResponse="true">
      <wps:Output>
          <ows:Identifier>urn:exa:wps:examind::marine:omics:output:dataset-collections</ows:Ident
      </wps:Output>
      <wps:Output>
          <ows:Identifier>urn:exa:wps:examind::marine:omics:output:dataset</ows:Identifier>
      </wps:Output>
  </wps:ResponseDocument>
wps:ResponseForm>
Execute>
```


Figure 11 OGC:WPS execution of "Marine Omics Visualization" workflow

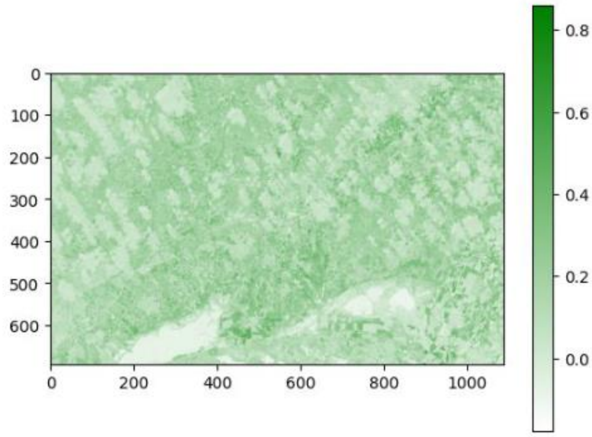
2.2.3.2.3. OpenEO

The latest development for the FAIR-EASE project concerns the openEO standard. OpenEO is an open standard designed to facilitate the processing of Earth Observation (EO) data in a consistent and interoperable manner. It provides a common interface and protocols for accessing, manipulating, and analyzing geospatial data from various sources. One of the key features of openEO is its support for chaining processes, allowing users to create complex workflows by linking multiple data processing steps together. This enables more sophisticated and automated data analysis pipelines. Another significant advantage of openEO is its capability to perform data processing directly on the server, close to where the data is stored. This approach, known as "processing at the source," reduces the need for large data transfers and enhances efficiency, especially when dealing with massive datasets. OpenEO is gaining widespread adoption and is increasingly being used by major institutions and initiatives, such as Copernicus (for exposing Sentinel data for example).

We therefore decided to integrate this standard into Examind to enable pilots to manage data in Examind, and to create process chains on this data to retrieve results (calculated on the server). Thanks to these developments, it is now possible for Examind to manage openEO workflows to, for example, calculate NDVI from Sentinel-2 data.

```
exec_data = {
    "id": "evi-execution",
    "summary": "Enhanced Vegetation Index",
    "description": "Computes the Enhanced Vegetation Index (EVI).
    "parameters": [],
    "returns": {
        "description": "Computed EVI.",
        "schema": {
            "type": "object"
        }
    },
    "process_graph": {
        "load": {
            "process_id": "examind-dynamic.openeo-evi-sentinel",
            "arguments": {
                "dataId": "sentinel_b03_b04_b08:1",
                "bbox": {
                    "west": 3.87,
                    "east": 4.5,
                    "north": 43.85,
                    "south": 43.56,
                    "crs": "urn:ogc:def:crs:OGC:1.3:CRS84"
                },
                "temporal": ["none"],
                "bands": [0, 2, 4]
            },
            "result": "true"
        }
    }
}
```

```
URL = SERVER_IP + "/examind/WS/openeo/process/WPS/result"

headers = {'Content-Type': 'application/json', 'Accept': '*/*'}

r = requests.post(url=URL, data=json.dumps(exec_data), headers=headers)
```



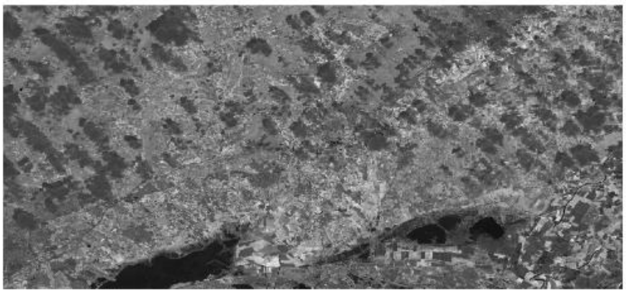


Figure 12 OpenEO request (json on the left) of an EVI computation, and the result (on the right)

2.2.3.3. What remains to be done

A number of tasks are in progress and are expected to be completed for the Fair-Ease project. On the datacube, updates are planned to extend aggregation to other dimensions. In addition, adaptation work is planned, if necessary, if the pilots using datacube require specific use cases.
On the implementation of openEO in Examind, the API must be finished (for the moment some APIs are not finished and bugs persist). There are also plans to carry out tests with openEO clients, to make the implementation compatible with clients of this standard. Still on openEO, work is planned, if necessary, to add processes to Examind that can be called via openEO.

Finally, work is underway on a data lakehouse within Examind, using Apache Iceberg. The aim of this work is to be able to process a large amount of geospatial data, possibly distributed, quickly and efficiently, in order to extend the possibilities offered by Examind when using large volumes of data.

2.2.4. webODV

2.2.4.1. Overview

WebODV is the online version of the popular Ocean Data View (ODV) desktop software for analysis and visualization of marine and other environmental data, currently used by more than 10,000 researchers worldwide. Like ODV, webODV provides an interactive graphical user interface and offers rich feature sets via context menus that are specific to the clicked or tapped object.

While the desktop ODV requires that the analyzed datasets reside on the end user machine or an attached storage device, webODV works differently. All datasets offered to users, as well as a special version of the ODV software reside on dedicated webODV servers. Users do not have to install any software or download the sometimes very large datasets. Instead, users simply connect to these datasets using the webODV server URLs in their web-browser. New browser tabs will open for every opened dataset, with each tab providing interactive user interfaces that look and feel like the ODV desktop interface, thereby allowing the large ODV user community to easily switch to webODV. Communication between the webODV server and the browser is via low-latency Websocket messages allowing a true interactive experience.

Subsetting of datasets using versatile station and sample filters, as well as the flexible arrangement of a wide range of data plot types works in exactly the same way as for ODV desktop. webODV processing speeds match ODV desktop levels, allowing creation of publication-ready graphics containing hundreds of millions of data points in just a few seconds.

2.2.4.2. What has been done

Development during the FAIR-EASE project has led to major advances: (a) webODV and ODV now also support netCDF classic files commonly used for satellite data as well as model and re-analysis output. (b) webODV/ODV now allows export of graphics in GeoTIFF format with geolocation information embedded inside the image file. These GeoTIFF images can be imported as separate layers in Examind and other GIS software. (c) OpenID Connect authentication functionality has been added and a webODV Docker image was created for a webODV deployment on D4Science infrastructure. This server will provide a variety of satellite sensor data as well as re-analysis and numerical model outputs, thereby complementing other existing webODV servers that host in-situ data for all compartments of the Earth system.

The figure below shows example webODV and ODV visualisations of multidisciplinary datasets obtained in separate browser tabs and separate local ODV instances. Each one of these multidisciplinary datasets can be explored online using the full set of interactive features.

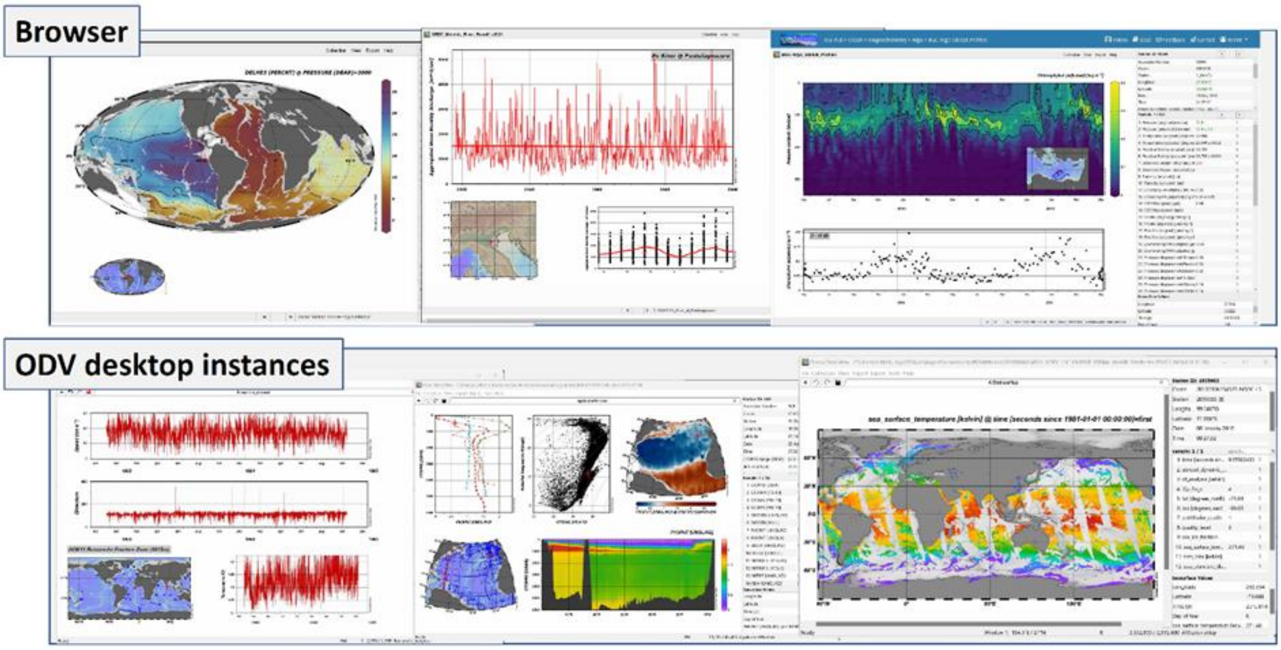


Figure 13 ODV and WebODV visualisations

2.2.4.3. What remains to be done

Development is now underway that will allow us to go beyond and link multi-disciplinary data by exchanging data blobs between different tabs in the web browser and/or local ODV desktop instances. The received data will be used in three modes: (1) *compose* (visually add the received data to the existing plot), (2) *compute* (use data of the plot and the received data to calculate and display new quantities, such as the difference), and (3) *correlate* (correlate the existing and received data using space and/or time information as linkage).

2.3. Core services

Core services act as the backbone linking all the applications that make up EAL, as well as the underlying infrastructures. These core services provide users with value-added functionality, streamlining access to data and computing resources while encouraging scientific collaboration efforts. To ensure the seamless integration of applications and services within the EAL, it is crucial to rely on standard protocols and services. As the implementation of these basic services is the responsibility of EAL implementers, this chapter concentrates on presenting the standards evaluated or approved by the FAIR-EASE project.

2.3.1. IAM and security

EAL and all its related services are accessible via a common Identity and Access Management (IAM) to guarantee technical interoperability and ease of use for all users and teams (i.e. Virtual Organizations). OpenID Connect (OIDC) is an identity layer built on top of OAuth 2.0, enabling secure user authentication and identity verification. It allows applications to obtain basic user profile information through an ID token, facilitating single sign-on (SSO) across systems.

The European Grid Initiative (EGI) provides services to communities, including EGI Checkin, which become the building blocks at the heart of the EOSC federation. EGI Checkin connects to a wide range of identity providers (IDPs), such as Edugain and ORCID. Technical solutions promoted by the FAIR-EASE project have been modified where necessary to interface with OIDC, and indirectly with EGI Checkin, such as WebODV, Galaxy and S3 storage.

To avoid making credentials visible, for example when sharing a notebook accessing a service requiring authentication, the EAL shall provide a solution for managing secrets and injecting values at runtime only. Due to time constraints, the FAIR-EASE project will not focus on secret management solutions. However, it should be noted that Galaxy offers natively secret management via Galaxy Vault.

Ideally, sensitive information should also be encrypted to limit the risk of leakage.

2.3.2. Storage and computing resources

Storage and compute resources are essential elements to process data and run workloads with flexibility and scalability.

Among the key elements, data grids play an essential role in interconnecting disparate data spaces. They facilitate the seamless integration and sharing of large-scale data between distributed systems, ensuring that information can be accessed and used efficiently across different platforms and geographical locations. This functionality is not addressed by the FAIR-EASE project, as there is no need to interconnect dataspaces. However, there are solutions already deployed in communities across Europe, such as iRods, Rucio and One Data. The latter is the solution implemented by EGI DataHub.

Cloud storage, exemplified by services such as Amazon S3, offers web-based data access with advanced features such as serverless subset or compatibility with OIDC. It facilitates access to data from data spaces but also data from user space from heterogeneous and distributed applications and infrastructures. As the FAIR-EASE project promotes standards and cloud-ready solutions, we believe that the S3 filesystem is a solution at the heart of the EAL. Analytics services, such as Galaxy, JupyterLab (e.g. via the boto3 library in python), WebODV and Examind are compatible with S3.

On the computing side, cloud environments offer versatile computing resources encompassing temporary storage, memory and processing power via CPUs and GPUs. These resources are tailored to the needs of a wide range of applications, whether through high-performance computing (HPC), high-performance data analysis (HPDA) or cloud computing clusters. The sometimes complex access to these resources should be kept as simple as possible, even hidden by web applications such as analytics services (e.g. Galaxy, JupyterHub, remote processing via Galaxy Pulsar or OGC:API processes).

2.3.3. Provenance and observability

2.3.3.1. Summary

The EAL and its ecosystem of tools and services require these to ensure FAIRness of data and workflows, as well as providing insights for open science and reproducibility. Establishing provenance is critical for ensuring transparency, reproducibility, and trust in scientific outputs, hence affecting the project outputs and their uptakes by communities. RO-Crate is a framework for packaging and sharing research outputs with rich, machine-readable metadata to ensure they are reusable and interoperable. It uses JSON-LD to describe datasets, software, and workflows, making them understandable to both humans and automated systems. This standard supports open science by enabling seamless sharing, collaboration, and long-term preservation of research. RO-Crate is notably used by the UC5.3 “Marine Omics Observation” as a container for describing their data. It is also used by Galaxy to export all the metadata about a workflow run.

Metrics are used to evaluate the performance and impact of the infrastructure. Accounting procedures support the project’s development teams by ensuring that adequate tracking and documentation of resource usage is set to optimize costs and allocation. This functionality is not addressed by the FAIR-EASE project. However, a solution such as OpenTelemetry could be considered. Indeed, OpenTelemetry is an open-source framework for collecting, processing, and exporting telemetry data such as traces, metrics, and logs, making it ideal for monitoring resource usage in environments like an Earth Analytics Lab. By standardizing observability data and supporting metrics collection, it enables tracking of resource consumption by users or teams, which can be integrated with tools for accounting and analytics.

2.3.4. Learn and collaborate

Sharing knowledge is important not only to improve technical and functional skills, but also to avoid reinventing the wheel. This involves sharing data or any other type of file, as well as sharing notebooks and workflows. Another important point is the sharing of learning materials, such as webinar or MOOC.

The Galaxy Training platform is a good example. All of these elements must be centralised in the resource catalogue in the same way as the data (see next chapter). There are now solutions for producing interactive and reproducible documentation from code and rich text (e.g. Markdown). They allow for flexible output formats, customizable designs, and seamless integration with various programming languages, ensuring dynamic and shareable content. As part of the FAIR-EASE project, we studied two solutions:

- JupyterBook, an open-source tool that allows users to write content in markdown and Jupyter notebooks, enabling seamless integration of live code and output. JupyterBook is commonly used for building data science reports, academic papers, and educational materials, with support for multiple formats like HTML, PDF, and ePub.
- Quarto, an open-source platform that supports multiple programming languages like R, Python, and Julia, and allows users to generate reports, presentations, websites, and books. Quarto is designed for seamless integration with various scientific computing workflows and can output to multiple formats such as HTML, PDF, and LaTeX.

Funded by
the European Union

We propose to create, in the first quarter of 2025, a Quarto website on the FAIR-EASE Github to document the technical achievements as well as the pilot projects. This site could also include functions for querying the catalogue of resources.

Collaboration capabilities should be driven by the User Space and analytics services.

2.3.5. Resources catalog and discovery portal

A resources catalogue is a centralised repository that organises and provides access to any relevant type of resources, such as data, software artefacts (e.g. software, tools, scripts) and learning materials. This catalogue shall provide metadata adapted to the different types of resources, as well as the ability to link metadata records together to provide associated resources.

We have not yet started work on this feature. However, several technical solutions will need to be considered:

- A graph-based solution, using semantic web technologies (i.e. ontologies) to extend the FAIR-EASE Research Knowledge Graph. This could be seen as an extension to the approach already taken up inside the IDDAS (based on RDF, DCAT and SPARQL). Alternatively separate graph technology solutions such as Neo4j could be applied.
- A metadata management solution, such as:
    - CKAN, an open-source data management system designed to catalogue, manage and share datasets and resources. It offers robust metadata management, advanced search and API access, making it ideal for organising diverse resources. Its extensibility and support for standards such as DCAT allow it to be used as a flexible resource catalogue for research projects.
    - GeoNetwork, an open-source cataloging tool that implements ISO 19115 and ISO 19115-3 metadata standards, making it ideal for managing geospatial data and services. It ensures interoperability by providing structured metadata and a user-friendly interface for easy search and discovery of resources. While optimized for geospatial data, GeoNetwork can be extended to manage non-geospatial resources by customizing metadata schemas to include diverse data types, tools, and learning materials.

The use of controlled vocabularies and various semantic artefacts (e.g. ontologies, mapping) ensures consistency in the description of the various resources and simplifies the search for and cross-referencing of information. The use of https://earthportal.eu is a response to this need.

Ideally, several types of search would be interesting to provide based on the resource catalogue, such as graph-oriented search, thesaurus search, full-text search or search via a chatbot (using Large Language Model). Access to these search APIs could be encapsulated in web components. This makes it easy to integrate the catalogue into any website.

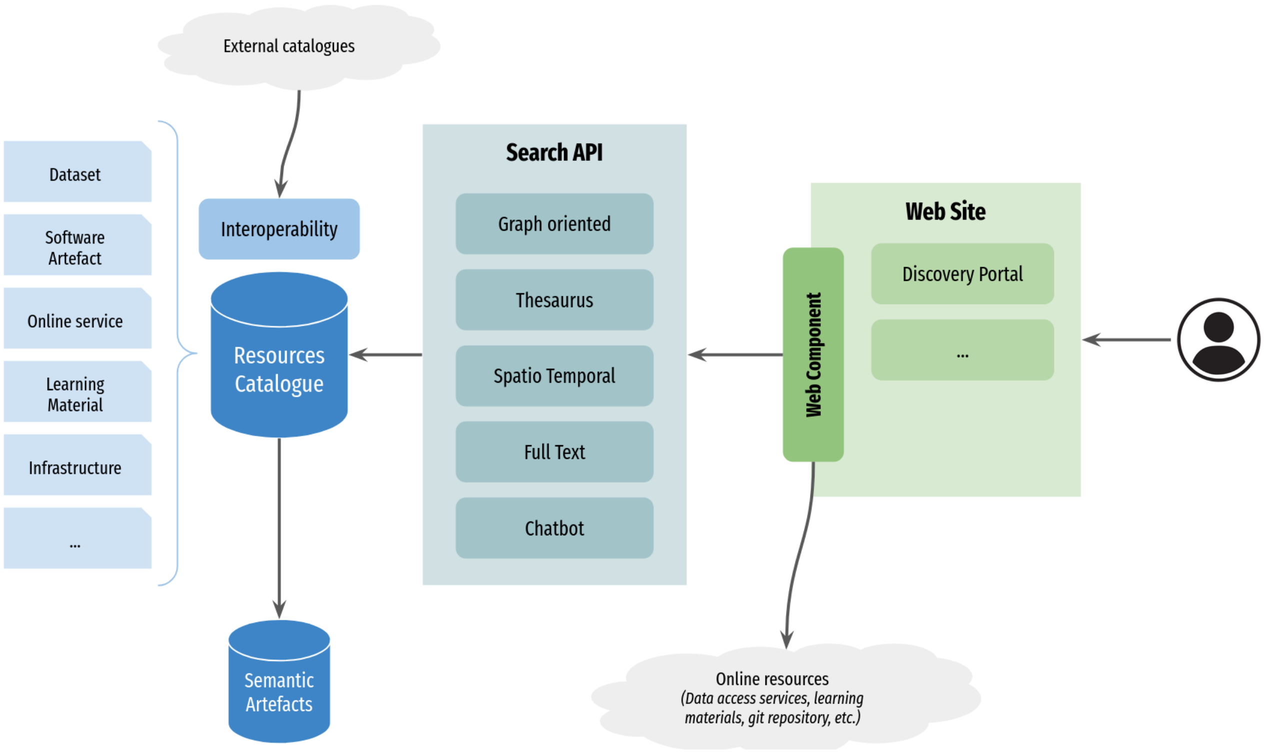


Figure 14 Resources catalog and usages

3. EAL implementations

Many implementations can provide some or all of the capabilities of EAL, offering flexibility tailored to the needs of the specific domain of use, while being complementary and compatible with similar implementations. As part of the FAIR-EASE project, three implementations are used:

- Galaxy Europe in collaboration with the EOSC EuroScienceGateway project ;
- D4Science Virtual Labs in collaboration with the EOSC Blue-Cloud 2026 project ;
- FAIR-EASE testbed, set up at the University of Clermont-Auvergne (UCA), in collaboration with Gaia Data project, which supports the Data Terra research infrastructure.

3.1. Galaxy Europe

3.1.1. Overview

Galaxy Europe is an open source and free access web-based platform, designed to provide accessible, reproducible, and collaborative data analysis tools to researchers across various disciplines. It provides free access to the Galaxy functionalities described earlier in this document. Hosted at https://usegalaxy.eu, it is based on a robust cluster infrastructure located in Freiburg, Germany, ensuring high computational capacity and reliability. The maintenance and administration of UseGalaxy.eu is done primarily by the Freiburg Galaxy Team in collaboration with other academic groups across Europe and with the US Galaxy team. The Galaxy Europe platform offers thousands of documented and maintained tools, free registration, 250 GB per user, and on-demand training capacity. It hosts several domain-specific versions of the usegalaxy.eu server. You can access any of these subdomains and find data and tools focused on the following research areas: Egology, Climate, Cancer research, Genome annotation, Earth System sciences, etc.

Galaxy Europe distributes computational tasks across a distributed network, including the University of Freiburg (Germany), the EGI ReCaS-Bari Cloud node (Italy), the French Institute of Bioinformatics cluster (France), the Vlaams Supercomputer Centrum (Belgium), and others. This distributed architecture enhances its ability to handle large-scale data analysis while promoting accessibility and collaboration. By integrating thousands of tools, workflows, and training materials, Galaxy Europe supports a wide range of research domains, fosters FAIR data practices, and advances open science across Europe and beyond.

3.1.2. What has been done

Within the FAIR-EASE project, Galaxy Europe made significant advancements to support Earth System Sciences by establishing a dedicated subdomain, https://earth-system.usegalaxy.eu tailored to this field. This effort included the implementation of over ten new tools and workflows within Galaxy Europe (as described in a previous chapter on Galaxy). These enhancements were designed to provide researchers with comprehensive resources for analyzing and visualizing complex Earth system datasets.

To maximize utility and interoperability, workflows were created that integrate these tools while reusing existing resources from communities like Pangeo. Pangeo tools, known for their efficiency in large-scale geospatial and climate data processing, were incorporated into Galaxy workflows to enable data analysis and visualization.

Funded by
the European Union

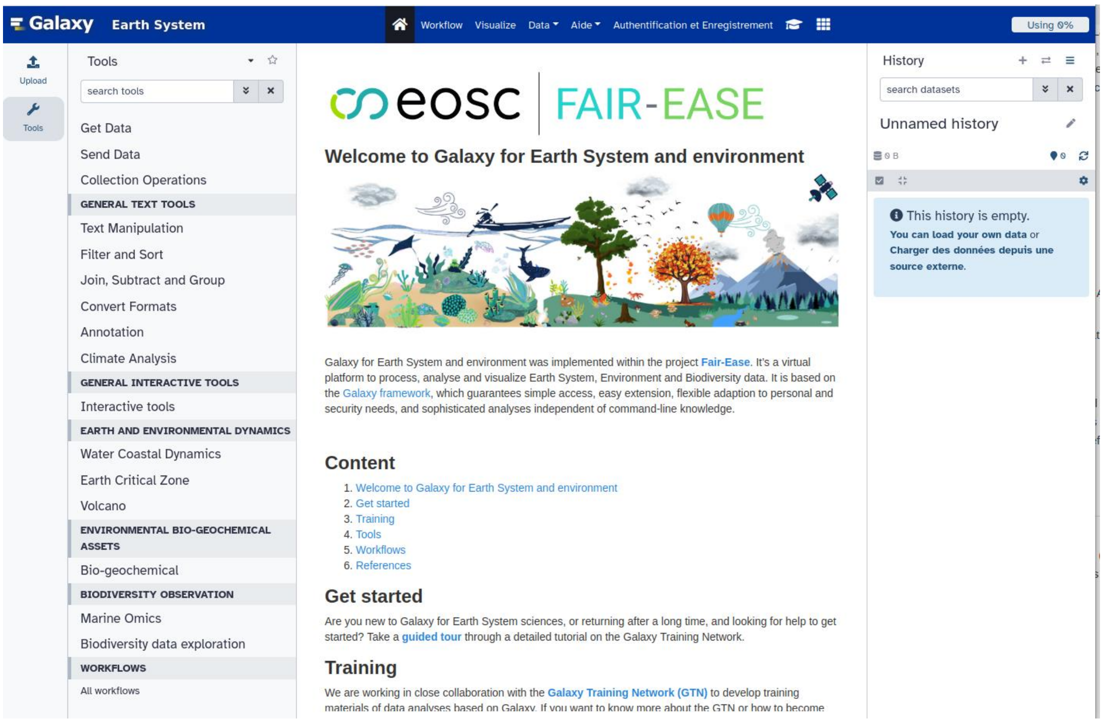


Figure 15 Galaxy Europe's Earth System sub-domain

Additionally, all developments were documented within the [Galaxy Training Network](), ensuring reproducibility for users, developers, and administrators.

Efforts to improve accessibility and findability extended to publishing select workflows in WorkflowHub (ongoing work), further aligning with FAIR principles. Indeed, WorkflowHub is an open, online platform designed to enhance the discoverability, accessibility, and interoperability of scientific workflows. It acts as a central repository where researchers can share, search for, and reuse workflows across different disciplines.

FAIR-EASE is now also including its own newly established endpoint at the Hellenic Center for Marine Research (HCMR), allowing Galaxy to run jobs remotely without needing a shared file system. This new node supports tasks for usegalaxy.eu and is secured with password-protected communication. It is equipped to handle containerized job execution and has already been successfully tested with tools like the OBIS Marine Indicators.

This distributed approach, combined with the tailored tools and workflows, exemplifies Galaxy's commitment to fostering a collaborative, reproducible, and accessible environment for Earth system science research.

3.1.3. What remains to be done

Continue to provide tools, workflows and training answering both FAIR-EASE's pilots needs and the Earth system sciences needs, in order to have the most complete Galaxy offer for Earth sciences.

3.2. D4Science

3.2.1. Overview

D4Science is a hybrid data infrastructure, hosted by CNR Italy, designed to support scientific collaboration and resource sharing across various domains, including biology, ecology, environmental studies, social sciences. Established in 2014, it connects over 24,000 scientists from 50+ countries, integrates data from 50+ providers, and provides access to over a billion records in global repositories.Its key feature is the provision of Virtual Research Environments (VREs) and Virtual Laboratory (VLab), web-based platforms tailored to specific community needs. Powered by gCube technology, D4Science ensures scalability, flexibility, and interoperability. This e-infrastructure emphasises robust service reliability and offers features such as dynamic software deployment, secure data sharing, large-scale analytics, and resource lifecycle management. It operates across two primary infrastructures in Italy, leveraging advanced cloud-computing technologies and stringent governance protocols to ensure efficient operations.

3.2.2. What has been done

FAIR-EASE has made steps to engage with D4Science in joint developments to serve the communities represented by the FAIR-EASE use cases. The aim is firstly to test the e-infrastructure proposed by D4Science and then to pursue future collaboration if the technical aspect proves useful. On the technical side, some elements have been provided by D4Science:

- Creation of a dedicated gateway: https://fair-ease.d4science.org/
- Creation of two Virtual Laboratories for Coastal Water Dynamics and Marine Omics Observation
- Deployment of two new services promoted by FAIR-EASE: Galaxy and WebODV

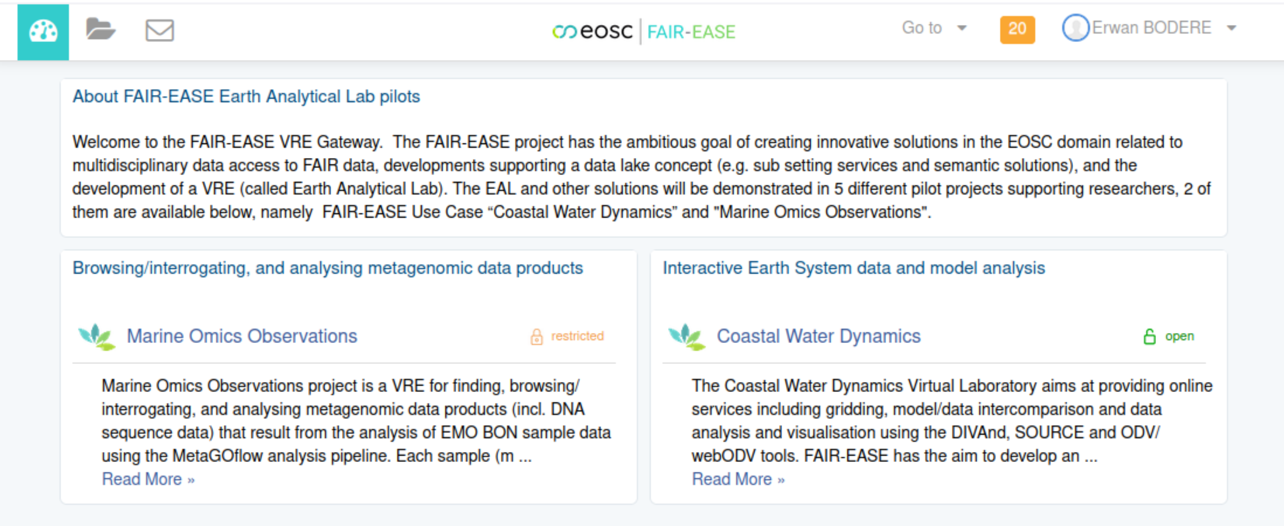


Figure 16 FAIR-EASE Gateway in D4Science e-infrastructure

A Memorandum Of Understanding (MoU) was signed between Blue Cloud 2026 and FAIR-EASE in November 2024, to strengthen collaboration between the two projects.

3.2.3. What remains to be done

The deployment of new FAIR-EASE services is delayed by the availability of the D4Science technical team and the process for integrating new services. However, thanks to the MoU between Blue-Cloud 2026 and FAIR-EASE, it will be possible to deploy services that will benefit both projects, such as the integration of IDDAS to select the files to be uploaded to one's workspace (with a specific development from the WP2). Another element of interest would be the deployment of Examind Community software to complete the geospatial services. Finally, the demonstrators for the two pilots ”Marine Omics Observations” and “Coastal Water Dynamics” hosted on D4Science still need to be finalised.

3.3. FAIR-EASE testbed

3.3.1. Overview

Along with the various existing infrastructures (D4Science, Galaxy EU, EGI), a testbed has been designed to encompass a maximum of data management and data services tools. The aim for FAIR-EASE is to provide an environment where pilots and technical WPs can test the deployment and the use of services in the same place with storage and computing resources. The testbed is deployed on the UCA mesocentre in Clermont-Ferrand (France) which has an OpenStack Iaas Cloud, a distributed Ceph storage and an HPC cluster (with CPU and GPU). The Ceph storage provides a production level S3 object storage and the OpenStack can provision virtual machines. Another interesting feature is that UCA mesocentre also hosts a Galaxy instance (https://galaxy.mesocentre.uca.fr/) with around 400 tools installed.

The first deployment of FAIR-EASE datalake testing version coincides with the first release of the FAIR-EASE EAL documented in this deliverable, also due at Project Month 24. The two deliverables have been edited in a closely coordinated way to reflect the thorough exchanges between the two technical WPs. The Deliverable 4.4 only solely aims at providing a status of the deployment of FAIR-EASE innovative services on the different infrastructures and in the different environments that have been selected for hosting them. The deployment status will be updated in Deliverable 4.5 that will describe the operational version of the data lake.

3.3.2. What has been done

Based on the requirements from the pilots and the key findings on the tests driven in former devCycles, a set of technologies and tools have been selected and tested.

Actually, the most of the work has been done on data use and data backend management in order to give new insights to pilots according to new approaches and particularly cloud approach for data use in end users practices. To this aim, UCA testbed had provided:

- A dedicated 1 To object space on the S3 UCA endpoint. The credentials are individual and the users can create buckets and objects they need. Data from copernicus and Glodap have been uploaded and manipulated
- Tutorial Jupyter notebooks on data creation and manipulation: starting from a local jupyter notebook (with files available locally but downloaded from Copernicus) to a cloud version

(using files on the S3), and including data transformation (from standard format to cloud optimize format), and validating the use of cloud optimized format in current practices.
- Validated the ability to automatically register new deposited files (in dedicated bucket) in a STAC catalog

Outside the UCA testbed, but linked to the forecast deployment, we have performed preliminary tests on deploying Apache Iceberg and MinIO object storages on local resources (Irish Marine Institute).

3.3.3. What remains to be done

From the cloud computing environment side, a set of VM machines are planned to be provided in order to deploy data services and some analytics services, namely Beacon, Examind, Apache Iceberg, WebODV, Galaxy, JupyterHub.

From the cloud data access, the S3 endpoint is to be continued, and its capacity to be extended if required. The public and private access will be explored deeper in order to promote the availability of intermediate data products.

From the computing side, Galaxy Pulsar is to be deployed to validate the use of computing resources in Galaxy workflows (as well as those provided by HCMR and also Datarmor at Ifremer), and combine it with S3 access from Galaxy workflows.

From the SSO/AAI side, EGI Checkin is to be used and technical contacts have been taken to make the deployment of this easier.

We have conducted initial tests to create VOs on the EGI environment and we will ensure that pilots have access on the longer term through a MoU that will be signed based on the work by WP1 task 1.4.

4. From hands-on to demonstrators

The technical WPs have proposed and promoted solutions to meet the needs of the five FAIR-EASE pilots. After a hands-on phase, the plan is to work together on scenarios that will demonstrate the usefulness and relevance of the technical solutions. Each demonstrator covers different components or technical solutions and will use different implementations of EAL.

4.1. UC5.1.1 - Coastal Water Dynamics

4.1.1. Overview

The Coastal Water Dynamics Pilot focuses on the coastal marine environment near river estuaries, where important processes, such as the evolution of plankton blooms or the transport and fate of nutrients, carbon and contaminants, critically depend on a large variety of factors, including river discharge, ocean circulation, meteorological conditions as well as internal biogeochemical processes in the water column.

As a consequence, important socio-economic and scientific targets, such as (1) marine biological productivity and fish stock buildup, (2) uptake of atmospheric CO_2 and the effects on the marine carbon cycle, as well as (3) input and off-shore transport of suspended material and hazardous substances by rivers can only be fully understood by combining and linking large multidisciplinary datasets and by developing effective and easy-to-use software tools that enable such combined exploitation.

This demonstrator focuses on the Po river estuary and the Northern Adriatic region as target areas. The aim is to extend the web-based tools webODV, DIVAnd and SOURCE to enable and facilitate the combined scientific usage of the diverse environmental datasets as well as of the output from the growing number of numerical models. Expected users include Earth System scientists across all disciplines, students of the environment at various levels, decision makers and the interested general public.

4.1.2. What has been done

All three tools were further developed and adapted to meet the Pilot's needs. In collaboration with WP2, 3 and 4 as well as with the Environmental bio-geochemical assets pilot, the tools were also integrated into a variety of larger data workflows and data quality control environments. Also, a new webODV server was deployed on D4Science infrastructure to serve satellite data as well as re-analysis and numerical model output to users. Details of these developments are described below.

The DIVAnd software tool, primarily available through JupyterLab, can now be used in bash mode (i.e., using scripts, instead of notebooks) as well as in the Pluto interface, a simple and reactive environment specific to Julia. The SOURCE software tool has been reorganised in Jupyter Notebooks and work is still ongoing to increase its flexibility, its FAIRness and to integrate some functionalities to allow the interaction with webODV and DIVAnd tools.

Funded by
the European Union

The webODV server software was extended to include an OpenID Connect authentication step and was packaged into a Docker container that is pulled and run on D4Science hardware. A selection of satellite data as well as re-analysis and numerical model outputs are now available via the D4Science webODV server. Together with the already operational webODV servers, the D4Science deployment will complement the distributed data system providing very low barrier online access to a wide range of multi-disciplinary community datasets.

Other major developments of the ODV and webODV software include: (1) creation of geo-referenced GeoTIFF images for use in QGIS or Examind, (2) extension of the station and sample filtering capabilities, and (3) simplified creation of data overlays and axis range synchronisation. These enhancements facilitate data quality control procedures established by the Environmental bio-geochemical assets pilot.

In Galaxy we updated the ODV, DIVAnd and SOURCE software to their latest versions, and we also implemented data workflows involving more than one of our tools. Some of these workflows are fully functional, some still need further improvements. HowTo documents describing how to use these workflows are available.

4.1.3. What remains to be done

Each one of the various multidisciplinary datasets available via the operational webODV servers can already be explored online using the full set of interactive features. As next step, we will go beyond and link multi-disciplinary data by implementing a new system allowing exchange of data packages between different tabs in the web browser and/or ODV desktop instances running locally on the user’s computer. The received data will be used in three modes: (1) *compose* (visually add the received data to the existing plot), (2) *compute* (use data of the plot and the received data to calculate and display new quantities, such as the difference), and (3) *correlate* (correlate the existing and received data using space and/or time information as linkage).

A set of Pluto notebooks are being prepared, with the goal of creating a high-resolution climatology in the Northern Adriatic Sea, based on the application of the DIVAnd software tool on observations obtained using different Beacon services. As the same observations can be present in different datasets (for instance, Argo and SeaDataNet), an appropriate method for the detection and removal of duplicates has to be implemented and tested. Finally, the climatology will be used to provide a list of potential outliers. The list could then be fed back to the original data providers so they can conduct more advanced quality checks, hence closing the loop between data providers and data users.

SOURCE tool has been further developed in order to make the code more flexible and robust. An input module to gather in situ observations and model data is under development, exploiting the FE services, i.e. UDAL and Beacon (WOD, CORA TS (CMEMS)). The deployment in Galaxy has to be finalized with the stable version of SOURCE and the workflow with DIVAnd to use the climatologies as additional reference data to compare with in situ time series will be finalized. The visualization of SOURCE output will be finally visualized by webODV.

Funded by
the European Union

4.2. UC5.1.2 - Earth Critical Zone

4.2.1. Overview

The Earth Critical Zone represents the dynamic interface where rock, soil, water, air, and living organisms interact. It is critical for understanding environmental processes and their influence on climate, ecosystems, and human activities. In FAIR-EASE, this use case focuses on creating a land degradation assessment tool and on embedding it within an innovative geospatial cyber-infrastructure called LandSupport. The ambition is to contribute to the United Nations Sustainable Development Goals (SDGs), in particular indicator 15.3.1, which focuses on land cover change, soil productivity and carbon sequestration.

This pilot will demonstrate how technical solutions can enhance data discovery, processing, and reproducibility. Specifically, the demonstration will use the IDDAS catalogue to search and retrieve relevant datasets, UDAL to manage uniform data access and subsetting, Galaxy for orchestrating and running scalable workflows, and Examind for advanced geospatial data cube management. An evaluation will be carried out to determine which components will be deployed within the LandSupport infrastructure (e.g. Galaxy is stand-alone), showcasing how the combined technical contributions can be put into practice to ensure FAIRness of data and improve the efficiency of land degradation assessments at both local and regional scales.

4.2.2. What has been done

- Code development: this activity started from the analysis of the Trends.Earth code, a plugin installed in Quantum GIS (QGIS) capable of running the land degradation assessment analysis. The main objective was to get a Python code that can make the same calculations of the three indicators (Productivity, Land Cover and Soil Carbon) while removing any dependence on Google Earth Engine and QGIS. The development is concluded.
- Code sharing: the code is on GitHub but the repository is not yet public, waiting to conclude the development stage.
- Galaxy: a close collaboration was established with the Galaxy team to add the land degradation assessment tool to the Galaxy suite.
- Tool development: early versions of the tool have been tested from the backend point of view, including the creation of raster datacubes about the input data, such as NDVI by MODIS, land cover by Copernicus, and soil taxonomic units from ISRIC.

4.2.3. What remains to be done

- IDDAS / UDAL: integrate the semantic search offered by IDDAS and the data access layer offered by UDAL in the main land degradation assessment code.
- Examind: possible solutions about the optimised management of raster datacubes will be explored using Examind.
- FAIRness: application of FAIR principles to the tool but also to the output data created by the tool after any user request.
- Tool testing: tests are going to be performed to ascertain that the three versions of the tool produce the same results. The three versions are the original Trends.Earth, the stand alone Python code, and this last code embedded within Galaxy.

Funded by the European Union

- Tool refinement: further code refinement is planned to enhance the flexibility and the modularity about data and GIS operations, and to enhance the computing performance.
- Data integration: expand data collections that can be included in the analysis, ensuring a more detailed assessment of land degradation.
- Outreach and training: embed the Python code within the LandSupport platform and dedicated Jupyter notebooks to reach the communities working on land and soil; prepare training modules to familiarize researchers and stakeholders with the platform and the tool.
- Scaling and deployment: scale the developed solution using the distinct capabilities offered by the LandSupport data center in the University of Naples and the huge Galaxy infrastructure. Scaling is crucial to get assessment results in a short and reasonable time.

4.3. UC5.1.3 - Volcano Space Observatory

4.3.1. Overview

The Volcano Space Observatory (VSO) aims at bringing together a broad set of satellite and ground-based observations relevant to the monitoring of volcanic activity and associated atmospheric hazards, serving the scientific objectives of the Atmospheric Science and Solid Earth communities. The VSO is built around a web-based portal that enables users to explore available datasets. Year 2 of the FAIR-EASE project has been devoted to the development of automatic interactive processing tools for the retrieval of volcanic emissions, mainly focused on atmospheric satellite products (Sentinel-5P/TROPOMI), building upon the database of the VSO, and their implementation into open-access web-based interfaces.

In parallel, developments for the processing of Sentinel-1 data, aiming at efficient data access and computation interferograms, has been conducted to nourish the Solid Earth component of the VSO for the monitoring of volcano ground deformation. Sections below specifically describe recent advances and future prospects on this component of the VSO pilot.

4.3.2. What has been done

Regarding the new developments achieved in Year 2 for the determination of volcanic gas emissions, an article was published to present a novel TROPOMI-based methodology for estimation of volcanic SO2 mass flux for any degassing volcano worldwide (Grandin et al., 2024, https://doi.org/10.1029/2024JB029309). Associated to these developments, a new open access web application named “SO2 Flux Calculator”, partner of the VolcPlume web platform, was released at https://dataviz.icare.univ-lille.fr/so2-flux-calculator (Grandin et al. 2024, https://dx.doi.org/10.25326/655). The service is based on an open-source package available from https://git.icare.univ-lille.fr/icare-public/so2-flux-calculator. The dataset of the publication, with application on Etna and Piton de La Fournaise volcano degassing activities, has been uploaded on an open-data repository, in compliance with FAIR principles (https://doi.org/10.57932/235f8c42-142b-40ee-9948-518e83554a7d).

Regarding our progress on Solid Earth data, the Copernicus Data Space Ecosystem (CDSE) has been selected for Sentinel-1 satellite data access for monitoring the “ground deformation by interferometry” sub-service.

The OpenEO service has been, at first, considered as a potential method for data access, but was discarded for two reasons: (a) OpenEO only provides data products that have been projected on a regular geographic grid, hence at a processing level that does not preserve the geometric information of the radar data, and does not preserve the phase (an essential ingredient for interferometry). (b) OpenEO offers limited data access, with quotas that do not allow for scaling up the service.

Alternatively, the Volcano Space Observatory has selected a low-level data access method via the S3-storage offered by the CDSE, combined with boto3 library for session handling and gdal/rasterio for raster access, to ingest Sentinel-1 images and process interferometric products from multiple Sentinel-1 images. Data identification and selection is based on the OpenSearch API.

In parallel, the choice of the software that will be used for the processing has been investigated, based on several criteria, including (a) accuracy, (b) portability, (c) computational efficiency, and (d) capability to leverage the data access service offered by the S3 system of the CDSE. The package EOS-SAR, co-developed in partnership with the Kayrros company, has been selected. An agreement is being secured between the parties, that will enable the VSO to operate an open-access web-based service focused on the generation of Sentinel-1 interferograms for volcano monitoring, with EOS-SAR as the computation engine under the hood.

A first prototype of a web-based interface for the service has been developed to allow the user to select (a) the volcano of interest, (b) the Sentinel-1 track and (c) the temporal window for the processing, relying on the OpenSearch API for granule selection.

4.3.3. What remains to be done

We plan to develop an API for the InSAR processing of Sentinel-1, using the EOS-SAR processing toolbox that integrates the capability, thanks to the gdal/rasterio libraries, to perform range requests on the CDSE S3 datalake. The API will trigger processing on the AERIS-ICARE-Univ Lille computation facilities. It will ingest a JSON file generated by the web-based interface (see description above). The JSON file describes the list of Sentinel-1 granules that need to be processed. The results will be returned to the user in the form of a datacube of coregistered SLCs, along with the corrections for each SLC (orbital, topographic and tropospheric). A second API will be developed to ingest this datacube and perform the interferometric calculations itself, by combining the coregistered SLCs and associated corrections. The selection of pairs will be defined based on a certain “bandwidth” around the diagonal of the InSAR covariance matrix. Interferograms will be unwrapped and geocoded, and eventually provided to the users as a geotiff. Depending on available time, further post-processing tools for analysis and visualization may be developed.

4.4. UC5.2.1 - Ocean BGC

4.4.1. Overview

The observation of marine biogeochemical processes (BGC) is useful to address fundamental scientific questions regarding the health of our ecosystem (e.g. ocean acidification, oxygen minimum zone, biological carbon pump, phytoplankton communities...) and needs for ocean resource management. Today, BGC sensors are deployed in the global ocean through various

Funded by the European Union

platforms (floats, gliders, sea mammals, moorings, etc.) which need to be adjusted and validated to allow proper scientific usage. Because of the diversity of the scientific tools, methods and data, the aim of this pilot is to provide a common platform to data scientists to qualify / calibrate / validate (BGC) data. It answers actual issues like data management, method standardization and interoperability between platform types.

In order to do so, UC2.1 proposes to build the platform as a web application offering three services: qualification (Q) / calibration (C) / validation (V), centralizing existing methods / scientific algorithms by building specific workflows depending on the variable to process.
In the framework of FairEase, we plan to deploy 2 QCV workflows for Nitrates and Chlorophyll-*a* on Testbed (through web application) and Galaxy Europe (on separate standalone tools). We also plan to complete the galaxy Europe toolkit with published and useful BGC tools such as canyon-B, canyonMed or Soca Light. Data visualization and qualification will be done by the webODV application. Also the data access follows UDAL specifications.

4.4.2. What has been done

- In a FAIR way, the app has been containerized as all its major components, which can be run independently now.
- A front-end has been developed as a job manager. It embarks multiuser authentication with Keycloak to ensure the shareability of results between users to avoid work duplicity. This front end is still under development.
- A data manager has been developed to harmonize and manage the multiple sources of in situ and gridded data. For now, the tool manages ARGO, GLIDERS, and WOA23 sources. This tool has been built with the UDAL specification in cooperation with Jorges to manage the data source distribution. (Ongoing development on Galaxy)
- In order to use webODV facilities, the tool creating custom collections for data qualification and visualization usage has been developed on top of ODV and is operational. This tool is also containerized to be shareable.
- Calibration tools for nitrates workflow as Neural Network tools, Climatological comparison tool and Gain estimator have been developed and containerized .
- Finally, an API ensuring communication between Backend and Frontend has been developed and is still under development.

4.4.3. What remains to be done

- Before the deployment on the testbed
    - finalizing the nitrate workflow by connecting all the components together through the API. Optimizing middleware with user cache mechanism.
    - Finalizing the data management section: calibration outputs and exports. Then, it is necessary to include more data sources: at least GLODAPv2 and Satellites data
- Realize the Chlorophylle workflow including satellite colocation inclusion in UDAL, SocaLight methods implementing in calibration methods
- Investigate use for a set of Beacon nodes for gaining access to subsets of Copernicus, EMODnet, WOA and Argo data.
- Continue the Q.C.V. tools conversion to be deployed as GALAXY tools.

4.5. UC5.3.1 - Marine Omics Observation

4.5.1. Overview

The European Marine Omics Biodiversity Observation Network (EMO BON) is an initiative of the European Marine Biological Resource Centre (EMBRC) to establish a persistent genomic observatory among designated European coastal marine sites, sharing the same protocols for sampling and data curation. Environmental samples are collected from the water column, and at some sites, soft sediments and hard substrates (Autonomous Reef Monitoring Structures - ARMS), together with a set of mandatory and discretionary metadata (including Essential Ocean Variables - EOVs). Samples are collected following specified protocols at regular and specified intervals, and sequenced in large six-monthly batches at a centralised sequencing facility. The use of standard operating procedures (SOPs) during data collection, library preparation and sequencing aims to provide uniformity among the data collected from the sites; coupled with strict adherence to open and FAIR data principles, the SOPs ensure maximum comparability among samples, and enhance reusability and interoperability of the data with other data sources. The observatory network was launched when the first sampling campaign took place, in June 2021. All EMO BON metagenomic samples are analysed with the sample analytical workflow constructed specifically for the project called MetaGOflow (MGF: https://github.com/emo-bon/MetaGOflow). MGF is based on the EMBL-EBI MGnify workflow (v.5) from which it was derived, and performs raw sequence curation, sequence assembly, and taxonomic and functional annotation based on the read sequences. A typical MGF analysis produces approx. 25GB of data products that are packaged in specialised ro-crates with the actual data files being stored in an S3 object store. The MGF data product ro-crates and an RDF triple store of the metadata describing the EMO BON sampling events are entered into the F-E IDDAS catalogue.

Within the FAIR-EASE project, the aim of this pilot implementation is to develop a VRE prototype on D4Science with links to the IDDAS catalogue. The IDDAS catalogue will contain a DCAT profile for the EMO BON RDF metadata triple store and the dataset records for each of the MGF ro-crates. The VRE will interface to IDDAS to find data and use the UDAL to search and sub-set the data based on the RDF triple store. The IDDAS/UDAL will also provide access to data sets, and subsets, of co-located data from other data spaces. The VRE will contain a number of Jupyter Notebooks that provide default, continuously integrated, data representations and analyses (e.g. alpha and beta diversity analyses between taxonomic and functional attributes of samples), plus a number of workflows demonstrating potential analyses that can be conducted on the EMO BON data (e.g. biosynthetic gene cluster analyses). Analyses in Jupyter notebooks use the Galaxy API to run analyses on the Galaxy backend which automatically designates a specific compute node within the Pulsar network that is local to the data.

4.5.2. What has been done

- Sampling event (meta)data validation (Pydantic), quality control procedure using explicit rules and GitHub actions, and definitions of the RDF triples (classification and semantics)
- RO-crate specification and generation software for the MGF data products
- Jupyter notebooks of data visualisations of taxonomic and genetic alpha and beta diversity
- Jupyter notebook (Python) workflows for biosynthetic gene cluster identification using software on the Galaxy backend via their API

- Galaxy workflow execution using newly installed Galaxy Pulsar node
- Execute workflow via OGC:WPS (on top of Galaxy specific API)
- Initial VRE construction on the BlueCloud platform

4.5.3. What remains to be done

- DCAT catalogue templating descriptions of EMO data assets, i.e. IDDAS ingestion framework for the metadata triple database and the MGF data product RO-crates
- Metadata RDF triple generation, and SPARQL endpoint construction
- Storage solution for MGF data products - S3 Object Store in testbed infrastructure
- IDDAS ingestion framework for the metadata triple database and the MGF data product RO-crates
- Incorporate our RO-crates and triple store into the UDAL code, provide named queries
- Documentation and landing pages from the VRE on BlueCloud2026 infrastructure.
- Build FAIR metadata records for all our datasets/triple store and have them assessed through tools like FUJI
- Ensure provenance is provided out of our Jupyter Notebooks

5. Conclusion

The first release of FAIR-EASE EAL services represents a transformative leap in supporting data-driven research and innovation within the Earth science community. By integrating advanced technologies and addressing real-user needs, the EAL offers researchers powerful tools to analyse complex datasets, foster interdisciplinary collaboration, and drive scientific discovery. The working prototypes and use cases developed during the implementation phase highlight the platform's potential to streamline Earth science analyses and catalyse new approaches to addressing global challenges.

Looking ahead, this initial release is just the beginning. The upcoming deliverable D3.3, for the final release of the Earth Analytics Lab, will bring enhanced integration of all components, including a comprehensive resources catalogue and enriched learning materials. These improvements will ensure the EAL is even more accessible and effective for its users. Additionally, D3.3 will deliver EAL implementations ready to be utilised by pilots for specific demonstrators, showcasing the practical applications of the platform.

Overall, the efforts and progress made by the FAIR-EASE project should make it possible to build one or more sustainable platforms, and to integrate new techniques and technologies of interest over time, in order to serve Earth system science.

Funded by
the European Union