Integrating a community

Earth sciences use cases

Samuel Keuchkerian - (CNRS)

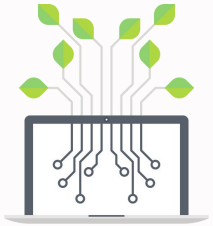
Marie Jossé - Data Terra (CNRS)

FAIR-EASE has received funding from the European Union's Horizon Europe Framework Programme (HORIZON) - under grant agreement No. 101058785.

eosc | FAIR-EASE

FAIR-EASE

Building an interdomain digital architecture for distributed and integrated use of environmental data



FAIR-EASE Data Discovery and Access Interdisciplinary Service

FAIR-EASE Virtual environments

eosc | FAIR-EASE

FAIR-EASE Earth sciences use cases

5 pilots for an earth system model

Coastal Water Dynamics: focuses on the coastal marine environment near river estuaries, where important processes take place.

Earth Critical Zone: monitors land and soil degradation.

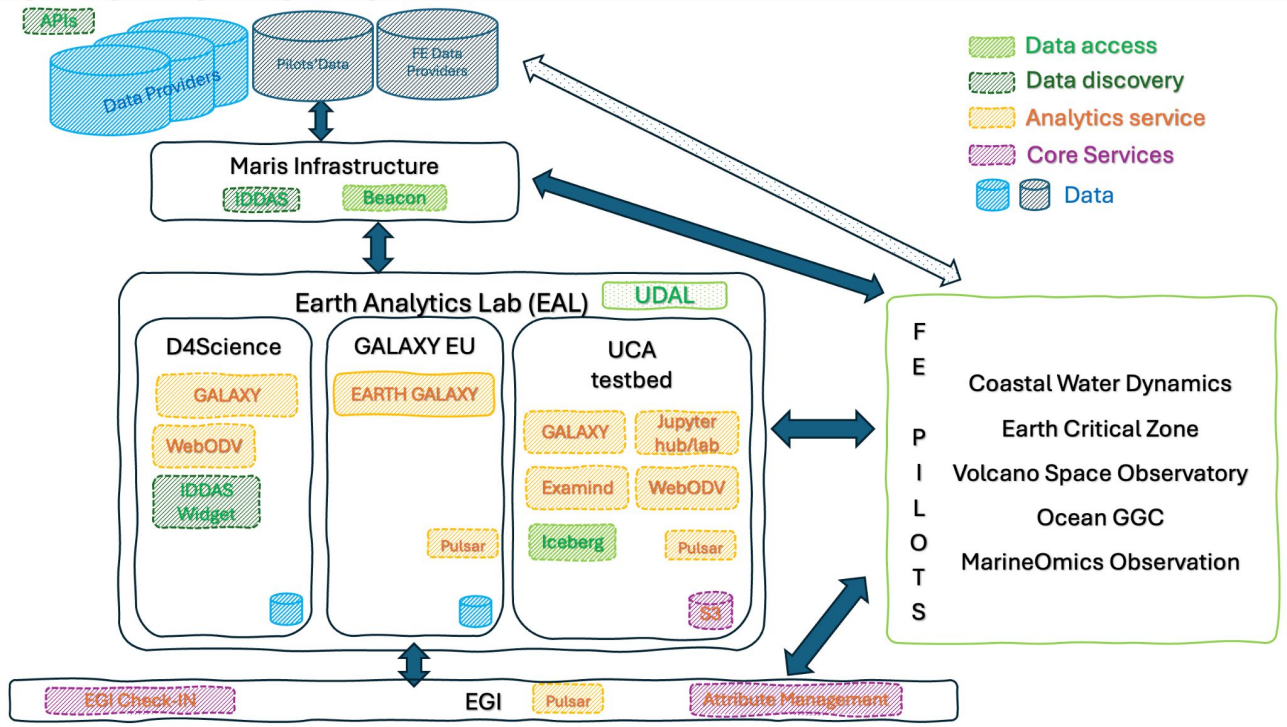
Volcano Space Observatory: monitors global volcanic activity, allowing the focus on any major volcanic eruption worldwide

Ocean Bio-Geochemical Observations: addresses fundamental scientific questions regarding the health of marine ecosystems (e.g. ocean acidification, ...) and needs for ocean resource management.

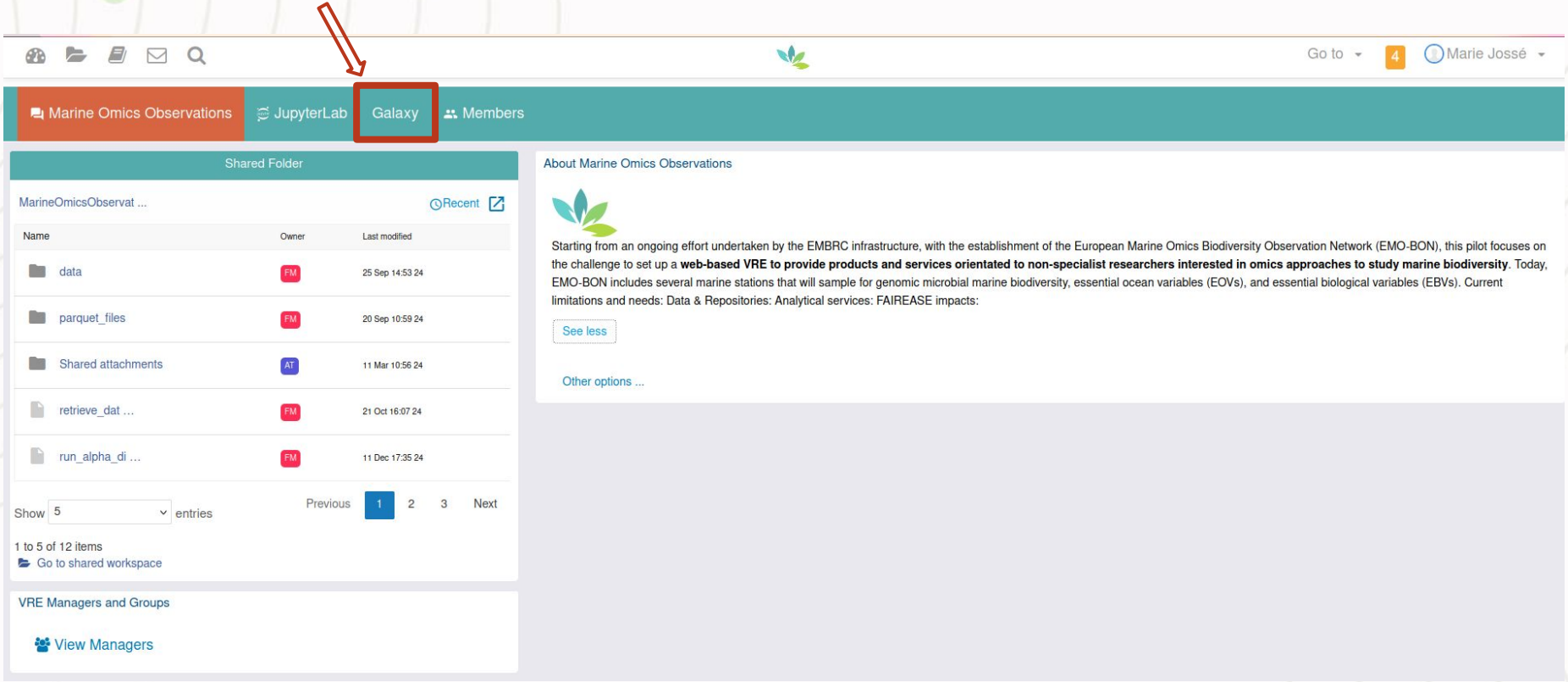
Marine Omics Observation:

analyses of spatial- and time-comparable marine microbial metagenomics data sets for the exploration of biodiversity and its correlations with environmental quality

FAIR-EASE datalake infrastructure



FAIR-EASE Pilots on D4science VLabs

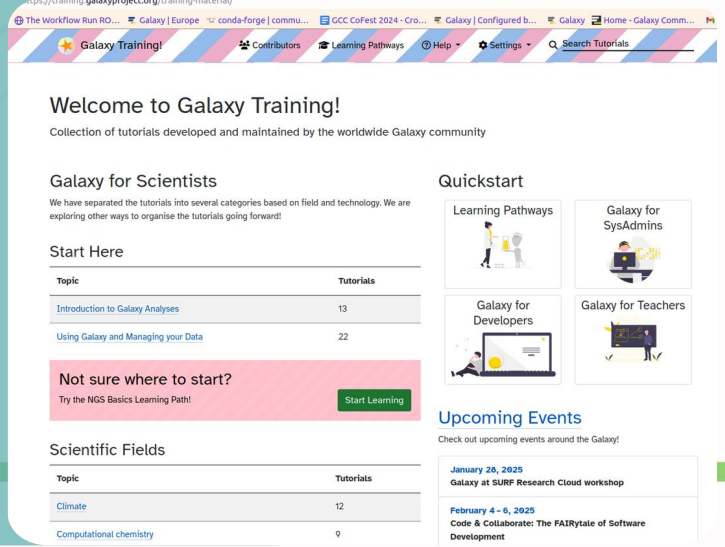


FAIR-EASE Pilots on Galaxy Europe

eosc | FAIR-EASE

Galaxy Training Network

An easy way to learn how to use Galaxy and improve your skills on various domains for instance a set of tutorials are available on FAIR management



- **A catalog of tutorials**
- **Pathways on a dedicated topic**
- **Classes, courses, webinars, and other interactive events**

A community for the community
https://training.galaxyproject.org/training-material/

Galaxy Training Network

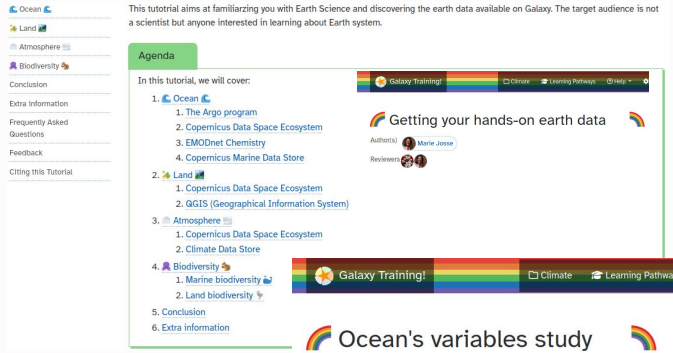
Galaxy Training Network

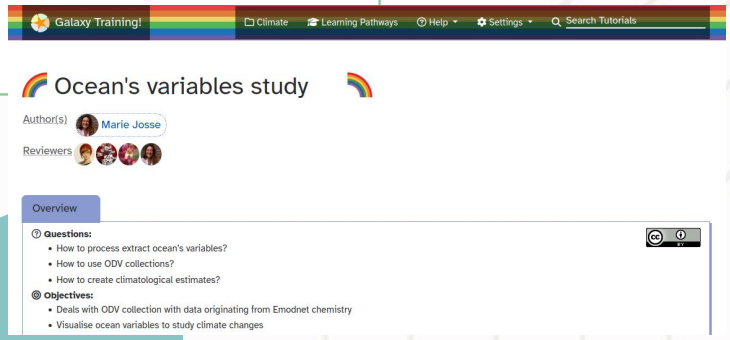
FAIR-EASE contribution (for earth sciences)

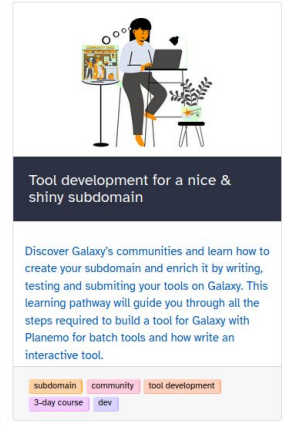
Earth sciences discovery tutorials

Thematic tutorials on ocean, land, atmosphere, and biosphere

Development tutorials end pathways to build a subdomain and a community with Galaxy







Importance of the team in the sustainability of Galaxy

A vibrant team of :

Administrators

Developpers

Trainers

BUT with a strong focus on life sciences

Freiburg Team

Björn Grüning
Dr. rer. nat., Researcher, Head of the Team
gruening@informatik.uni-freiburg.de
bgruening
+49(0) 761 - 203 54130
Build.: 079, Room: -1006

Bérénice Batut
Ph.D., Researcher
bebatut@informatik.uni-freiburg.de
bebatut
+49(0) 761 - 203 54126
http://research.bebatut.fr

Anika Erxleben-Eggenhofer
Dr. rer. nat., Researcher
erxleben@informatik.uni-freiburg.de
erxleben
+49(0) 761 - 203 54130
Build.: 079, Room: -1006

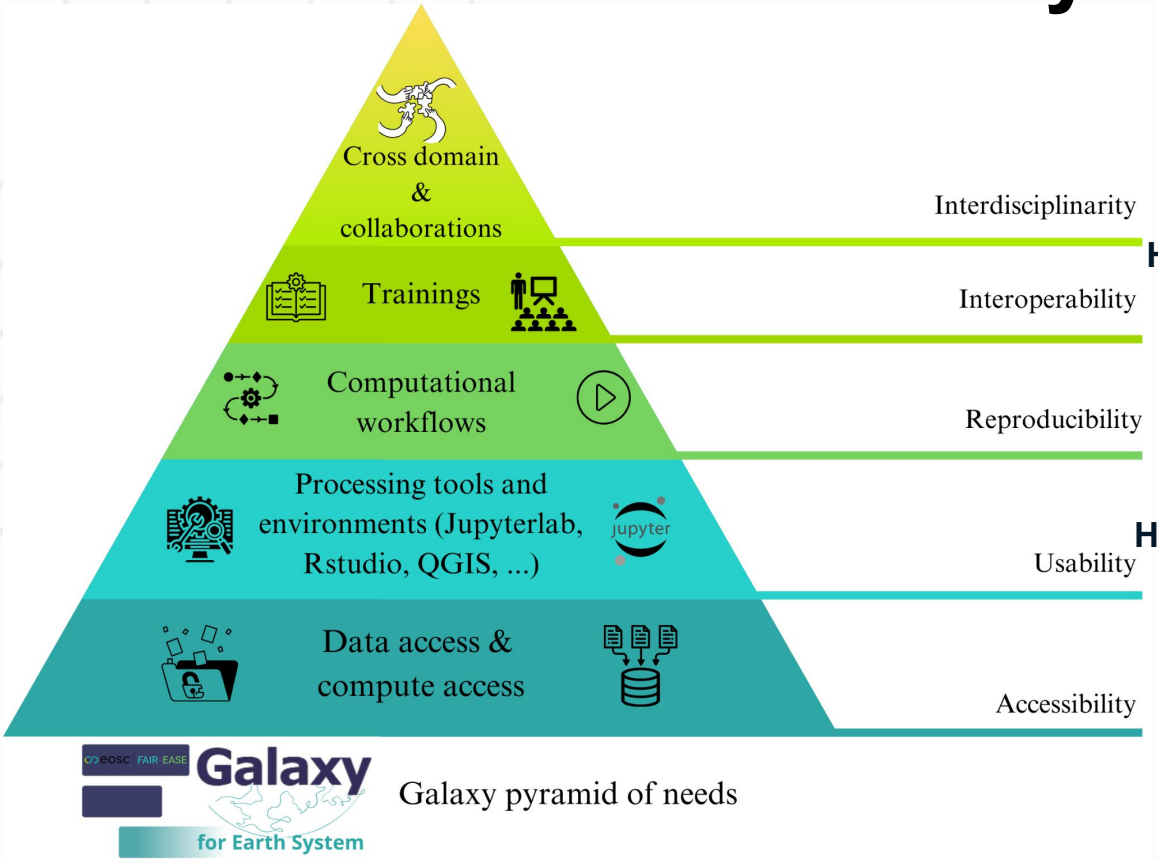
Wolfgang Maier
Dr. phil. nat., Researcher
maierw@informatik.uni-freiburg.de
wm75
+49(0) 761 - 203 54126

Helena Rasche
B.Sc. Biochem., Technician
hxr@informatik.uni-freiburg.de
hexylena
Galaxy Administrator

Paul Zierep
Dr. rer. nat., Researcher
zierep@informatik.uni-freiburg.de
paulzierep
+49(0) 761 - 203 54130

Need the same team for earth sciences

Earth community



Services for EOSC in the proposal for a node (and the EOSC EU node)

French national digital infrastructures

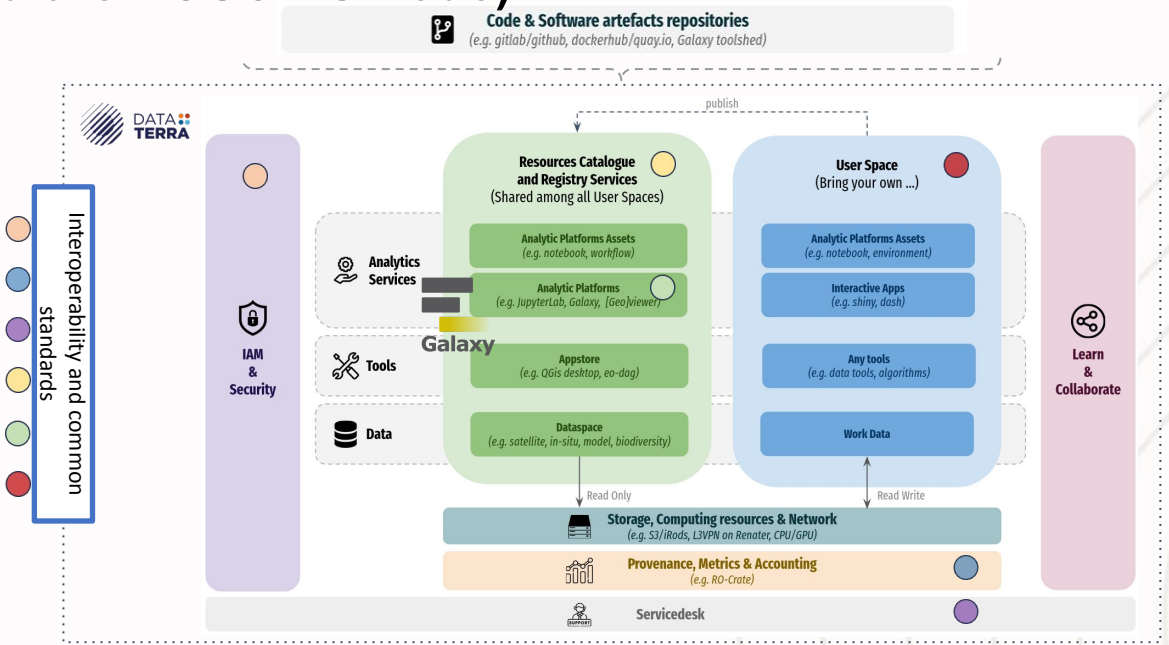
- Renater (Geant);
- France Grille (EGI), & Mesonet > GENCI / EuroHPC: IDRIS, CINES, TGCC;
- National & regional labelled data centres and meso-centres.

EOSC Federation

- D-T data and services accessible through the EOSC EU node;
- Services interoperability with thematic cluster nodes and related national nodes.

Core services

- Distributed data storage and management;
- Large data transfer (files, objects);
- User spaces (interactive notebooks, virtual machine, container images);
- HPC/Cloud computing services
- Federated AAI.



Gaia data timeline allows to co-develop with EOSC EU node and third party (EOSC Federation)

Data Terra services designed to support EOSC EU Node and third party core services

EOSC Federation (EOSC EU node and other nodes)

- Identity Management
- Monitoring and Accounting
- Service Management System
- Resources Catalogues and Registry Services
- Application Workflow Management
- User Space

eosc | FAIR-EASE

Integrated Earth System Observation - Data Terra

Research Infrastructure to access, process and combine data

MAIN MISSION

Develop a global system for accessing and processing observation data (satellite, in situ), value-added products and services to observe, understand and predict in an integrated manner the functioning and evolution of the Earth system.

€42m (2020)	+1000 products & services	+15,000 users	100,000 TB (2022/2023)

Partners

French scientific organizations and universities

5 thematic hubs

- Atmosphere — AERIS
- Ocean — ODATIS
- Solid earth — FORMATER
- Land Surfaces — Theia
- Biodiversity — PNDB (Pôle National de Données de Biodiversité)

Cross-cutting device

- Satellite images — DATATERRA DINAMIS

- A multidisciplinary approach because it calls on work in several areas of Earth System sciences
- An inclusive project that goes beyond the scientific circle with an approach also oriented towards field actors and participatory data

OUR SERVICES

Facilitating the cross-referencing of observations and the modeling of Earth System data

The IR Data Terra offers services around Earth system observation data. The objective is to provide interoperable and interdisciplinary services at all levels.

Connection with producers of data

1. **Data Access**
2. **Daily production of data**
3. **On-demand analysis and processing**
4. **User support services**

Software sharing, Analysis platform, Model evaluation

DATA TERRA

wednesday session

Data storage and handling

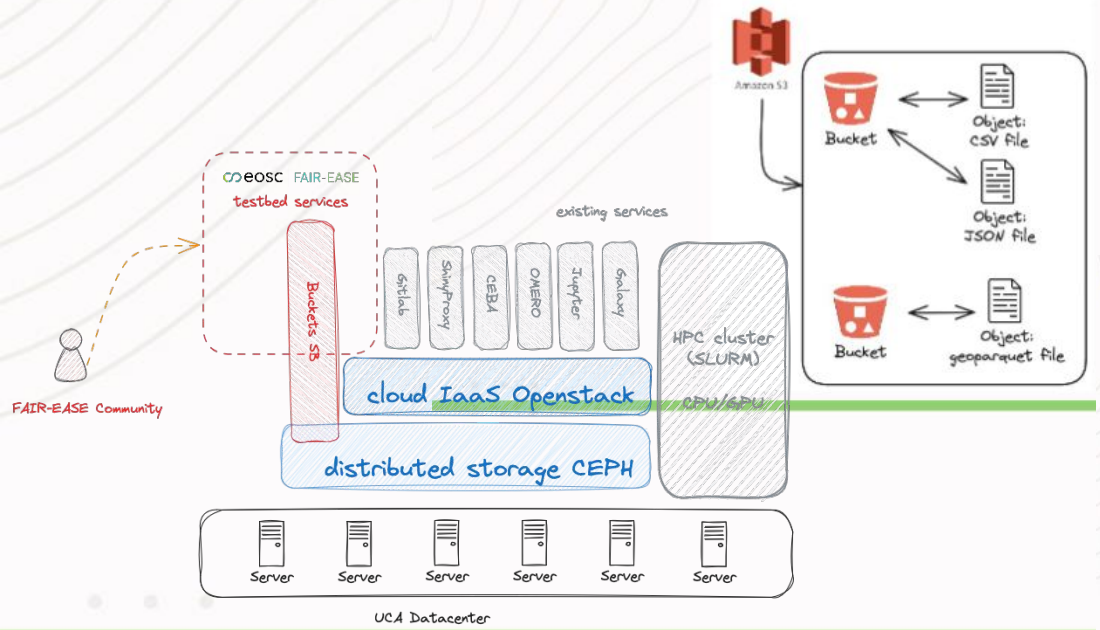
Samuel Keuchkerian - (CNRS/FAIR-EASE)
Marie Jossé - Data Terra (CNRS)

FAIR-EASE has received funding from the European Union's Horizon Europe Framework Programme (HORIZON) - under grant agreement No. 101058785.

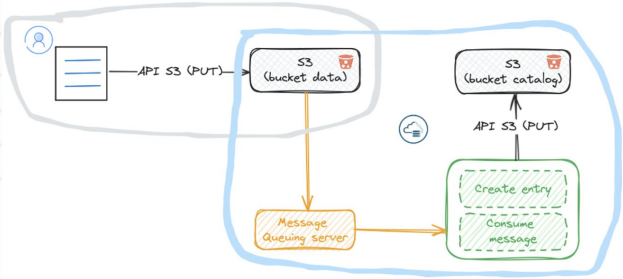
Integrating New Data Storage and Access Paradigms in pilot's practices

"S3 Testbed Tests" Antoine Mahul, David Sarramia, Damian Smyth

S3 concepts reminder

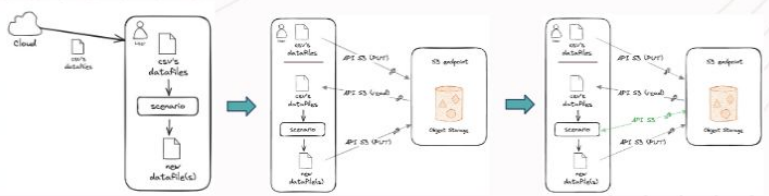


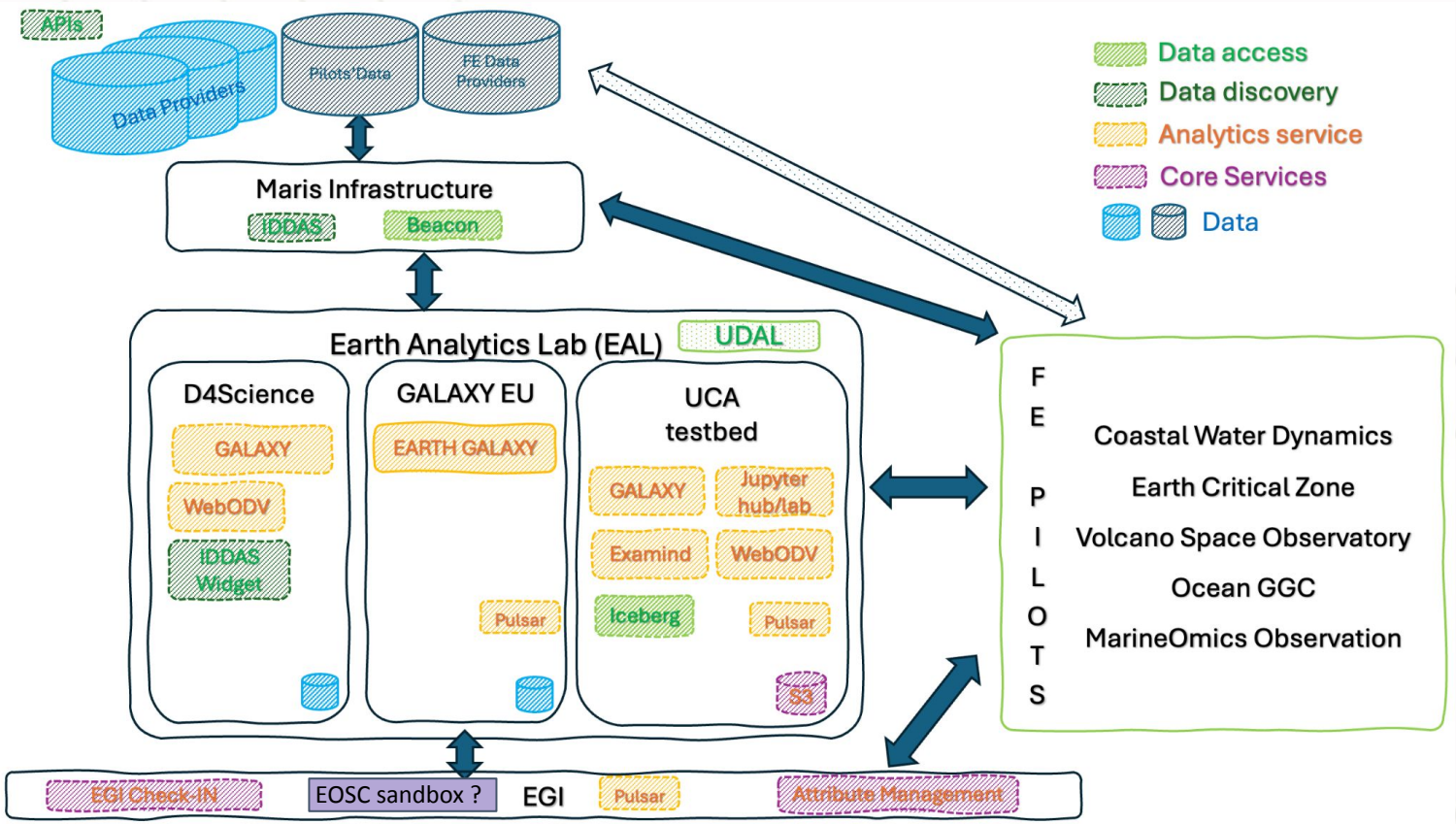
Testing interactions with a static STAC



Can we manipulate a static STAC catalog directly on S3?
Can we use S3 notifications to automatically update a static catalog?
Can we manage metadata automatically from S3 to STAC?

Testing interactions between jupyter notebook, third party softwares and S3




APIs
Data Providers
Pilots' Data
FE Data Providers
Maris Infrastructure
IDDAS
Beacon
Data access
Data discovery
Analytics service
Core Services
Data
Earth Analytics Lab (EAL)
UDAL
D4Science
GALAXY
WebODV
IDDAS Widget
GALAXY EU
EARTH GALAXY
Pulsar
UCA testbed
GALAXY
Jupyter hub/lab
Examind
WebODV
Iceberg
Pulsar
S3
FE PILOTS
Coastal Water Dynamics
Earth Critical Zone
Volcano Space Observatory
Ocean GGC
MarineOmics Observation
EGI Check-IN
EOSC sandbox ?
EGI
Pulsar
Attribute Management

Adding file-sources to Galaxy

Under Development!

This tutorial is not in its final state. The content may change a lot in the next months. Because of this status, it is also not listed in the topic pages.

Marie Josse

Overview

Questions:
- How to set up an S3 bucket

Objectives:
- Add your S3 bucket on Galaxy

Time estimation: 15 minutes
Supporting Materials:
Published: Jan 20, 2025
Last modification: Jan 20, 2025

Revision: 1

Find the information you need
Add the S3 bucket
Conclusion

This tutorial demonstrates how to implement an S3 bucket as a Galaxy file-source within Galaxy. We will add here the public Argo data Amazon S3 bucket. Argo is an international program that observes the interior of the ocean with a fleet of profiling floats drifting in the deep ocean currents (https://argo.ucsd.edu). It started 20 years ago and is a dataset of 5 billion in situ ocean observations from 18.000 profiling floats (4.000 active). The Argo GDAC dataset is a collection of 18.000 NetCDF files. It is a major asset for ocean and climate science and a contributor to IOCCP reports.

Find the information you need
Add the S3 bucket
Conclusion
Frequently Asked Questions
Feedback
Citing this Tutorial

Agenda

In this tutorial, we will cover:

1. Find the information you need
2. Add the S3 bucket
3. Conclusion

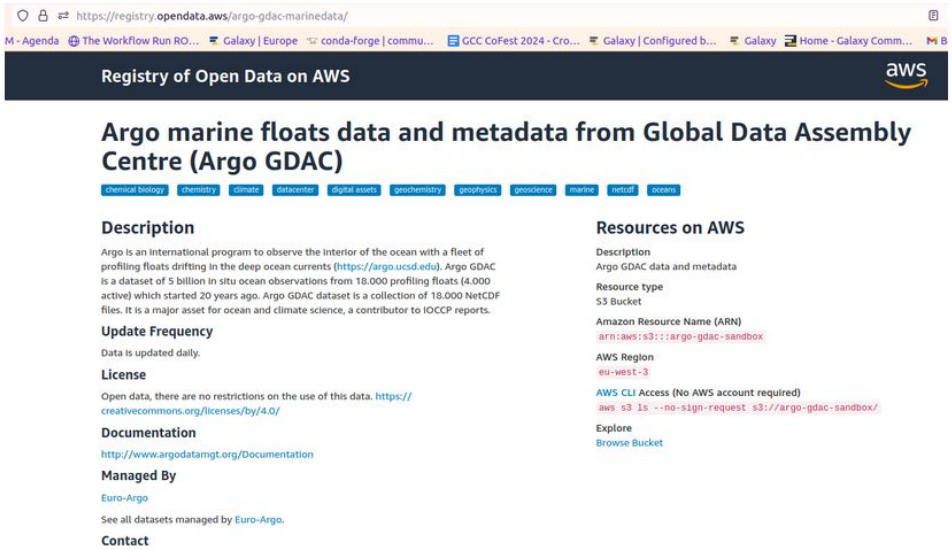
Find the information you need

Hands-on: Find an S3 bucket

Go on Amazon Sustainability Data Initiative.

There you can visit the catalog of data, and by searching for Argo you can directly get to the Argo registry.

On this last page you'll find all the information you'll need to add the S3 bucket to Galaxy



you need
Add the S3 bucket
Conclusion
Frequently Asked Questions
Feedback
Citing this Tutorial

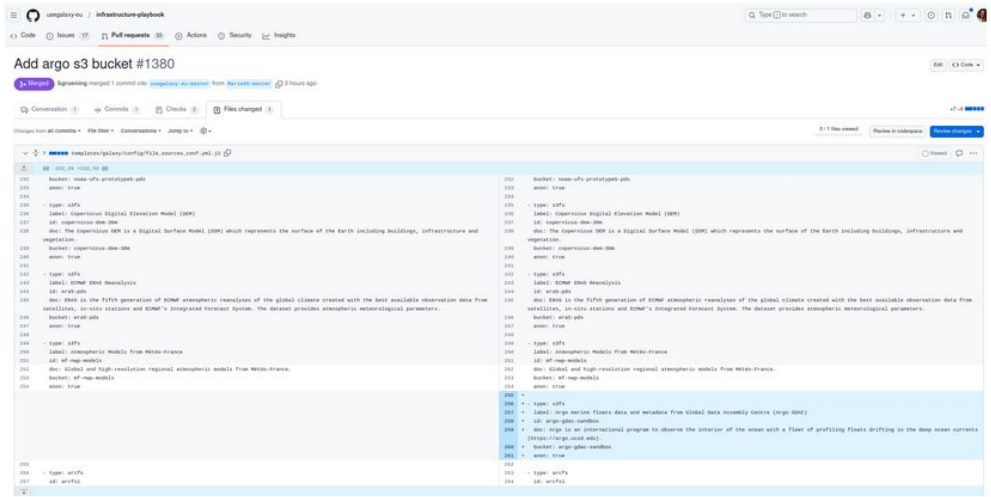
Add the S3 bucket

Hands-on: Add on Galaxy

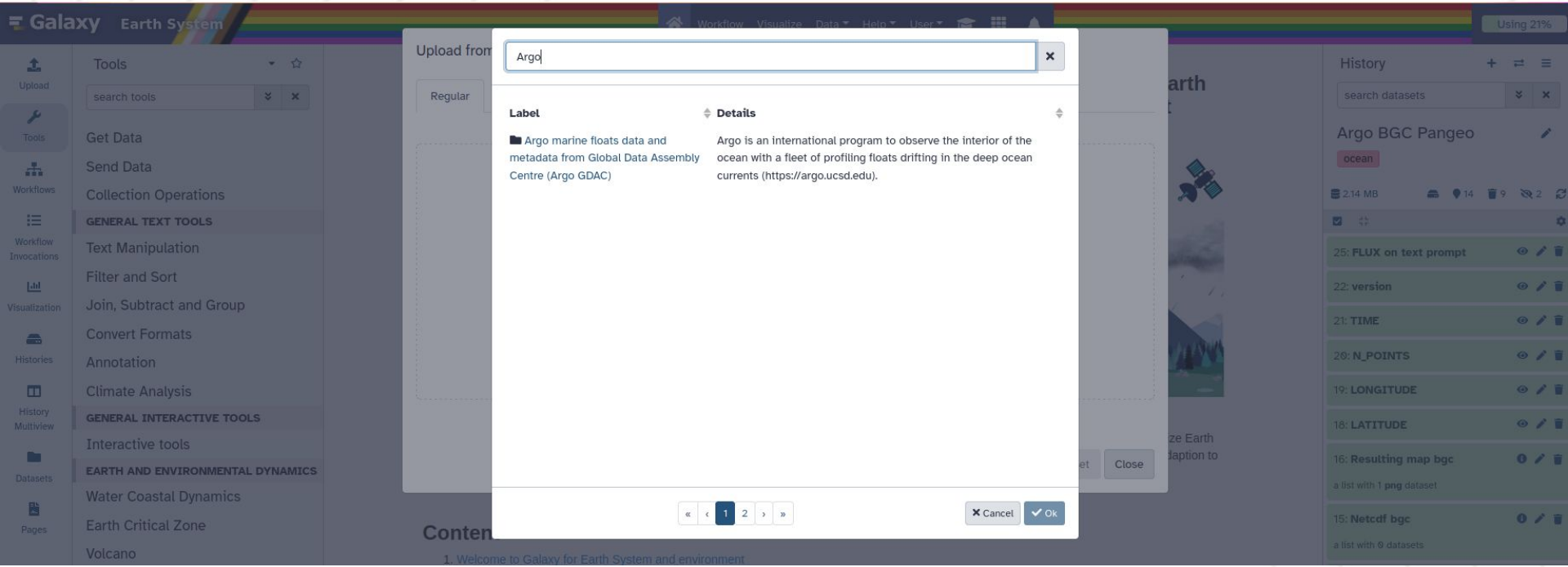
- If not already done clone the Galaxy Europe Infrastructure-playbook repo
- Create a branch on your fork
- Go to the file **file_sources_conf.yml.j2** in **templates/galaxy/config/**

There you can edit the file and add your S3 bucket by adding a Argo specific section, like in the following:

```
- type: s3fs
  label: Argo marine floats data and metadata from Global Data Assembly Centre (Argo GDAC)
  id: argo-gdac-sandbox
  doc: Argo is an international program to observe the interior of the ocean with a fleet of profiling fl
  bucket: argo-gdac-sandbox
  anon: true
```



Finally, commit your changes and write a nice message for the admin when you open your Pull Request.


Galaxy
Earth System
Workflow
Visualize
Data
Help
User
Using 21%
Upload
Tools
Workflows
Workflow Invocations
Visualization
Histories
History Multiview
Datasets
Pages
Tools
search tools
Get Data
Send Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Annotation
Climate Analysis
GENERAL INTERACTIVE TOOLS
Interactive tools
EARTH AND ENVIRONMENTAL DYNAMICS
Water Coastal Dynamics
Earth Critical Zone
Volcano
Upload from
Regular
Close
Argo
Label
Details
Argo marine floats data and metadata from Global Data Assembly Centre (Argo GDAC)
Argo is an international program to observe the interior of the ocean with a fleet of profiling floats drifting in the deep ocean currents (https://argo.ucsd.edu).
1
2
Cancel
Ok
History
search datasets
Argo BGC Pangeo
ocean
2.14 MB
25: FLUX on text prompt
22: version
21: TIME
20: N_POINTS
19: LONGITUDE
18: LATITUDE
16: Resulting map bgc
a list with 1 png dataset
15: Netcdf bgc
a list with 0 datasets
Content
1. Welcome to Galaxy for Earth System and environment

← Back to News

Bring Your Own Storage with Galaxy

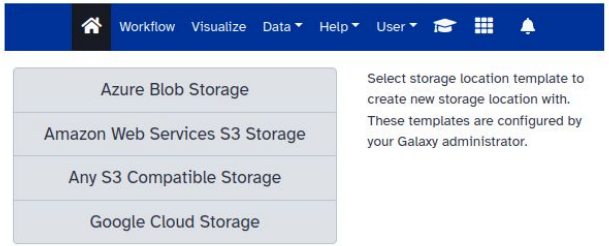
Connect Galaxy with external storage resources easily

September 20, 2024

Introduction

The EuroScienceGateway Project is streamlining the way that users *Bring their Own Storage (BYOS)* to Galaxy. This post covers a specific case in which a user has access to storage resources in the EGI Federated Cloud and wants to connect it to the EU Galaxy instance. However, similar steps can be followed to connect a different Galaxy instance with storage capacity using Azure Blob Storage, Amazon Web Services S3 Storage, and Google Cloud Storage.

In order to connect Galaxy with external cloud storage, the user needs to click on the `User` drop-down menu, select `Preferences` and then `Manage Your Storage Locations`. This will display a list of the cloud storage services already connected to Galaxy. The first time it will be empty and the user can connect to a new cloud storage service clicking on `Create`. The following options are given to the user:



Generally speaking the user needs the following information to connect Galaxy with cloud storage:

1. Name of the bucket
2. Access key
3. Secret key

Getting S3 storage from EGI

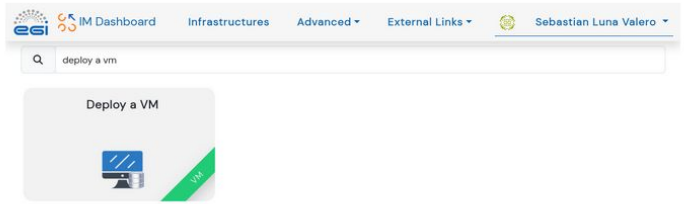
Users with access to cloud resources in the EGI Federated Cloud can benefit from the available TOSCA template in Infrastructure Manager (or IM) for the automated deployment of MinIO.

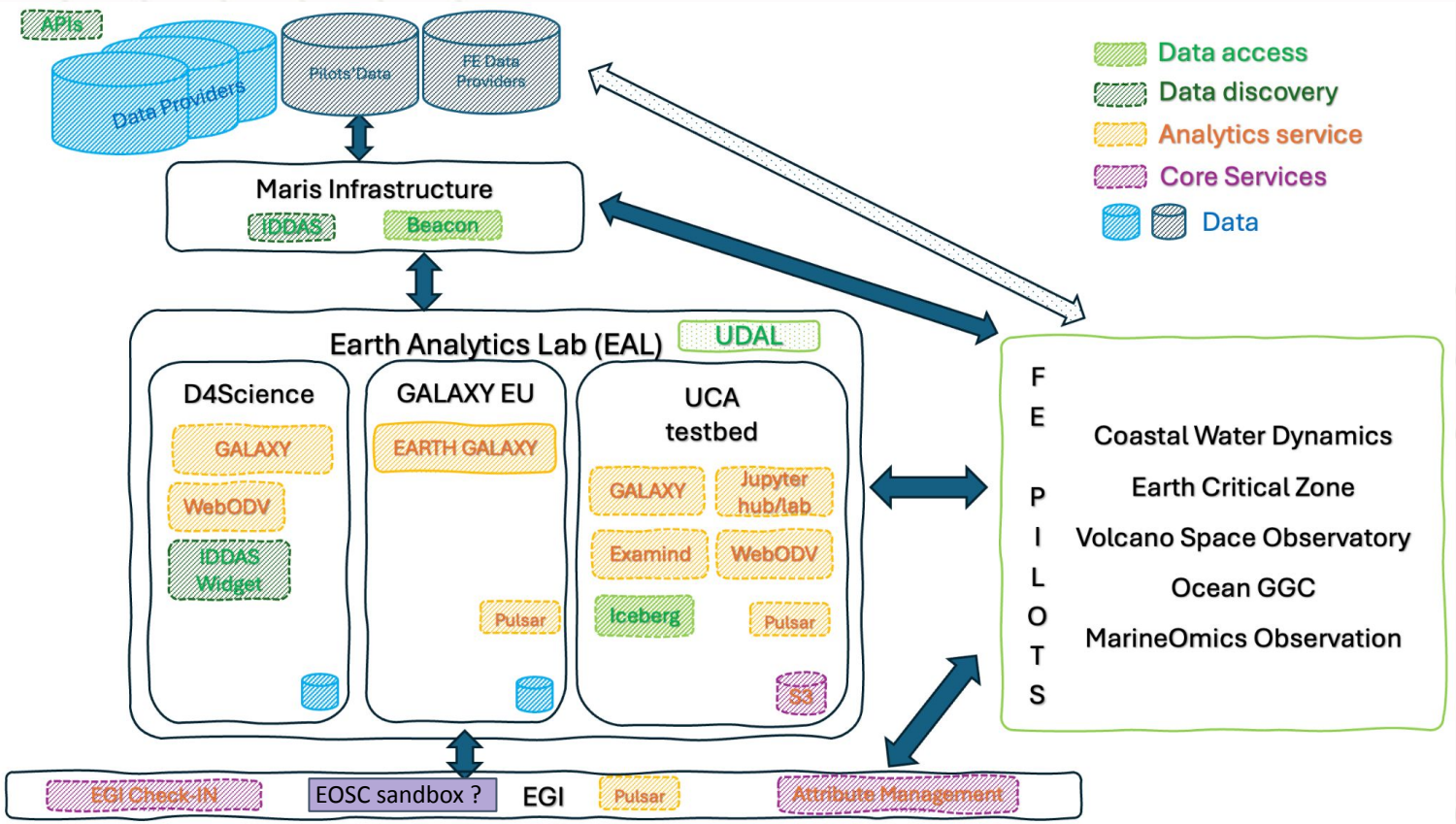
Below are the steps that a user needs to follow to make use of computing resources in the EGI Federated Cloud:

1. Create an EGI Check-in account.
2. Enroll in a Virtual Organization (VO). You need to wait for approval before moving on.
3. Once you are a member of a VO, configure credentials in Infrastructure Manager.
4. In Infrastructure Manager select `Deploy a VM` and click `Configure`.
5. Select `MinIO` and click `Add`.
6. Configure deployment details for MinIO and click `Submit`.

A prerequisite for the deployment of MinIO with IM is the use of some sort of Dynamic DNS service. EGI also offers a Dynamic DNS service for free for users with an EGI Check-in account (see a tutorial for further help). Using a Dynamic DNS service the user needs to register two DNS hostnames: one DNS hostname for the `MinIO Console`, and another one for the `MinIO API endpoint`.

IM will deploy a Virtual Machine with Docker and MinIO as a containerised service. Therefore, the first step for the user is to select `Deploy a VM` in IM:




APIs
Data Providers
Pilots' Data
FE Data Providers
Maris Infrastructure
IDDAS
Beacon
Data access
Data discovery
Analytics service
Core Services
Data
Earth Analytics Lab (EAL)
UDAL
D4Science
GALAXY
WebODV
IDDAS Widget
GALAXY EU
EARTH GALAXY
Pulsar
UCA testbed
GALAXY
Jupyter hub/lab
Examind
WebODV
Iceberg
Pulsar
S3
FE PILOTS
Coastal Water Dynamics
Earth Critical Zone
Volcano Space Observatory
Ocean GGC
MarineOmics Observation
EGI Check-IN
EOSC sandbox ?
EGI
Pulsar
Attribute Management

UCA TESTBED INFRASTRUCTURE

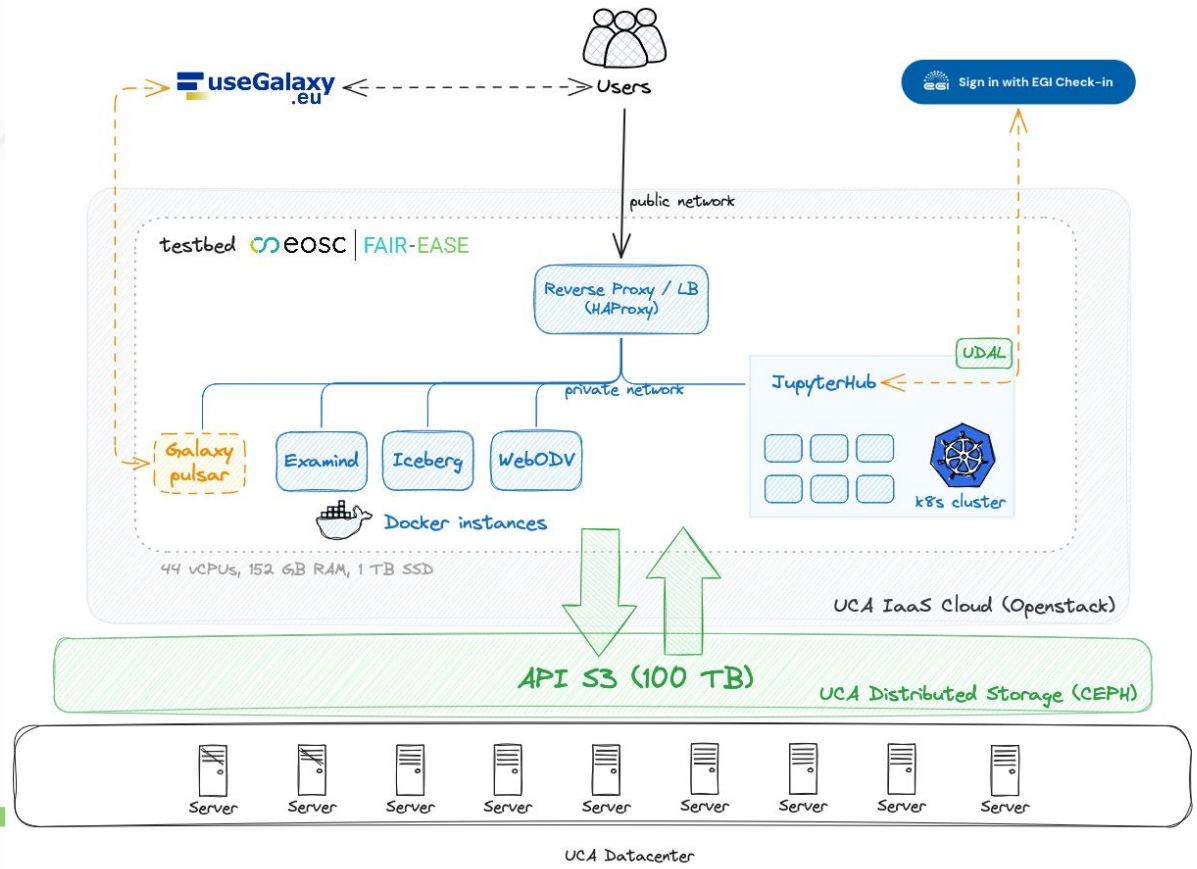
Examind / Iceberg/WebODV effective from this month end.

More precise elements:
Cloud IaaS : Openstack (community version) https://www.openstack.org/
 total : 48 hypervisors, 800 CPU oores, RAM 7.6 TB,

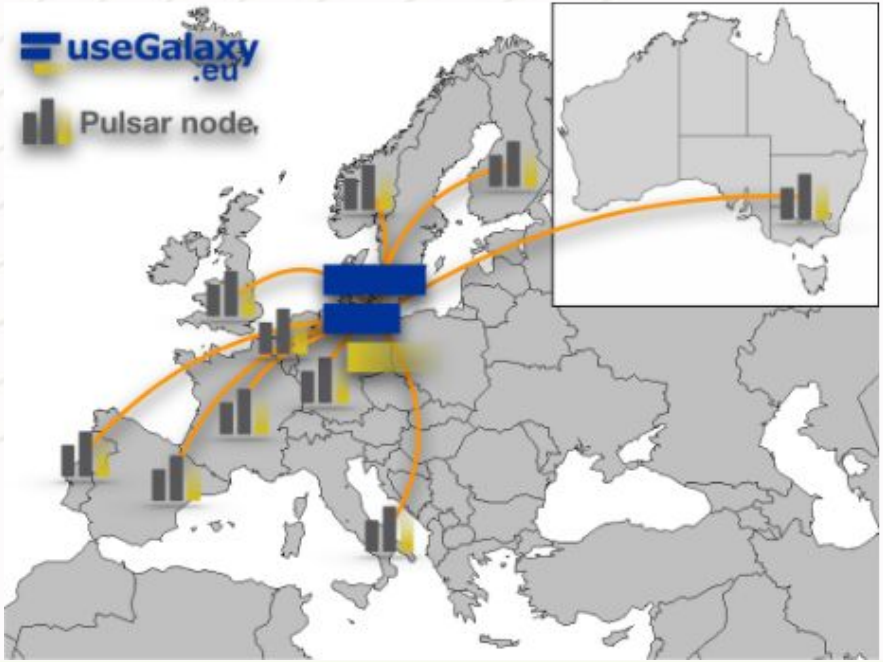
Distributed storage/ S3 : Ceph (community version) https://ceph.io/
 total : 600 TB SSD (for vm openstack), 5.5 PB HDD (S3)
 (replicated 3 times so /3)

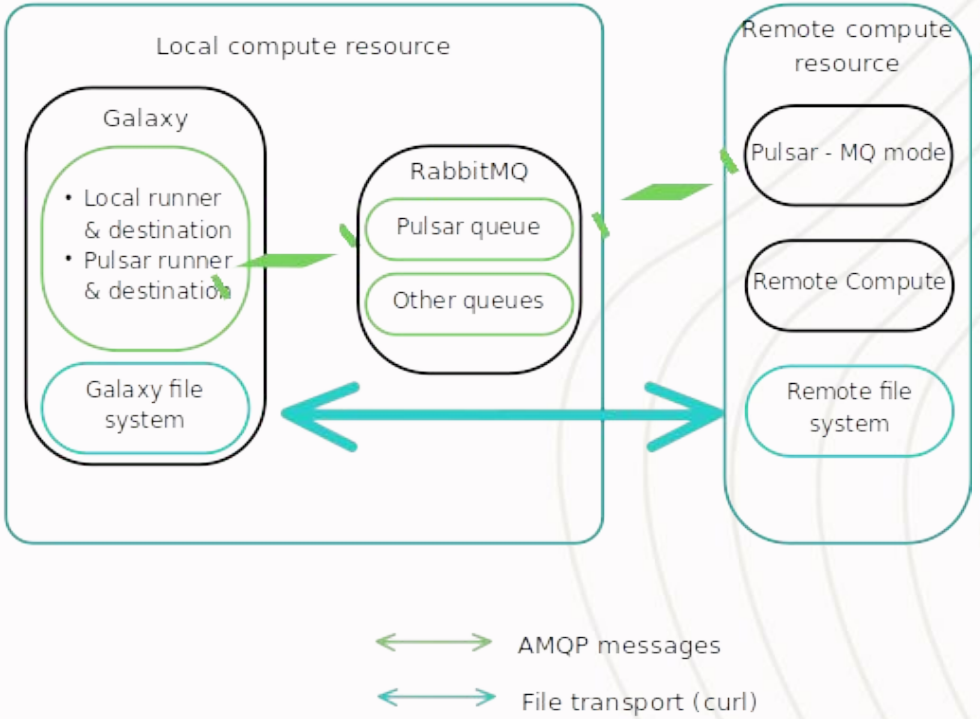
FAIR-EASE specifically allocated resources: 44 CPUs, 152 GB RAM, 128 GB SSD, 100 TB via API S3

Cluster kubernetes FE : kubernetes v1.31, déployed withTerraform et Talos Linux (https://www.talos.dev/)
 3 vm for control plane
 3 vm for workers executed on client services instances (JupyterHub/JupyterLab)



Galaxy deals with remote compute resources and remote files system





eosc | FAIR-EASE

Apache Iceberg an open table format for huge analytic datasets (including metadata)

Iceberg adds tables to compute engines including Spark, Trino, PrestoDB, Flink, Hive and Impala using a high-performance table format that works just like a SQL table. https://iceberg.apache.org/docs/1.5.2/

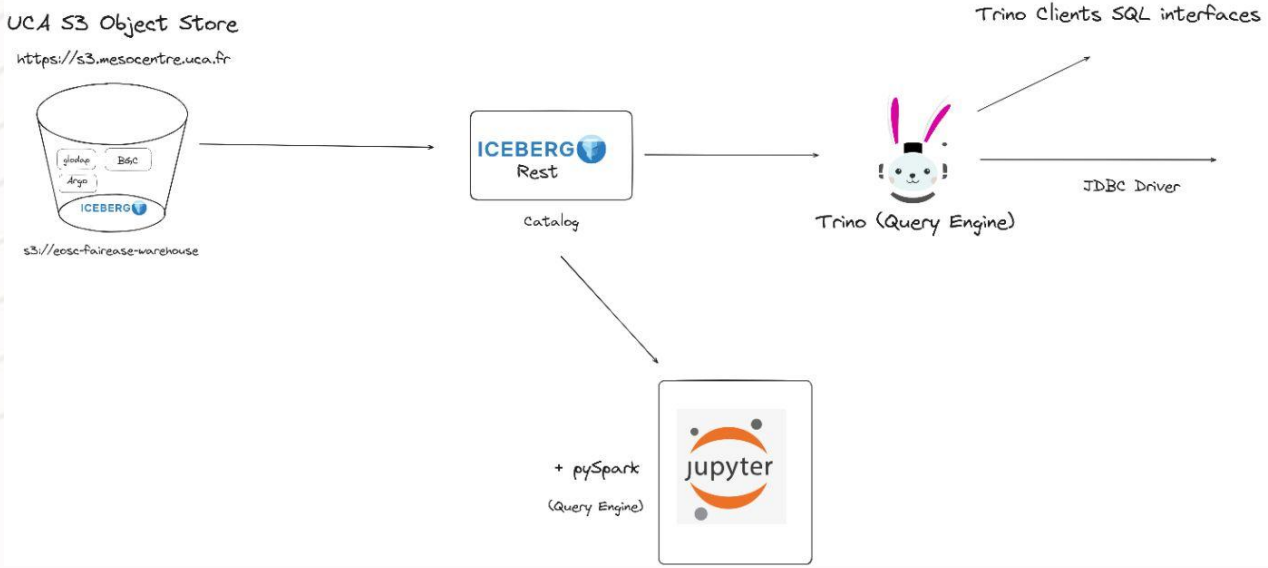
- Iceberg avoids unpleasant surprises. Schema evolution works and won't inadvertently un-delete data. Users **don't need to know about partitioning to get fast queries.**
- Iceberg is used in production where a **single table can contain tens of petabytes** of data and even these huge tables can be read without a distributed SQL engine.
- Open standard designed and developed to be an open community standard with a specification to ensure compatibility across languages and implementations. **Apache Iceberg is open source**, and is developed at the Apache Software Foundation.

Summary

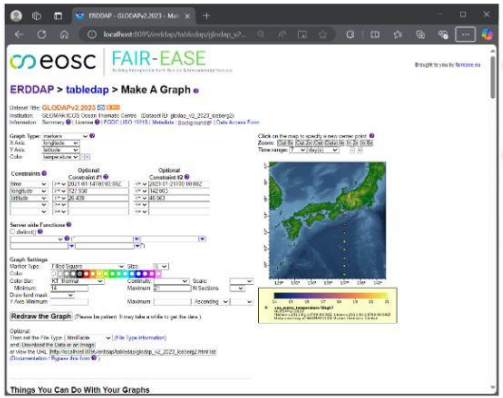
Iceberg excels in decoupling storage from compute, managing evolving schemas, and enabling scalable, FAIR-compliant data workflows, making it ideal for research infrastructures.

For FAIR, open data infrastructures like FAIR-EASE, Iceberg offers flexibility and adaptability while ensuring long-term usability and performance at scale.

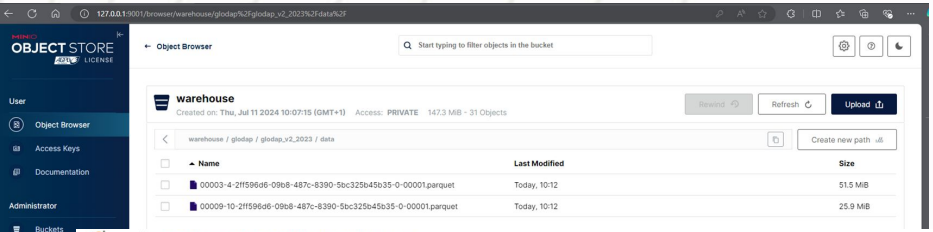
Iceberg to ERDDAP data publication

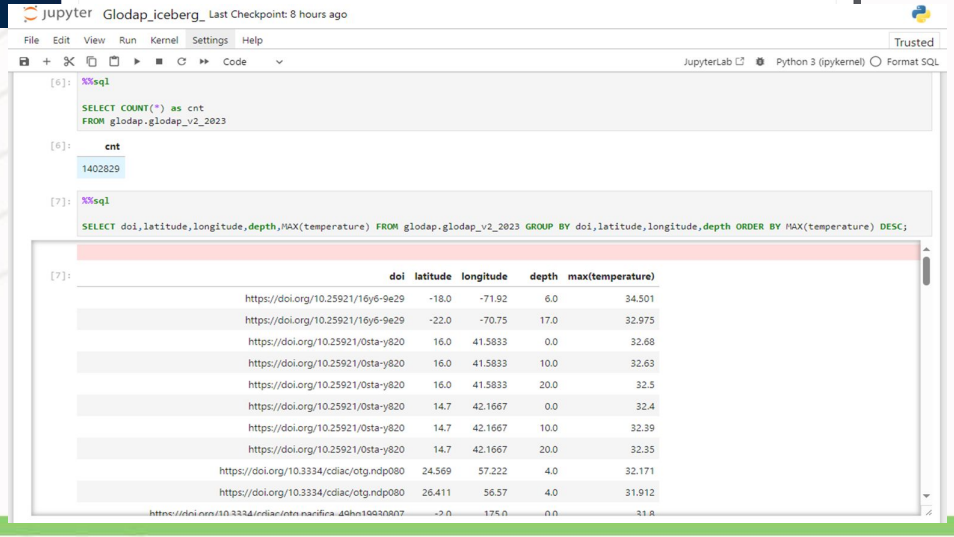


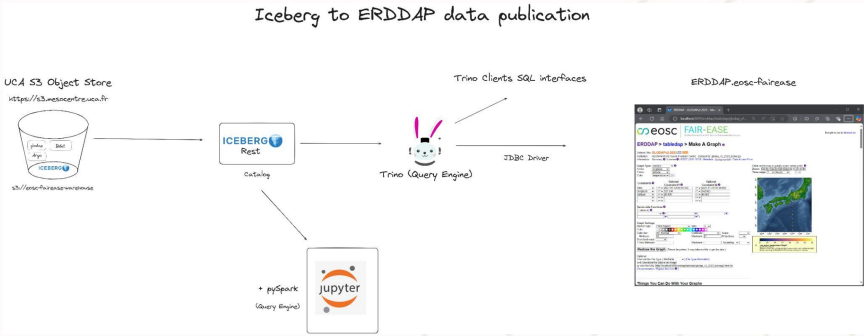
ERDDAP.eosc-fairease



Integrating New Data Storage and Access Paradigms in pilot's practices







"Open Data Lake House Format - Apache Iceberg - Quick start" Damian Smyth:

eosc | FAIR-EASE

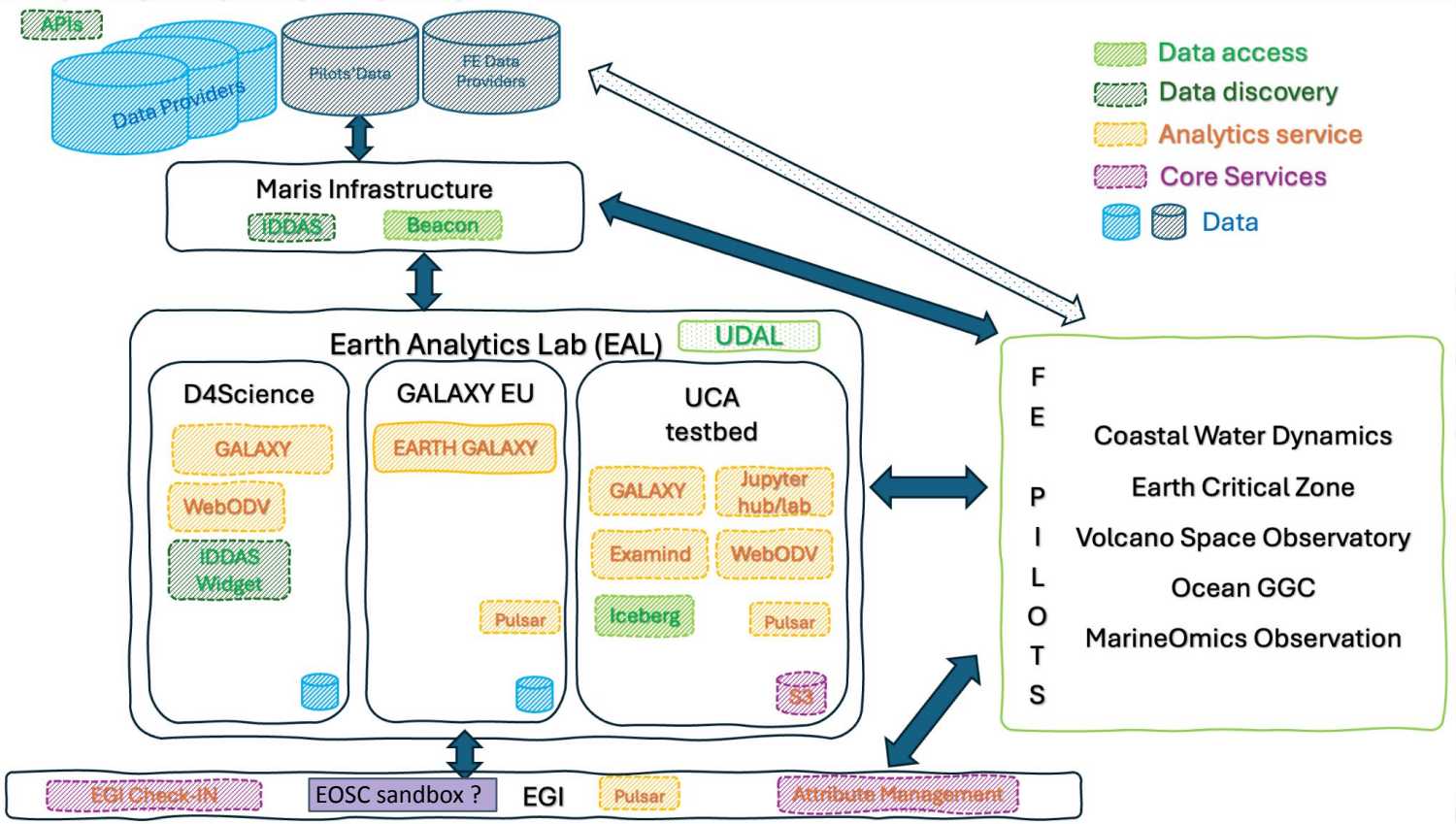
EXAMIND

The Examind provides all the features you need to build a Geographic, Hydrographic, Oceanographic and Meteorologic geospatial data processing infrastructure. Designed for **interoperability**, all the products in the suite comply with **OGC standards**, and integrate OGC Web services

Examind community is a **open-source platform / server.**

- It manages multiple data formats (Netcdf, geotiff, etc.), clouds native data formats (COGs, GIMI), multiple OGC (Open Geospatial Consortium) standards (WCS, WMS, WPS, OGC API, etc.)
- It has a number of ready-to-use processes for various uses in the geospatial world.
- It 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 (via WPS, OGC API Process, openEO); all via different paradigms, such as data cube structuring.

Examind supports several data storage options local storage, FTP, S3 (AWS / Minio), via HTTP/HTTPS, or storage from a WMS / WMTS service, etc.
The service can be deployed via a docker image and a docker-compose file on kubernetes infrastructures.


APIs
Data Providers
Pilots' Data
FE Data Providers
Maris Infrastructure
IDDAS
Beacon
Data access
Data discovery
Analytics service
Core Services
Data
Earth Analytics Lab (EAL)
UDAL
D4Science
GALAXY
WebODV
IDDAS Widget
GALAXY EU
EARTH GALAXY
Pulsar
UCA testbed
GALAXY
Jupyter hub/lab
Examind
WebODV
Iceberg
Pulsar
S3
FE PILOTS
Coastal Water Dynamics
Earth Critical Zone
Volcano Space Observatory
Ocean GGC
MarineOmics Observation
EGI Check-IN
EOSC sandbox ?
EGI
Pulsar
Attribute Management

Summary

Iceberg excels in decoupling storage from compute, managing evolving schemas, and enabling scalable, FAIR-compliant data workflows, making it ideal for research infrastructures.

For FAIR, open data infrastructures like FAIR-EASE, Iceberg offers flexibility and adaptability while ensuring long-term usability and performance at scale.

eosc | FAIR-EASE

eosc | FAIR-EASE

WEDNESDAY AFTERNOON

FAIR-EASE in the context of the European green deal


Mobilising research and fostering innovation
Transforming the EU's economy for a sustainable future
Increasing the EU's Climate ambition for 2030 and 2050
A zero pollution ambition for a toxic-free environment
Supplying clean, affordable and secure energy
The European Green Deal
Preserving and restoring ecosystems and biodiversity
Mobilising industry for a clean and circular economy
From 'Farm to Fork': a fair, healthy and environmentally friendly food system
Building and renovating in an energy and resource efficient way
Accelerating the shift to sustainable and smart mobility
Financing the transition
Leave no one behind (Just Transition)
The EU as a global leader
A European Climate Pact

FAIREASE in the context of the european green deal

European green deal topics

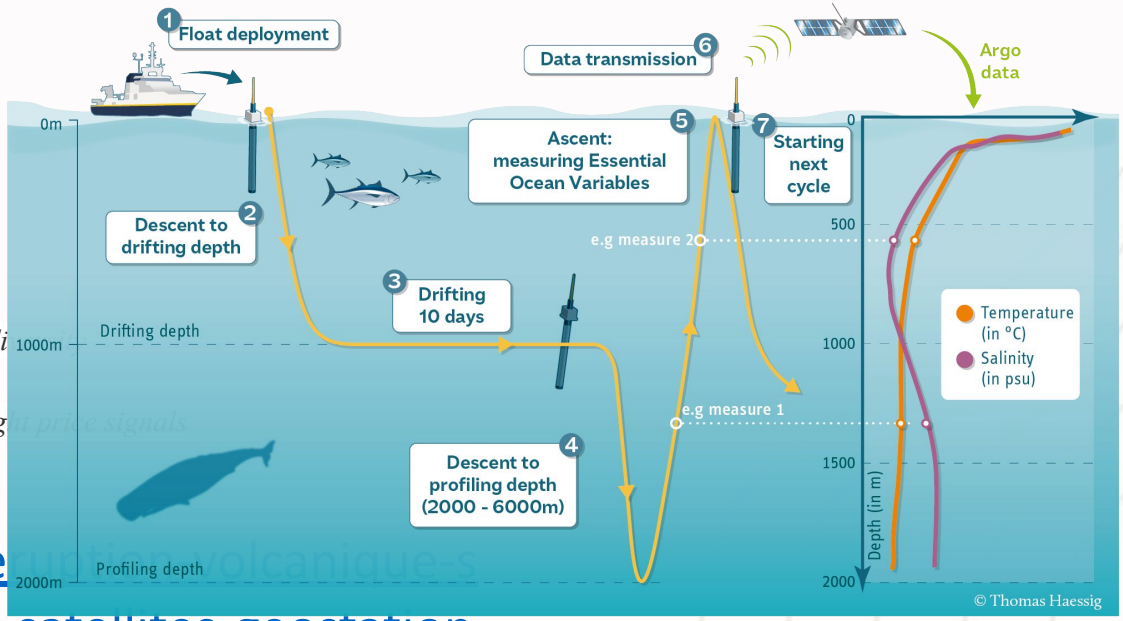
- **Designing a set of deeply transformative policies**
- **Mainstreaming sustainability in all EU policie**

- *2.1.7.*
- *Preserving and restoring ecosystems and biodiversity*
- *2.2.2.*
- *Greening national budgets and sending the right price signals*
- *2.2.3.*
- *Mobilising research and fostering innovation*

- *2.1.7.*
- *Preserving and restoring ecosystems and biodiversity*
- *2.2.2.*
- *Greening national budgets and sending the right price signals*
- *2.2.3.*
- *Mobilising research and fostering innovation*

https://www.aeris-data.fr/leruption-volcanique-sur-les-iles-tonga-vue-par-les-satellites-geostationnaires/

- *2.1.7.*
- *Preserving and restoring ecosystems and biodi*
- *2.2.2.*
- *Greening national budgets and sending the right price signals*
- *2.2.3.*
- *Mobilising research and fostering innovation*



https://www.aeris-data.fr/le[illegible]volcanique-sur-les-iles-tonga-vue-par-les-satellites-geostationnaires/