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

D3.1 - Specifications of FAIR-EASE Earth Analytics Lab and implementation plan

Work Package	**WP3, Earth Analytics Lab**
Lead Authors (Org)	**Erwan Bodéré (Ifremer)**
Contributing Author(s) (Org)	**Reiner Schlitzer (AWI), Jérôme Detoc (Ifremer), Dorian Ginane (Geomatys), Marie Josse (CNRS), Cymon J. Cox (CCMAR/EMBRC.PT), Charles Troupin (ULiege), Simona Simoncelli (INGV), Claudia Fratianni (INGV), Nicolas Pascal (CNRS), Maria-Luisa Chiusano (UNINA), Christelle Pierkot (CNRS), Marine Vernet (Ifremer), Jean-François Piollé (Ifremer), Joël Sudre (CNRS), David Sarramia (UCA)**
Due Date	**31.08.2023**
Date	**26.10.2023**
Version	**V1.0**

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 ID	Notes
0.1	27.06.2023	Erwan Bodéré (Ifremer)	ORCID	ToC and V0.1
0.2	13.07.2023	Erwan Bodéré (Ifremer)	ORCID	Write chapter 2, 3, 4, 5
0.3	21.08.2023	Dorian Ginane (Geomatys)		Added Phidias project feedback, Examind description
0.4	22.08.2023	Marine Vernet (Ifremer)		Add Pillar project feedback
0.5	26.09.2023	Erwan Bodéré (Ifremer)	ORCID	New diagrams (design, content)
0.6	29.09.2023	Cymon J. Cox (CCMAR/EMBRC.PT) Maria Luisa Chiusano (UNINA) Jean-François Piollé (Ifremer)	ORCID	Revision of all text
0.7	01.10.2023	Erwan Bodéré (Ifremer)	ORCID	Adapt chapter 4 according to new diagram
0.8	02.10.2023	Reiner Schlitzer (AWI)	ORCID	Added WebODV description
0.9	02.10.2023	Christelle Pierkot (CNRS)	ORCID	Added collaboration part, FAIRness, revision of the document
0.10	02.10.2023	Simona Simoncelli (INGV)	ORCID	Comments/su to the general structure of the deliverable
0.11	02.10.2023	Marie Josse (CNRS) Jerome Detoc (Ifremer)		Added Galaxy description
0.12	03.10.2023	Erwan Bodéré (Ifremer)	ORCID	Complete chapter 4
0.13	04.10.2023	Charles Troupin (ULiege)	ORCID	General revision
0.14	05.10.2023	Erwan Bodéré (Ifremer)	ORCID	Write chapters 5 & 6
0.15	11.10.2023	Claudia Fratianni (INGV)	ORCID	General review
0.16	12.10.2023	Reiner Schlitzer (AWI)	ORCID	Added draft Introduction, revised 4.2.4, and 6

0.17	13.10.2023	Erwan Bodéré (Ifremer)	ORCID	Review Reiner inputs Update 4.2.3
0.18	16.10.2023	Erwan Bodéré (Ifremer)	ORCID	General Review
0.19	23.10.2023	Joël Sudre (CNRS)		Internal Review
0.20	25.10.2023	David Sarramia		Internal Review
0.21	25.10.2023	Erwan Bodéré (Ifremer)	ORCID	Consideration of internal review (reduce chapter 2)
1.0	26.10.2023	Erwan Bodéré (Ifremer)	ORCID	Final Revision
Final	26.10.2023	Erwan Bodéré (Ifremer)	ORCID	Final edition for submission

Table of Contents

Table of Figures

Terminology

Terminology/Acronym	Description
API	Application Programming Interface
ARCO	Analysis Ready, Cloud Optimised
BGC	Bio-Geochemical
CCP	Cloud Computing Platform
CWL	Common Workflow Language
EAL	Earth Analytics Lab
EOSC	European Open Science Cloud
FAIR	Findable; Accessible; Interoperable; Reusable
IAM	Identity and Access Management
IDDAS	Interdisciplinary Data Discovery and Access Service
IDE	Integrated Development Environment
M2M	Machine To Machine
MVP	Minimum Viable Product
NGO	Non-Governmental Organisations
OGC	Open Geospatial Consortium
RO-CRATE	Research Object Crate
TRL	Technology Readiness Level
VDAP	Virtual Data Analytic Platform
VRE	Virtual Research Environment
VNC	Virtual Network Computing
UC	Use Case
UDAL	Unified Data Access Layer
UI	User Interface
WPS	Web Processing Service
WMS	Web Map Service

Executive Summary

The overall objective of the FAIR-EASE project is to customise and operate integrated and distributed services for observation and modelling of the Earth System, environment, and biodiversity by improving the technology readiness level (TRL) of their different components implemented in close cooperation with user communities, the EOSC, and research infrastructures in their design and sustainable availability.

Among these services, an Earth Analytical Lab (EAL), with EOSC connectivity, through web-based interfaces, predefined processing tools and on-demand data visualisation services for remote analysis and processing of heterogeneous data facilitating cross-disciplinary collaboration, reducing the time to produce results, and increasing efficiency will be set-up in the WP3.

This report outlines the specifications of the EAL, including details about its main modules, a macroscopic implementation plan, and collaborations with other EOSC projects.

1. Introduction

The study of the Earth System and the documentation and understanding of natural or man-made changes on various scales require large, multi-disciplinary datasets from different data streams and observation technologies covering all compartments of our environment. Most of the time, researchers require not just a single data source, but need to exploit many multi-disciplinary datasets simultaneously to make progress.

The FAIR-EASE use cases exemplify this need for multi-disciplinary datasets in areas such as land-ocean interaction, ocean biogeochemical cycles, land use change, volcanic activity, biodiversity, and climate change in general. In all these research areas it is critical to combine in-situ data with satellite observations as well as model simulations. The FAIR-EASE Earth Analytics Lab described in this document addresses this fundamental need aiming to facilitate the use and exploitation of multi-disciplinary datasets by providing web-based interfaces, processing tools, interactive analysis and visualisation tools.

1.1. References

- D5.1 - Report on key requirements from Use Cases / Pilots (https://zenodo.org/record/7588904)
- D2.1 - Environmental Data Infrastructures : services description report (https://zenodo.org/record/7920551)
- D4.1 - Landscaping exercise : the (meta)data, software, and cloud needs for the data lake (https://zenodo.org/record/7965398)
- D4.2 - Landscaping exercise : the inclusion of special use-case datasets in the data lake (https://zenodo.org/record/7957747)
- MS06 - FAIR-EASE initial evaluation of the FAIRness of digital resources and services

2. Context

2.1. Project ambition

2.1.1. Promoting open science

Open Science is based on the premise that scientific knowledge should be freely available and shared with the global community. The principles of Open Science include:

- Open Access: openly providing free access to scholarly articles, research papers, and other scientific publications to the public, by removing paywalls and subscription barriers.
- Open Data: sharing research data and making it freely available for others to analyse, reproduce, and build upon. This promotes transparency, reproducibility, and the potential for new discoveries, and promotes integration of diverse data sources
- Open Collaboration: encouraging scientists to collaborate and share their findings, methodologies, and resources openly. This fosters interdisciplinary research and accelerates scientific progress.
- Open Source: releasing software, algorithms, and tools used in scientific research under open source licences, allowing others to use, modify, and distribute them freely.

The FAIR principles prescribe that data should be Findable, Accessible, Interoperable, and Reusable (see The FAIR Guiding Principles for scientific data management and stewardship). While the FAIR principles were initially formulated with a focus on data management and sharing, they have been extended to include services (see Framework for assessing FAIR Services) and analytics workflows (see FAIR Computational Workflows) as well. This expansion recognises that the accessibility and interoperability of scientific services are crucial for advancing research and enabling reproducibility.

When applied to research software related to the EAL, the FAIR principles can be understood as follows:

- Findable: services should have clear and unique identifiers, making them easily discoverable and citable. Metadata about the service, including its purpose, functionality, input/output requirements, and associated resources, should be provided.
- Accessible: services should be openly available to users, preferably with clear usage policies and appropriate access mechanisms. Accessibility may involve providing access to the service through web interfaces, APIs (Application Programming Interfaces), or other means, ensuring that users can utilise the service without unnecessary barriers.
- Interoperable: services should be designed to be interoperable with other services, platforms, or tools. They should follow common standards and specifications, allowing seamless integration and interaction with other services. This enables the composition and chaining of services to create more complex workflows or analysis pipelines.

- Reusable: services should be well-documented, providing comprehensive descriptions, instructions, and examples for their use. Documentation should include information about the service's versions, inputs, outputs, parameters, dependencies, and any associated data or resources. This facilitates the reuse of services by others and ensures transparency and reproducibility.

Applying the FAIR principles to services helps to enhance their discoverability, accessibility, interoperability, and reusability. This promotes collaboration, facilitates the combination of services into complex workflows, and supports the development of more efficient and transparent research processes.

By promoting Open Science and FAIR principles, FAIR-EASE contributes to a more transparent and inclusive scientific ecosystem, fostering innovation, accelerating research, and maximising the societal impact of scientific discoveries. The EAL will facilitate collaborative work by scientific teams.

2.1.2. Allow multi and cross domain analysis

The Earth System encompasses various interconnected components, including the atmosphere, hydrosphere (water), biosphere (living organisms), lithosphere (land), and anthroposphere (human activities). Each of these domains has its own sets of data, measurements, models, and methodologies.

Cross-domain interactions aim to break down the traditional boundaries between these domains enabling scientists researchers to inherit, exchange and explore different approaches, feedback loops, and complex relationships that exist within and between different Earth System components, sharing their specific expertises. By integrating data and knowledge in the same place from multiple domains, scientists can gain a more comprehensive understanding of Earth's functioning and the impacts of various processes.

Here are a few examples:

- Atmosphere-Ocean Interactions: investigating the impacts of long-term meteorological patterns (e.g., trade winds, westerlies) as well as short-term weather phenomena (e.g., hurricanes, droughts) on ocean circulation, greenhouse gas cycle, nutrient cycles and biological productivity. This requires combining meteorological data with ocean in-situ observations as well as satellite data to understand the connections and impacts.

- Land-Ocean Interactions: examining the influence of land processes, such as deforestation or agricultural practices, on the ocean ecosystem. This involves analysing land-use data, nutrient runoff, ocean temperature, and biodiversity data to assess the impact of terrestrial activities on marine ecosystems.

- Earth-Atmosphere Interactions: studying the interactions between land surface properties, atmospheric dynamics, and weather patterns. By integrating data from remote sensing, atmospheric measurements, and surface observations, scientists can explore how changes in land cover or land use affect local and regional weather patterns.

- Solid earth-atmosphere interactions: investigating the phenomena that couple atmosphere and solid-earth systems. For instance, volcanic eruptions have their source in the solid earth but release huge amounts of gases and aerosols in the atmosphere. Studying this kind of events needs to access, process and analyse data issued of land based and satellite remote sensing observations. Of course, the type of satellite instruments that are used for solid earth and atmosphere study are very different, but they have in common to generate an important volume of data that needs to be accessed and processed efficiently for interactive analysis.
- Human, environment, natural systems interactions: understanding earth erosions and relationships with anthropogenic or environmental effects. This includes analytics for recovery and sustainability.

By addressing multi and cross domain analysis, FAIR-EASE aims to provide virtual environments for holistic understanding of the Earth System, considering the interconnections, feedback loops, and interdependencies between its various components. This integrated approach will enable scientists to tackle complex environmental challenges, develop more accurate models, and provide insights for informed decision-making and sustainable management of Earth's resources.

2.1.3. Improve community applications TRL

The TRL scale provides a common language and framework for assessing the readiness of technologies, enabling decision-makers, investors, and researchers to evaluate and track the progress of innovations throughout the development cycle. It helps in determining the level of risk, resource allocation, and further steps required to advance the technology towards widespread implementation or commercialisation.

Figure 1 - Technology Readiness Level scale[1]

[1] Credits: https://www.twi-global.com/technical-knowledge/faqs/technology-readiness-levels

The TRL scale typically ranges from 1 to 9, with each level representing a specific stage in the technology development process. The aim of FAIR-EASE is to rely on services with a high TRL (from 7 to 9) in order to provide more reliable, efficient, and user-friendly tools that can better address the needs and challenges of the communities. That means FAIR-EASE will mostly rely on pre-existing solutions, integrating them to facilitate and speed up their use and increasing their reliability and interoperability, applying the philosophy of re-using first, re-do only if necessary.

2.1.4. Provide services on EOSC marketplace

The EOSC marketplace (https://marketplace.eosc-portal.eu/) serves as a central hub where researchers, institutions, and service providers can discover, access, and utilise a wide range of services and resources. These services and resources can include data repositories, computing resources, software tools, data analysis services, and more. In the framework of EOSC-Pillar, guidelines and recommendations for the technical integration of resources and services into the EOSC Catalogue have been published (https://zenodo.org/record/5648215).

By providing the Earth Analytics Lab service on the EOSC marketplace, FAIR-EASE contributes to the growing ecosystem of research services and resources available to the scientific community. This facilitates collaboration and the adoption of best practices, which in turn advances scientific research and innovation.

2.2. Gathering pilots requirements

Interviews with the pilot representatives have revealed several key findings that are central to the success of our data-driven initiatives.

What emerges from the pilot interviews is a clear need for simple, efficient and fast data access. Participants highlighted the importance of accessing heterogeneous and distributed data sources that align with their project objectives. This involves not only the ability to perform data subsetting but also data harmonisation. Furthermore, there's a demand for remote processing capabilities, allowing to execute the code as close as possible to the data and computing facilities.

Another critical insight is the need for a common data space shared among various applications, tools, and scripts within a pilot. This common data space serves as a unifying environment that enables seamless data exchange and supports the creation of integrated workflows. Participants emphasised the need to break down data silos and foster collaboration across teams and tools.

The interviews underscored the importance of providing a Data Sciences Integrated Development Environment (IDE) like JupyterLab. Such platforms enable users to execute interactive notebooks and conduct exploratory data analysis. This capability is seen as essential for data scientists and analysts to efficiently work with data and code.

Participants expressed the need for a dedicated workflow management platform. This platform should facilitate the chaining of tasks and automate the execution of sequential data processing steps. A robust workflow management system streamlines processes, reduces manual effort, and enhances reproducibility.

Lastly, the pilot interviews identified the requirement for flexible IT resources tailored to the specific needs of each pilot. This includes provisions for scalable storage, adequate CPU capacity, and sufficient RAM resources. Providing the proper IT infrastructure ensures that the technical requirements of each project are met effectively.

These elements underscore the importance of providing a holistic and integrated data environment that empowers pilots to access, analyse, and manage data efficiently while fostering collaboration and flexibility in analytics workflows.

2.3. External context

2.3.1. Distributed data - A complex challenge

Having distributed data aggregators and providers in Earth System sciences introduces a set of complexities and challenges due to the nature of the data, the diversity of sources, and the need for interoperability and collaboration.

Despite these complexities, having distributed data aggregators and providers in Earth System sciences offers significant benefits. It enables a comprehensive understanding of global environmental processes, facilitates collaboration among researchers worldwide, and supports evidence-based decision-making for addressing pressing environmental challenges.

Addressing the complexities requires a collective effort from the scientific community, policymakers, and stakeholders to develop standardised practices, e-infrastructure, and policies that foster responsible data sharing and use in Earth System sciences.

The pilot interviews underscore the demand for straightforward and effective and fast access to diverse data sets and subsets within this distributed ecosystem. These requirements lead us to acknowledge that the FAIR-EASE project may not offer a one-size-fits-all solution. However, it can certainly make valuable contributions by providing guidance, recommendations, and alignment in several crucial areas :

- Interoperability: facilitate seamless machine-to-machine access by enhancing interoperability among different data sources and services, notably using semantics. This ensures that data can flow efficiently between systems, promoting accessibility and usability.
- Environmental impact mitigation: contribute to the optimisation of data and process movements in order to address the significance of environmental concerns. Reducing unnecessary data transfers allow optimisation of dataflows, avoid duplicating all data, and minimise the environmental footprint.
- Data synchronisation: efficient data synchronisation between various data centres is a solution to enhance data availability and reliability, ensuring that relevant data is consistently accessible.
- Data transformation: transform data into analysis-ready formats that are aligned with the varying needs of different communities, optimising their usage within the EAL. This includes considerations for data in its raw form versus data presented as a datacube, catering to diverse user or community preferences.

In essence, while there might not be a single "magic solution" to address all the complexities within this distributed data landscape, the FAIR-EASE project is committed to offering valuable support through guidance and collaboration in key areas, ultimately promoting greater data accessibility, sustainability, and usability for all stakeholders involved.

2.3.2. Feedback from previous projects

2.3.2.1. EOSC Pillar

The EOSC-PILLAR project, a European H2020 project aimed at supporting the implementation of the EOSC, which concluded in September 2022, had started to explore issues related to FAIRification, access, exploration, and analysis of heterogeneous and distributed data within the Earth System community. Indeed, the Agile FAIR Data for Environment and Earth System communities Use Case worked on a prototype of a VDAP that allows users to easily discover, explore, analyse, and reuse data from various Earth System disciplines and repositories. The actions taken and the challenges encountered are described in Chapter 2.2 of the following deliverable: EOSC-Pillar D6.3 - Final Report on Use Cases and Community Involvement (1.0).

This VDAP prototype deploys a Conda environment on a JupyterLab instance with various Jupyter notebooks using libraries and science modules from the Python ecosystem and Pangeo community[2]. It also provides access to climate and marine datasets converted into Big Data analysis-ready formats (Zarr, Parquet, etc.).

Two data synchronisation services were tested: OpenStack Object Storage (Swift) and iRODS. For the latter, several solutions were explored to enable on-demand access to distributed data: on-the-fly data subsetting or direct data reading on demand via a Python-iRODS client. The last solution was chosen, but due to security issues, it could not be fully implemented.

This prototype was built on a VRE platform offered by the D4Science infrastructure (also available through the EOSC Marketplace) and using computing resources provided by D4Science (CPU & RAM).

The tools used and the proposed functionalities are described in the following documentation: EOSCPillar4EarthScience VRE - Quick Start.

To improve the VDAP service, the deliverable recommends exploring a solution that offers additional computing and storage resources in an e-infrastructure better suited for parallel computing, improving on-demand access, and exploring containerisation on multiple clusters.

2.3.2.2. Blue Cloud

Between October 2019 and March 2023, Blue-Cloud deployed a cyber platform with a smart federation of multidisciplinary data repositories, analytical tools, and computing facilities to explore and demonstrate the potential of cloud-based Open Science and address ocean sustainability.

[2] https://pangeo.io/packages.html

The Blue Cloud VRE provides services to promote the collaboration among its users, to support the execution of analytics tasks embedded in a distributed computing infrastructure, and to enable the co-creation of entire Virtual Laboratories.

This solution is built on the gCube open source technology and hosted in D4Science infrastructure. The full description of this architecture can be found in the following document : D4.4 Blue Cloud VRE Common Facilities.

2.3.2.3. EOSC Phidias

PHIDIAS, standing for "Prototype of HPC/Data Infrastructure for On-demand Services", is a 3-year, EU-funded project that ended in September 2021. The main goal of the project was to develop services allowing researchers to discover, manage and process spatial and environmental data, in different scientific domains: earth surface, atmosphere and ocean.

In order to share all the data from the different scientific work packages (Ocean Infrastructure, Atmospheric infrastructure, Land Earth Infrastructure), it was decided to set up an iRODS federation between the different remote sites and the CINES (the HPC infrastructure). This solution aims to make available all the scientific data to all partners and no longer work in silos in each scientific work package individually.

With iRODS, files stored in heterogeneous storage resources deployed at different remote sites are exposed to users in a single unified namespace. However, it was required that each partner can have private data with strictly limited access while sharing other data more openly. The principle of iRODS federation which applies to partners in pairs meets this requirement of data privacy management. The following deliverable explains the architecture applied: D3.2.1 - Use cases for processing data from different Earth System compartments.

In order to execute remote processing near the data or for accessing important external hardware resources (e.g HPC), the WPS standard was evaluated. WPS defines a set of common HTTP API for describing then executing distant processes. If the standard describes common operations, each process has specific inputs and outputs that the client has to know for launching a process. This state of fact, limits either the ready-to-use part of the standard and the flexibility of the process.

The Phidias feedback tends to show that a ready-to-use implementation of this standard for accessing important hardware capabilities is probably the most pragmatic approach.

- It allows developing agnostic clients from whom processes can be launched by users with minimal knowledge about process specificity.
- Combined with a mechanism that allows users to select a specific Spatio-temporal Region of Interest, it permits to provide results on demand only for the area where users really need it.

When such an interface provides access to important hardware capabilities, it has to be associated with a quota mechanism that limits the amount of resources an user can mobilise. Such mechanisms or policies are not part of the standard but as to be implemented in practice.

3. Specifications

An Earth Analytics Lab is a virtual space where scientists and researchers study, analyse and process various aspects of the Earth System. It serves as a hub for exploring, processing data, and generating insights related to Earth Science and Environmental research. The EAL can enable multi- and cross-domain analysis by encompassing different disciplines such as geology, climatology, oceanography, ecology, and atmospheric science.

3.1. Users

An EAL can be of interest to various individuals, organisations, and sectors that rely on data-driven insights and analysis related to the Earth System. Here we list some of the stakeholders who may become users of the EAL proposed by FAIR-EASE:

- Scientists and Researchers: Earth scientists (geographers, ecologists, climatologists, and other domain related researchers) can benefit from an EAL. It provides them with access to data, tools, IT facilities and resources for conducting analyses, developing models, and advancing scientific knowledge.
- Environmental and Conservation Organisations: organisations focused on environmental protection, conservation, and sustainable development can use an EAL to understand ecological processes, monitor environmental changes, assess biodiversity, and support evidence-based decision-making.
- Government Agencies and Policy Makers: government agencies responsible for natural resource management, land-use planning, climate policy, disaster management, or environmental regulations can leverage the insights generated by an EAL to inform policy decisions, resource allocation, and long-term planning.
- Industries and Businesses: industries such as agriculture, forestry, energy, mining, urban planning, and transportation can benefit from EAL to analyse geospatial data, assess environmental impacts, optimise resource utilisation, and mitigate risks associated with their operations.
- Education and Academia: an EAL can serve as valuable resources for educational institutions, providing students and researchers with hands-on experience in data analysis, geospatial technologies, and interdisciplinary research related to Earth System.
- Non-Governmental Organisations (NGOs): NGOs working in areas such as disaster response, humanitarian aid, climate change adaptation, or sustainable development can use an EAL to analyse data, support decision-making, and design effective interventions.
- Public and General Audience: the insights and visualisations generated by an EAL can be communicated to the general public through media, websites, or public outreach initiatives, raising awareness about Earth Systems, climate change, or environmental issues.

The wide range of stakeholders interested in an EAL reflects the interdisciplinary nature of Earth sciences and the importance of data-driven analysis for understanding and addressing challenges related to the Earth System.

Users may have different technical skills and knowledge. Collaboration and interdisciplinary work among these user groups, but also training and educational resources, can lead to more comprehensive and impactful analysis.

3.2. Data and the research processes

There are typically several distinct stages comprising the data analysis and research process. These stages can vary depending on the specific goals, objectives, and methodologies employed in the lab. However, here are some common stages: problem definition, data discovery and access, data exploration and preprocessing, data analysis, interpretation and modelling, visualisation and communication, etc. (see appendix 7.2 for more details).

The specific steps and their sequence may vary depending on the objective, the research goals and the nature of the Earth-related data being analysed.

Data analysis involves various operations to manipulate and extract meaningful insights from scientific data: subsetting, filtering, aggregation/reduction, resampling, colocation, etc. (see appendix 7.3 for more details). These operations are not exhaustive but provide a good starting point for understanding the different ways scientific data can be processed and analysed. The choice of operations depends on the specific research questions, data characteristics, and analytical goals.

Data visualisation plays a crucial role in understanding and communicating complex data related to Earth science and environmental research. It can be useful at all stages of the research process. Some key functions and benefits of data visualisation are explained in appendix 7.4.

As part of FAIR-EASE, the three technical working groups provide technical solutions to meet the needs of pilots and beyond. Below is an example of data flow from the EAL user's point of view, highlighting the main functions that need to be implemented.

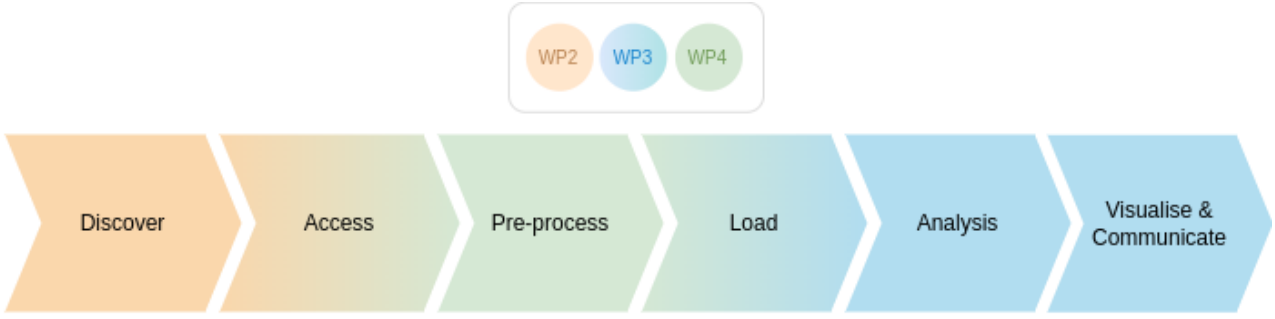


Figure 2 - Example of data flow from the EAL user's point of view

3.3. Definition and requirements

The Earth Analytics Lab to be deployed by the project is an **integrated, collaborative and FAIR, analytic web-based platform** that enables users to process, analyse, and derive insights from large volumes of data. It is not just a set of analytical tools and services. It is a value-added platform that facilitates data analysis processes, data-driven decision-making, teamwork, and knowledge sharing.

The EAL is built upon standard IT infrastructure (e.g. storage and compute). It is agile enough to work with heterogeneous underlying technologies, but also able to manage distributed or federated infrastructures (e.g. remote processing).

Features of the EAL:

- Data ingestion and harmonisation: The platform allows for the ingestion of data in various formats (e.g. NetCDF, CSV, PNG) via different protocols (e.g. HTTP, S3, OPeNDAP, STAC, or custom API's). This process includes the capability of transforming and harmonising diverse data sets into a unified format, thereby ensuring data consistency. The platform can be used to retrieve raw data and produce optimised, pre-configured, data sets. Such curated Analysis Ready Datasets often constitute a set of key reference datasets whose quality has typically been assessed in a documented and traceable way and which has been the subject of a consensus within the user community.

- Data and files management: The platform relies on infrastructure to store large volumes of structured, semi-structured, and unstructured data. It incorporates data governance and security mechanisms to ensure data protection and compliance. It also provides features to catalogue, index, and synchronise datasets.

- Data preprocessing and transformation: The platform provides tools for data curation, data filtering, data normalisation, and data transformation that are used to prepare data for analytical processing.

- Data analytics and processing: The platform offers a range of analytical capabilities, including tools for exploratory data analysis, statistical analysis, machine learning, and data-mining techniques. These tools enable users to uncover patterns, relationships, and insights from the data.

- Scalability and distributed computing: The platform is designed to handle large-scale data processing requirements. It leverages distributed computing frameworks, parallel processing, and cluster computing to achieve scalability and performance in analysing massive data sets.

- Visualisation and reporting: The platform provides data visualisation tools and reporting capabilities to create interactive visualisations, charts, dashboards, and reports. These visual representations help users understand and communicate insights effectively.

- Collaboration and sharing: The platform incorporates features for collaboration, allowing multiple users to work together on data analytics projects. It can include functionalities for

sharing code, visualisations, and insights among team members, promoting collaboration and knowledge sharing.

- Data governance and security: The platform prioritises data governance, confidentiality and security. It implements access controls, encryption, data anonymisation and compliance measures to ensure the protection and ethical use of data.
- Integration with other systems: The platform integrates with other systems, such as data sources, tools, and data visualisation software. This enables seamless data flow and integration in a distributed ecosystem (i.e. federated data, services, and platforms).
- Scalability and extensibility: The platform is designed to accommodate evolving requirements of data analytics. It can scale horizontally and vertically to handle growing data volumes and accommodate new analytical techniques, algorithms, and tools.

The EAL will abstract the technical complexity presented to the end user as far as possible by:

- Deploying tools and applications on heterogeneous infrastructures (HPC, Cloud, computer): The platform provides automated processes to simplify the deployment of software as a service or execution of a software as a job.
- Providing execution and monitoring of tools: The platform offers a simple way to configure the inputs and retrieve outputs. It also provides the ability to monitor processing progress and easily and quickly identify any potential problems.
- Enabling data access and loading: The platform offers simple and efficient access to data, wherever it is located and whatever its format.

Finally, the EAL is based on FAIR principles, for data, software, and middleware. The latter are special software that connects and facilitates communication between different components (hard or soft).

4. Design

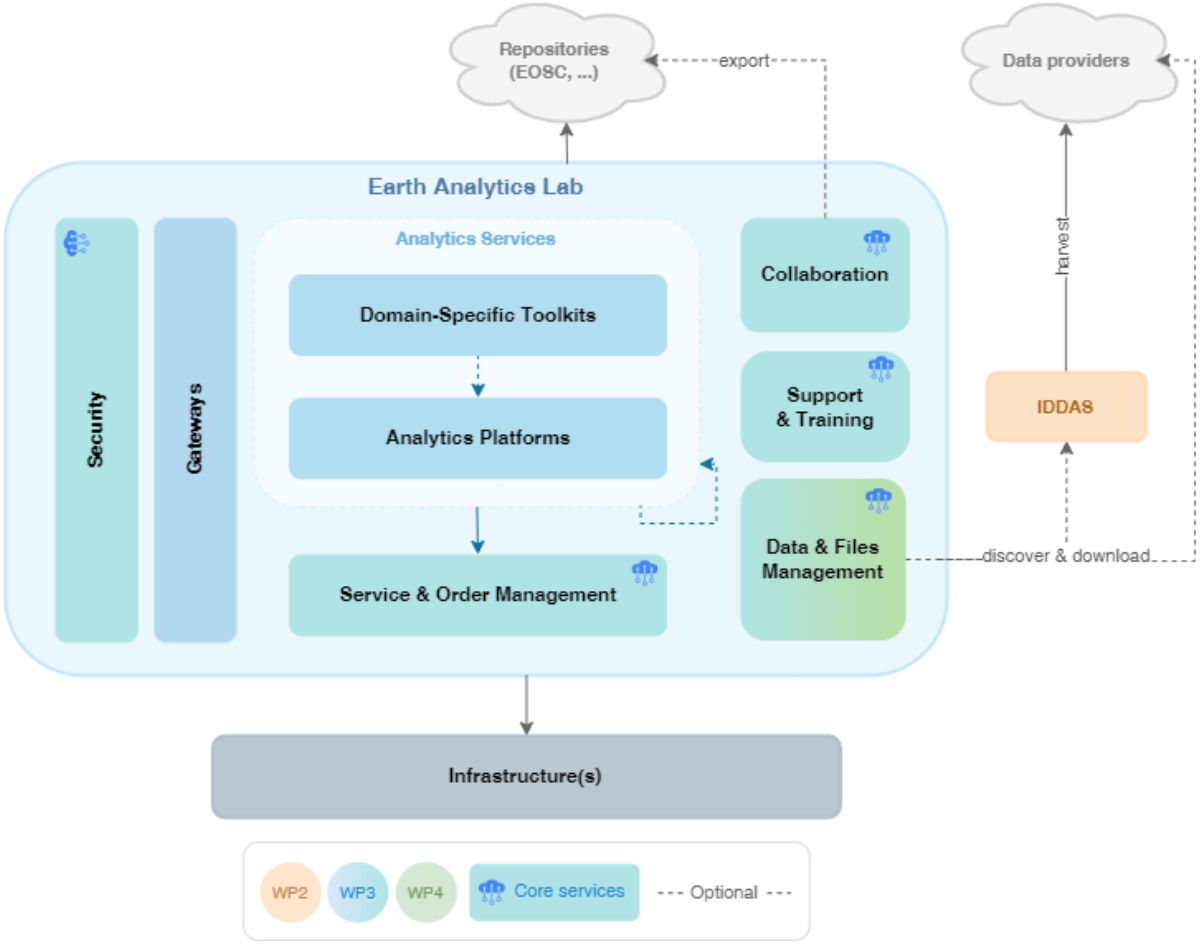


Figure 3 - Earth Analytics Lab architecture simplified diagram

The general architecture diagram in Figure 3 presents the expected features of an Earth Analytics Lab (e.g. analytics, core, and gateway services) and its external interfaces (i.e. data, IT resources, various repositories, and catalogues, including those of EOSC).

This general architecture diagram also outlines the responsibilities of each FAIR-EASE technical work package (WP2, WP3 and WP4), particularly with regard to data discovery, access, and management.

The following chapters present these various components, starting with analysis services, then core services. and finally gateways. A complete diagram of the FAIR-EASE architecture is provided in the 7.1.

4.1. Analytics Services

Analytics services refer to a range of tools and analysis capabilities that enable the processing and interpretation of Earth-related data for various scientific, research, and application purposes. These services are designed to facilitate the extraction of insights and knowledge from complex geospatial and environmental datasets.

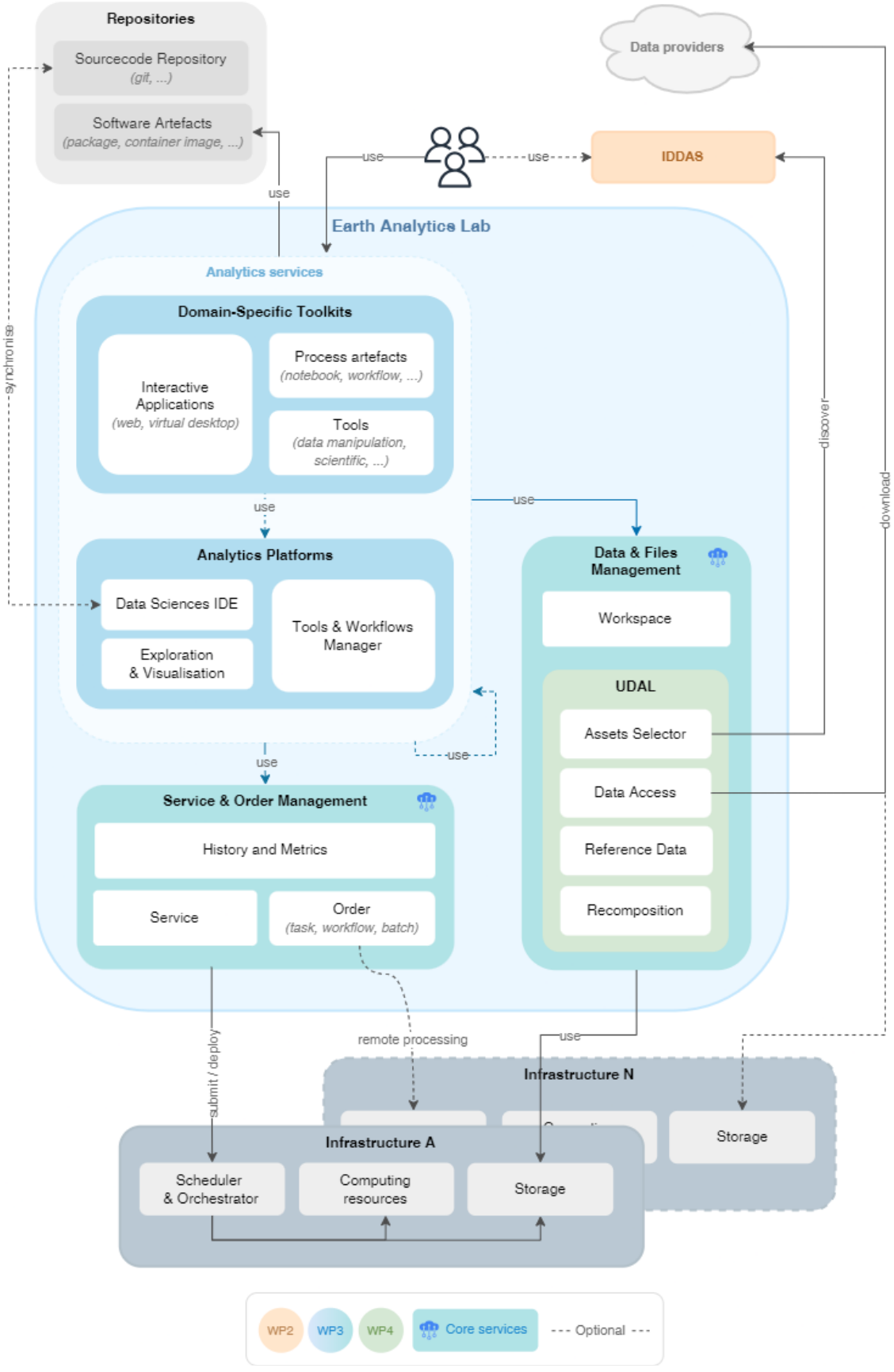


Figure 4 - Analysis services diagram

Interactive applications are user-friendly web applications that allow researchers, scientists, and analysts to directly engage with data and perform real-time analyses. These applications enhance the interactivity and flexibility of the analytical process. The types of interactive applications that will be commonly used in the Earth Analytics Lab include data sciences IDE's, exploration and visualisation applications, as well as domain-specific applications. These applications can be web-based applications or desktop applications encapsulated in a virtual desktop.

On-demand processing involves executing predefined tasks or workflows on datasets without requiring constant user interaction. These services are particularly useful for resource-intensive or time-consuming tasks. In scientific research and data analysis, the ability to reproduce results is critical. The history of tasks and provenance of data enable researchers to recreate experiments and analyses, fostering transparency and peer review.

Access to data is crucial for delivering a robust and high-performance service, whether in terms of data locality and data format, as well as the technologies used to store the data. Within the Earth Analytics Lab, data can come from the workspace of a user or a group of users, as well as from reference data. The latter avoids duplicating data useful to several users in their workspace. Usage of IT resources can be limited by individual or collective quotas (e.g. storage, computing resources). The Earth Analytics Lab will provide metrics on usage and quotas.

The use of software artefacts (e.g. language or conda packages, container images such as from Docker or Apptainer) simplifies the deployment of environments and applications on heterogeneous platforms (e.g. HPC and Cloud).

4.1.1. Analytics Platforms

4.1.1.1. Data Sciences IDE

A Data Sciences IDE is a platform that provides an environment for data exploration, manipulation, and modelling. It provides a comprehensive set of tools and functionalities to facilitate data analysis and development. A Data Sciences IDE typically includes the following components:

- Data manipulation and analysis: the Data Sciences IDE provides tools and libraries for loading, exploring, cleaning, transforming, and visualising data. It often integrates popular data manipulation libraries such as pandas/xarray in Python.
- Interactive development: the Data Sciences IDE provides an interactive environment for executing code, exploring results, and experimenting with different data analysis techniques. It may support interactive notebooks, such as Jupyter notebooks, which allow for combining code, text, and visualisations in a single document.
- Model development: the Data Sciences IDE offers tools and libraries for building, training, and evaluating machine learning models. This includes access to popular machine learning frameworks, algorithms, and libraries, such as scikit-learn, TensorFlow, or PyTorch.
- Collaboration and Version Control: the Data Sciences IDE can provide features for collaborating with other team members, including code sharing, version control integration (e.g., external Git Repository).

4.1.1.2. Data Exploration and Visualisation Applications

Data exploration and visualisation applications allow users to easily discover, combine, and visualise different data layers.

Data visualisation applications refer to generic web applications used to represent and communicate complex Earth-related data in a visually intuitive and informative manner. These applications help transform data from various Earth Sciences sources, such as satellite imagery, climate data, environmental data, or geospatial information, into meaningful visual representations and synthetic information. This can include maps, charts, graphs, and 3D models.

4.1.1.3. Tools and Workflows Manager

User Interfaces (UIs) are essential in this context. They simplify task and workflow configuration, design and submission, provide execution status and logs, and offer easy access to analysis or visualise outputs.

4.1.2. Domain-Specific Toolkits

Domain-Specific Toolkits refer to software packages and resources tailored to address specific Earth-related research and data analysis. They are designed to cater to the unique requirements of distinct Earth science domains and research areas.

Within these toolkits, users and communities can provide or use :

- Interactive applications: web-based or desktop applications.
- Scientific tools: tools developed with a deep understanding of the specific scientific field they serve, providing specialised functions and algorithms.
- Data tools: tools designed to manipulate files or data sets and provide value-added data that can be used by scientific tools (e.g. transformation and colocation)
- Process artefacts: set of files and configurations (e.g. scripts, notebooks, workflows, and environments) enhancing automation and collaboration.

For effective deployment and use, analytical processes need to be broken down and simplified. This process is called *atomisation* and *generalisation*. By doing this, one can make sure that the tools are reliable and easy to work with.

While primarily designed for domain-specific use, these applications may also facilitate interdisciplinary collaboration by providing tools that bridge multiple Earth science disciplines.

4.1.3. Interaction between analytics services

The interaction between interactive applications and on-demand processing streamlines the Earth Analytics workflows. It empowers researchers to seamlessly transition from exploration to large-scale processing, monitor progress, and integrate results, all within a unified and user-friendly environment. This dynamic interaction improves the efficiency of the user experience.

For example, users will be able to submit a task or a workflow from an interactive application (e.g. interactive notebook). Similarly, users will be able to submit interactive tasks (e.g. virtual desktop).

Moreover, this interaction can bring added value on streamlining recurrent and automated processing, for example by preprocessing workflow of some continuously updated input data to feed a model or interactive applications (e.g. data exploration and visualisation).

4.2. Core services

Core services are designed to enhance collaboration, reduce duplication of effort, and provide researchers with the infrastructure and support they need to conduct their work more efficiently and effectively. They are a key component in research and innovation environments, such as the Earth Analytics Lab, to enable scientists to focus on their primary research goals while benefiting from centralised resources and expertise.

Core services act as an interface between the Earth Analytics Lab and the underlying infrastructures, as well as with data providers and repositories of scientific assets (e.g. tools, workflows, and data). These services are used by the analytics applications and tools for building durable services.

4.2.1. Security

The security module ensures that only authorised individuals or groups can access protected or sensitive data and resources. It can also help facilitate the efficient management of distributed IT environments.

Figure 5 - Security module

4.2.1.1. Virtual Organisation

Virtual Organisations (VO) have become increasingly prevalent as technology has made it easier for individuals and groups to connect and collaborate remotely. They offer a flexible and efficient way to harness collective expertise and resources to address complex challenges and pursue shared objectives.

In the framework of the Earth Analytics Lab, each VO has access to a common space, including a workspace, analytics services, social information, and metrics.

4.2.1.2. Identity and Access Management

Identity and Access Management (IAM) is a framework of policies, technologies, and processes used to manage and control user access to any protected or sensitive resources.

The primary goal of IAM in the Earth Analytics Lab is to ensure that the right individuals or VO have the appropriate access to the right resources at the right time, while also ensuring the security and compliance of the environments and data, especially in complex and distributed IT environments.

Key components of IAM typically include:

- Authentication: process of verifying the identity of a user or system. It involves proving that an individual is who they claim to be.
- Authorization: determine what actions or operations a user or system is allowed to perform once their identity has been authenticated. It involves defining and enforcing access policies and permissions based on roles, attributes, or other criteria.
- Directory Services: store and manage user identities and attributes, typically in a central repository in order to provide a centralised source of user information and authentication.
- Access Control: enforce the authorization policies defined for users and resources. They may include access control lists (ACLs), role-based access control (RBAC), and attribute-based access control (ABAC). These mechanisms ensure that users only access resources they are authorised to use.
- Single Sign-On (SSO): allow users to access multiple applications or services with a single set of login credentials. This simplifies the user experience and reduces the need to remember multiple usernames and passwords.
- Identity Lifecycle Management: manage the entire lifecycle of user identities, including user provisioning (creating accounts), de-provisioning (disabling or deleting accounts), and managing changes to user roles and permissions as needed.
- Audit and Compliance: capabilities to track user activities and access events. These logs are crucial for compliance with regulatory requirements and for investigating security incidents.
- Password Management: include password policies, password reset mechanisms, and self-service password recovery to enhance security and user convenience.

4.2.1.3. Vault

Vaults are essential components of modern IT security practices, helping users to protect sensitive information, maintain compliance, and reduce the risk of data breaches and security incidents.

A vault refers to a secure and centralised repository or storage system designed to store sensitive information such as passwords, cryptographic keys, API tokens, and other confidential data. The primary purpose of a vault is to provide a secure and organised way to manage, store, and retrieve these secrets while ensuring their confidentiality and integrity.

By providing a vault solution, the Earth Analytics Lab enables users to avoid hard-coding confidential data in process artefacts, such as in interactive notebooks, or in the configuration of tasks and workflows.

4.2.2. Service and Order Management

This core service is responsible for deploying and managing analysis services as a service or task on the underlying infrastructure(s).

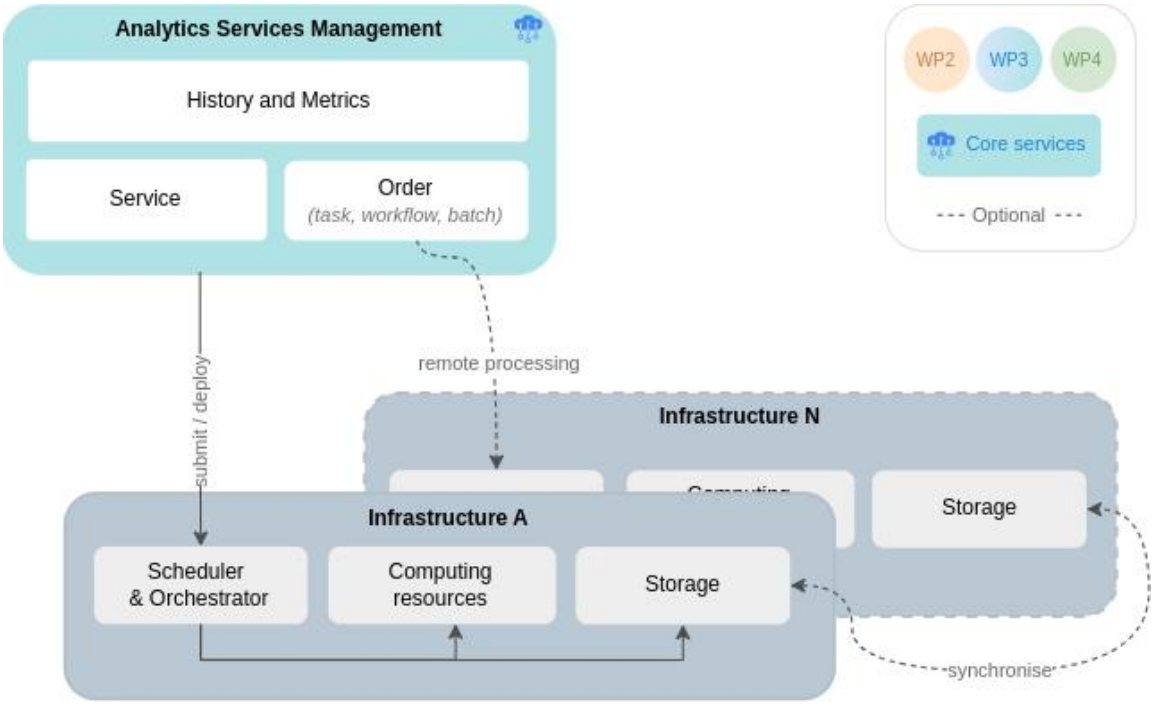


Figure 6 - Analytics Services Management module

4.2.2.1. Service Management

This component is designed to deploy and manage containerised interactive applications (i.e. web applications or virtual desktop applications) in a scalable and secure manner (e.g. read-only or read-write mode mounting point, inject secrets and configuration). This component acts as a proxy, routing the web request to the internal URL of the service.

Access to logs and metrics by service managers is an asset to identify and fix issues.

4.2.2.2. Order Management

The main elements managed by this component are as follows:

- Task Execution: a task represents the submission of an order, along with the necessary tools and parameters, to be processed as a job. A task encompasses the entire lifecycle of requesting, processing, and obtaining the outcomes of a particular job or operation. It can be initiated as part of a specific operation or set of instructions, often as part of a workflow or larger system.

- Workflow Execution: chaining the execution of sequential or parallel tasks. One of the outputs of a task can become the input of another task. Workflows streamline complex analyses by automating the data analysis pipeline from start to finish.
- Remote Processing: the ability to perform data processing and analysis on remote servers or cloud infrastructure, preferably on infrastructure physically near the location of where the data is stored.
- Batch Processing: automating the execution of multiple tasks or workflows on large volumes of data in a queue-scheduled mode. This approach optimises computational resources and efficiency.

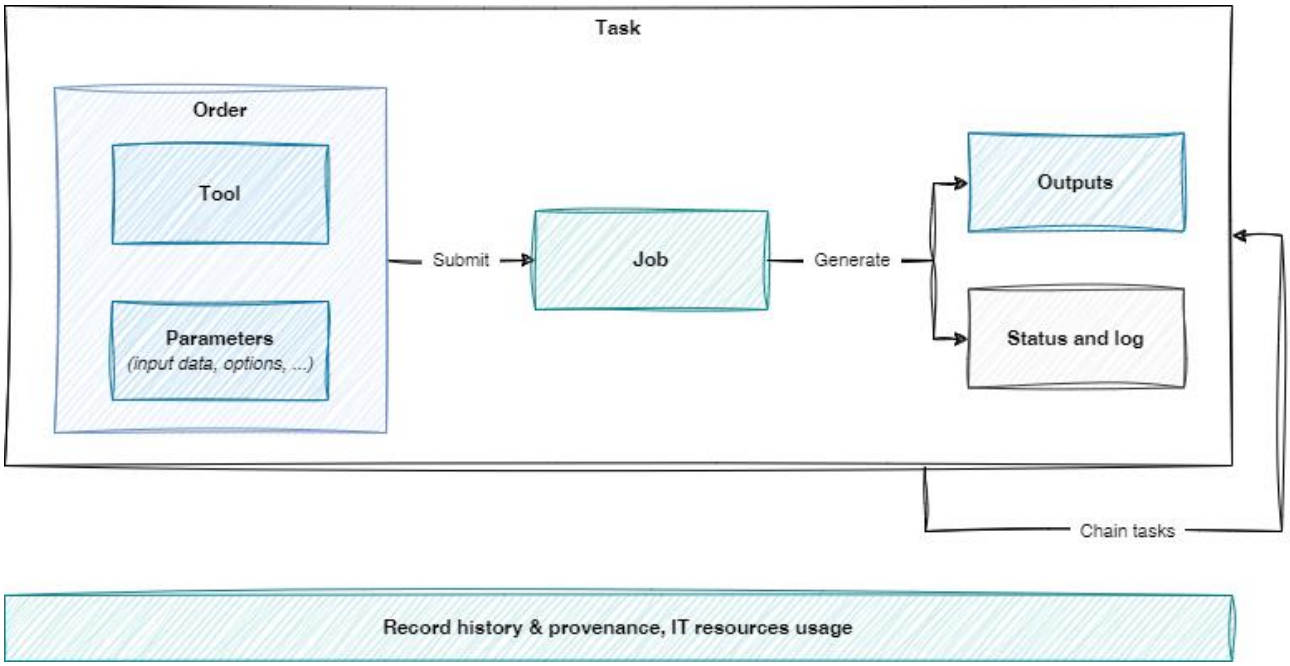


Figure 7 - Task and workflow management

4.2.2.3. Task History and Metrics Management

Recording the provenance metadata of tasks, including the details of software versions, execution parameters, and data sources, is important for traceability, reproducibility, and data reuse. But it is also useful for optimising the use of IT resources, and fostering transparency during the peer review process.

Usage of IT resources can be limited by individual or collective quotas (e.g. storage, computing resources). The component will provide metrics on usage and quotas.

4.2.3. Data & Files Management

Data and files management is a crucial module within the EAL. Thanks to the FAIR-EASE implementation of a data lake, called Uniform Data Access Layer (UDAL), users also will be able to access local or remote data transparently, without having to worry about the location or format of the data. They will also be able to bring their own data by uploading files.

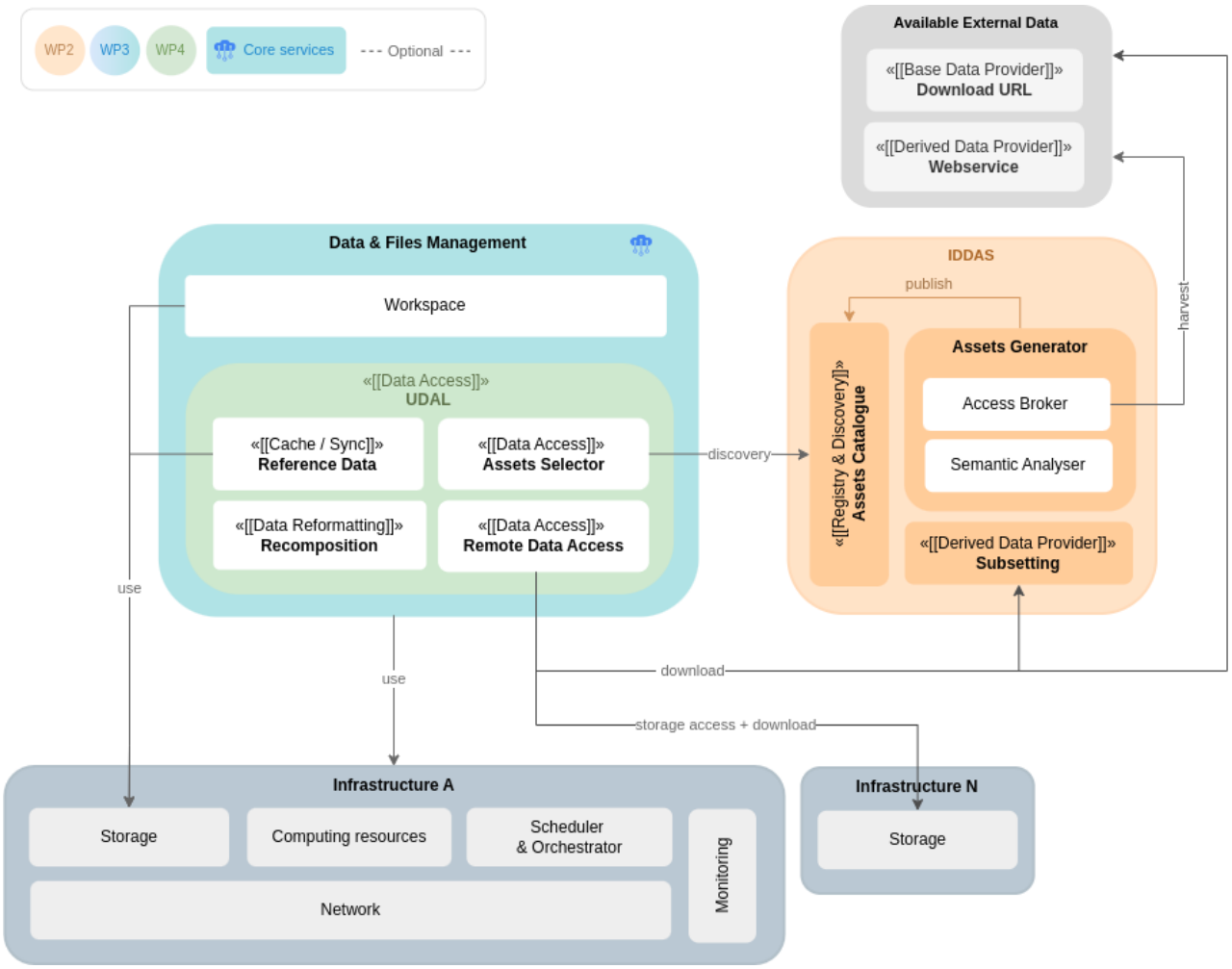


Figure 8 - Data & Files Management module

In the previous figure, stereotypes have been added to make the link with the functional blocks identified in FAIR-EASE deliverable D4.1.

4.2.3.1. Workspace

A workspace is a virtual space allowing individuals or VOs to store any files (e.g. data, documentation, configuration, and processing outputs) with a limited storage quota.

4.2.3.2. UDAL

The UDAL includes mechanisms for discovering and collecting sets (i.e. [[Base Data Provider]] ; e.g. S3 protocol) or subsets (i.e. [[Derived Data Provider]] ; e.g. OPEnDAP protocol) of data from external data providers, local storage (i.e. Reference Data) or from other infrastructures. It can also request any relevant web service (e.g. remote processing via OGC:WPS, image generation via OGC:WMS).

The UDAL uses a local shared and cached storage to optimise access to recurrent data and also to provide up-to-date data with synchronisation mechanisms

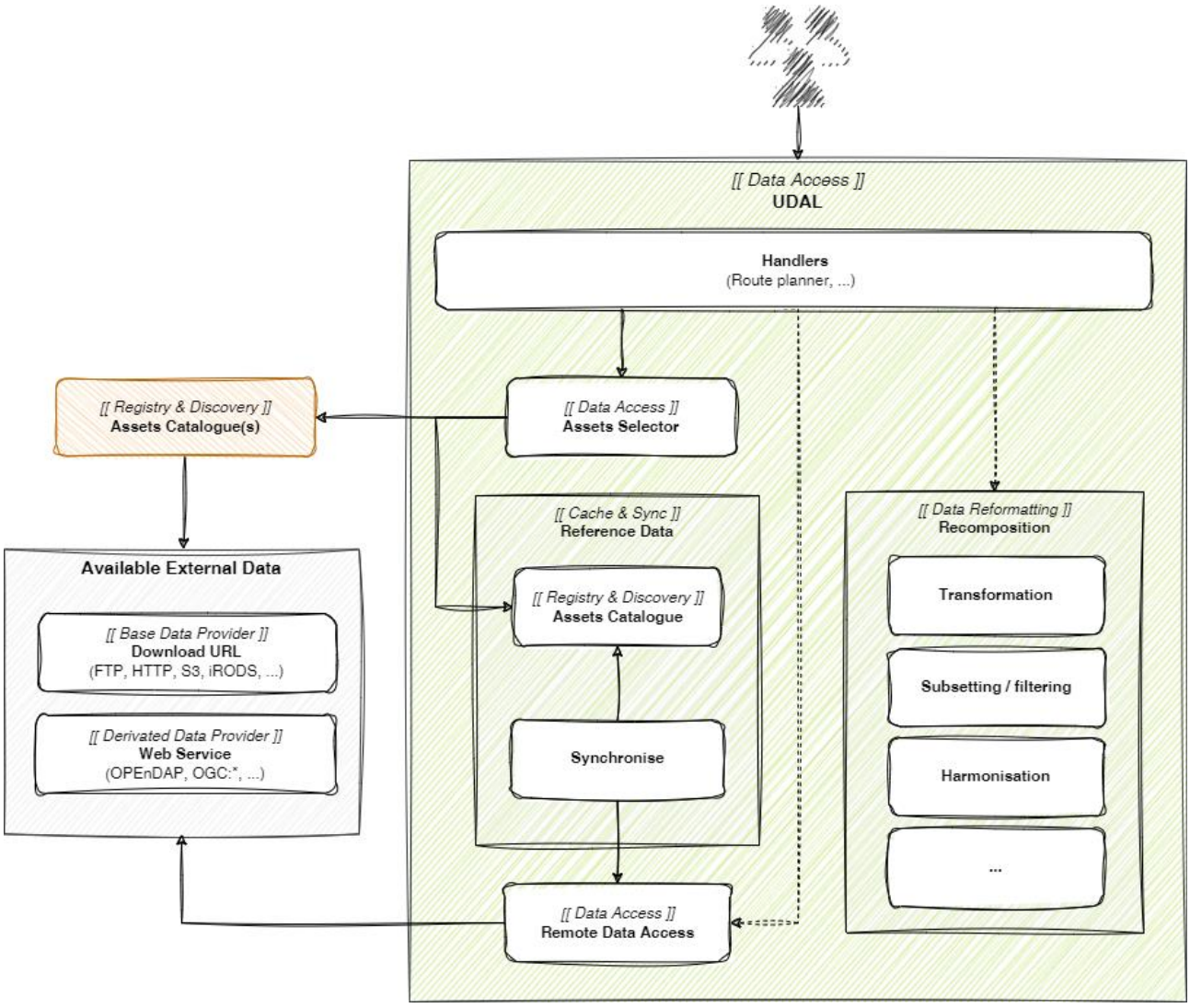


Figure 9 - UDAL draft architecture (aka datalake)

The UDAL will use an Assets Selector to request Assets Catalogues (e.g. IDDAS) in order to discover relevant data collections and retrieve the corresponding assets. Then it will use a flexible solution (i.e. plugin-based) to access the files or request the services.

Reference Data is a virtual area for storing, indexing, and cataloguing data sets in raw or analysis-ready format that are most relevant to research projects, ensuring data consistency, curation, and accessibility. This avoids data duplication on different workspaces, and also avoids reducing user/team quotas.

Data recomposition refers to the process of restructuring or reformatting data to make it more suitable for analysis, visualisation, or modelling. This process is crucial for handling and working with diverse Earth and environmental data sets efficiently. If some functions are not available through UDAL, users can always use their own solution as an analytics service (see chapter 4.1.2).

4.2.3.3. IDDAS

The Interdisciplinary Data Discovery and Access Service (IDDAS) will include the assets provided by the Access Generator under one banner using a to be determined DCAT standard. It is important that in the EAL the user/or machine knows what type of access can be expected and how to respond to retrieve the asset. This means that there will need to be a clear description for each of the assets indicating the type of (meta)data access that it provides.

The assets that are provided by the different components that are included in the asset catalogue will be described using a to be determined DCAT standard, which is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web. The assets will be harvested from, or manually created for, the various components behind it.

For the Access Broker, the metadata is harvested and converted to the FAIR-EASE metadata profile, that contains identifier, title, keyword, bounding box, temporal extent, parameter, instrument, platform, organisation, date-stamp, revision date, and resource identifier. This set of information is then ultimately converted to the DCAT standard, and it should generate access information for the assets from the metadata it already holds.

The goal of the semantic analyser is to automatically analyse data and metadata records in order to identify if certain metadata fields include semantic artefacts. Consequently it will be mapped to other semantic artefacts from different repositories to create crosswalks and support interoperability and discoverability.

A subsetting service will be provided to make up for the lack of this type of service from certain data providers (e.g. data only available through an FTP server). This service will also be registered in the Assets Catalogue.

As an independent service, the IDDAS does not need to be deployed on the same infrastructure as EAL.

4.2.4. Support & Training

Training material will be created to help researchers make effective use of the EAL, from analytics platforms and domain-specific toolkits to data & files management. This will include guidelines and best practices, webinars, notebooks, workflows and different topics to promote responsible and efficient research practices.

Once operational, we envision that the EAL will also be used in hands-on courses that teach scientific data analysis procedures and data management in general.

4.2.5. Collaboration

The Earth Analytics Lab will provide a set of interactive and community-oriented functionalities that promote communication, knowledge sharing, and teamwork among researchers, scientists, and analysts. These features enhance collaboration, foster a sense of community, and facilitate the exchange of ideas and information in the pursuit of Earth science research and analysis.

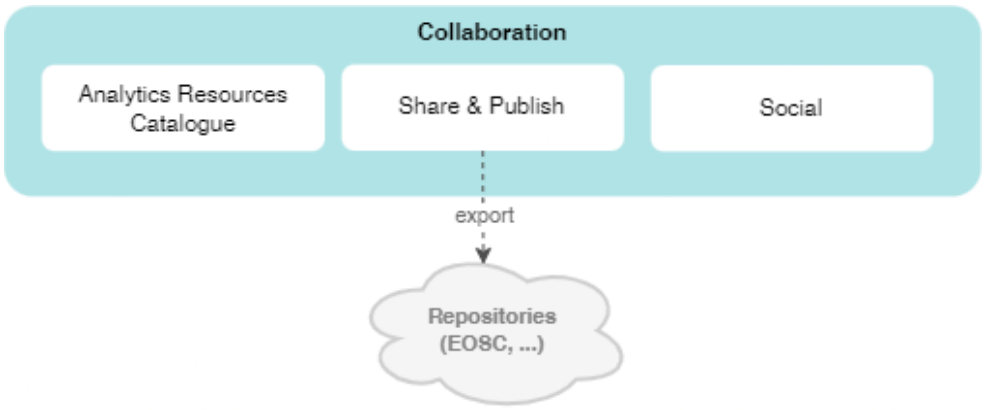


Figure 10 - Collaboration module

4.2.5.1. Analytics Resources Catalogue

An Analytics Resource Catalogue is a centralised digital platform designed to store, manage, and provide access to a diverse range of assets, tools, training materials, and data resources that support the activities of data analysts, data scientists, and other professionals involved in analytics and data-driven decision-making.

This catalogue serves as a comprehensive and easily accessible collection of resources, enabling users to find, share, and utilise the necessary tools, knowledge, and data for conducting data analysis, modelling, and other analytics-related tasks. This catalogue will help users navigate through a variety of resources, from training materials to data and analytics services. These assets can be ordered and filtered with relation to a specific usage, community or thematic.

It plays a crucial role in streamlining analytics workflows, promoting collaboration, and facilitating the efficient use of resources within a community of analytics users.

4.2.5.2. Share and Publish

This component enables sharing files between user or VO workspaces: datasets, research findings, process artefacts, etc. It will also enable workspace data to be published as Reference Data.

This component will allow users to export analytics resources (e.g. notebook, tools, workflows) from the EAL to external repositories, such as EOSC Marketplace, Thematic Trust Repository (e.g. https://zenodo.org/), scientific computational workflows (e.g. https://workflowhub.eu/).

4.2.5.3. Social

The social functions contribute to cooperation between users and the feeling of belonging to a community of users. Here are some examples of social features that can be provided:

- User Profiles: Provide a way for EAL members to introduce themselves, showcase their expertise, and connect with others who share similar research interests.
- News, Notification and Alerts: Receive notifications and alerts for important updates, discussions, or events, thereby ensuring users stay informed and engaged.

- Social Networking Integration: Integration with social media platforms allows EAL members to share content and updates with a broader audience, extending the reach and impact of their research.

4.3. Gateway services

Analytics services and Core services will be accessible via a single web UI called the EAL Web Portal. These services will also be accessible via a web service called EAL Web API.

These two gateways are accessible behind the Security module.

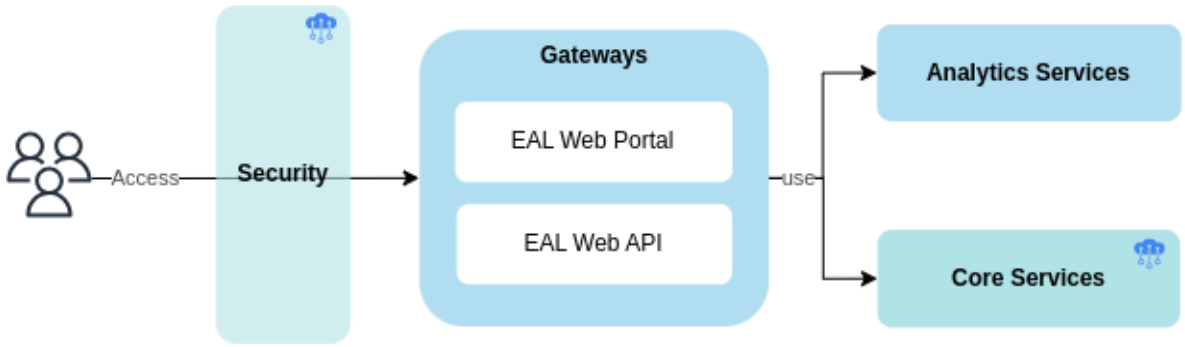


Figure 11 - Gateways module

4.3.1. Web API

The web API will be used to integrate analytics and core services within the EAL, but also to access these services outside the EAL (e.g. retrieving the user's workspace, submitting a task or workflow). The web API will allow machine to machine (M2M) access by providing a REST API that complies with HATEOAS[3] constraints and FAIR principles[4].

4.3.2. Web Portal

The Web Portal will allow users to access EAL internal services via a centralised Web UI.

The analytics resources catalogue (e.g. reference data, analytics services, support and training materials) will be accessible publicly.

A Virtual Space will be provided for each user and VO. It will group all the services tailored for this dedicated space :

- Workspace: browse the content (owned or shared) and visualise online if possible (e.g. image, date file)
- Data: discover, explore and visualise data from Reference Data

[3] https://doi.org/10.1109/CINTI-MACRo57952.2022.10029427

[4] https://doi.org/10.5281/zenodo.6656431

- Analytics services: manage and use interactive applications, manage and execute on-demand processing (i.e. tasks or workflows).
- Collaboration: social and sharing features
- Metrics : IT resources usage (i.e. storage and computing resource)

4.4. Infrastructures

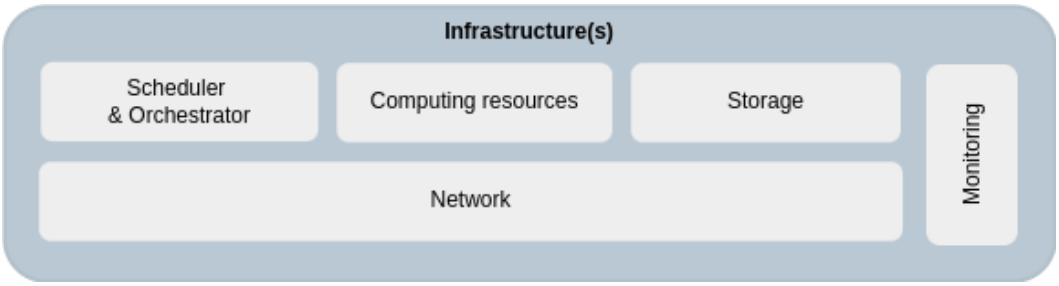


Figure 12 - Infrastructures components

The EAL will rely on scalable infrastructures and will be able to be used for distributed and heterogeneous infrastructure (from computer to private or public Cloud/HPC clusters).

The underlying infrastructure shall provide :

- Storage: workspace, reference data, software artefacts, configuration files
- Scheduler and orchestrator: deploy applications, submit jobs, access status / logs
- Computing resources: CPU, GPU, Memory, local temporary storage
- Network: efficient data transfer between storage and processing resources, between external services (e.g. data, map layers) and EAL
- Monitoring: status and availability, accounting

5. Implementation plan

During this first year of the project, the analyses of requirements, specifications, and design phases, notably with this deliverable, were completed. The implementation (including evaluation) and testing phases will take the form of an iterative process that will begin by defining a minimum viable product (MVP) based on the needs expressed by the pilots and proceed through to the implementation of a complete and integrated Earth Analytics Lab.

A lot of initiatives have been launched during the previous years by different projects and communities. The EAL does not reinvent the wheel but aims to exploit, re-use and integrate solutions that already exist. Therefore, it is crucial to establish partnerships with other projects and collaborations with external communities, to have support but also to propose enhancements.

The EAL will use open standards in order to enable and simplify interoperability, but also portability of the solution.

A dedicated space has been created on Github (https://github.com/fair-ease) to share the code produced during the project lifecycle.

5.1. Partnerships and synergies

5.1.1. FAIR-IMPACT

A collaboration with the FAIR-IMPACT project (https://fair-impact.eu/) has been initiated around the FAIRness of the FAIR-EASE research objects. This collaboration focuses on the FAIRness of data, research software, and semantic artefacts.

The research software that will be used as a service in the EAL has been evaluated according to a methodology developed in the project (see Milestone MS06 for details) and the results will be presented at the FAIR-IMPACT annual meeting.

5.1.2. EuroScienceGateway

A collaboration with the EuroScienceGateway project (https://galaxyproject.org/projects/esg/) has been initiated around the evaluation, usage, and contribution to Galaxy (https://galaxyproject.org/).

Galaxy is an open-source platform for FAIR data analysis that covers a wide range of modules and components identified during the design of the EAL. It is particularly useful for on-demand processing, as it enables non-interactive and interactive tasks to be carried out from a catalogue of tools, as well as designing and executing workflows. Thanks to its modularity, Galaxy can also interface with a large number of core service implementations.

During this first year, in collaboration with EuroScienceGatweay, tools for, and from, pilot projects were deployed on Galaxy Europe. The methods used to implement interactive tools (e.g. JupyterLab,

Virtual Desktop) on Galaxy were also improved. For example, a docker image was used to deploy QGIS[5] on Galaxy (docker-qgis).

Our collaboration is reflected in our participation in joint events, for example: two-days training, Galaxy Community Conference 2023, European Galaxy Days (EGD).

Galaxy will be one of the EAL solutions. To enable the use of the tools produced by the FAIR-EASE pilots, but also by all the communities working within Earth system sciences, we will create a dedicated sub-domain on Galaxy Europe. The following link provides the existing subdomains on Galaxy Europe: https://galaxyproject.org/eu/subdomains/.

5.1.3. Blue-Cloud 2026

A collaboration with the Blue-Cloud 2026 project (https://blue-cloud.org/about-blue-cloud-2026) is currently under negotiation.

CNR, a partner of Blue-Cloud 2026, provides an e-infrastructure called D4Sciences, linked to EOSC, which offers an integrated solution close to the EAL design as well as computing resources (https://dev.d4science.org/). In return, the FAIR-EASE project could enhance the existing BC data discovery and access components, and provide a Galaxy platform, as well as other domain-specific toolkits.

5.2. Macroscopic Roadmap

During the first year, as a result of the meetings and consultations with the FAIR-EASE pilots and through the collaboration with EuroScienceGateway, some required Analytics Platforms have already been identified:

- Data sciences IDE: JupyterLab and Pluto.jl
- Data exploration and visualisation: Examind Community and WebODV
- On-demand processing: Galaxy

Parallel actions will be carried out to evaluate, test, and implement solutions in different areas, in line with the organisation of the project's development cycle.

Apart from the need to identify the target infrastructure, a focus on the EAL as a whole will be realised by evaluating existing core services integrated or developed by Galaxy and D4Sciences, but also include other core services that could be provided by EOSC:

- Interactive applications deployment: JupyterHub (https://jupyter.org/hub), ShinyProxy (https://www.shinyproxy.io/) used by D4Sciences or other solutions such as in Galaxy
- Training: Galaxy training Materials (https://training.galaxyproject.org/) or other solutions

[5] https://www.qgis.org/en/site/about/index.html

- Catalogue : CKan (https://ckan.org/) used by D4Sciences or other solutions
- Security: EOSC solution such as EGI Check-in (https://www.egi.eu/service/check-in/) or other solutions

Work will be needed to identify and develop EAL gateways and integrate services. Finally, the EAL will be registered on the EOSC Marketplace. This action should be anticipated because the process can be complex and time consuming.

Support to pilots will be provided on packaging and deployment of the applications and tools (i.e. domain-specific toolkits) they need or provide, but also by improving data access. Here are some relevant topics:

- Taking part in the Uniform Data Access Layer (UDAL) prototype leaded by the WP4 (see deliverable D4.1)
- Providing a set of tools and best practices used by the various communities, such as Analysis-Ready Cloud Optimised (ARCO) and STAC catalogue with Pangeo Ecosystem or intake, EODAG, and openEO.
- Evaluating and implementing technical solutions for remote processing, such as Galaxy Pulsar, OGC:WPS, or OGC:API Processes. The latter is implemented by D4Sciences on their framework called the Cloud Computing Platform (CCP). The openEO gateway also enables the execution of remote processes (https://openeo.org/documentation/1.0/processes.html).
- Evaluating efficient remote data access frameworks, such as iRODS, , Rucio, or ARCO data (e.g. Cloud-optimised GeoTIFF, Zarr, parquet) coupled with object storage (e.g. S3 protocol) which has been used by various communities.

In order to simplify the use of the EAL, additional utilities will be developed or integrated. For example, these will enable the EAL to reproduce a complete environment from a git repository (e.g. Repo2Docker), or (pre)generate metadata from data file(s) in an human readable format such as YAML.

5.3. Data Sciences IDE

The EAL will be flexible and propose the Data Sciences IDEs that meet the needs of the users. The choice between the Data Sciences IDEs often depends on the preferred programming language.

5.3.1. JupyterLab

JupyterLab (https://jupyterlab.readthedocs.io/en/latest/) is a web-based interactive computing environment that serves as the next-generation user interface for Jupyter Notebook. It provides a flexible and powerful environment for data analysis, scientific computing, and collaboration. JupyterLab builds upon the core functionality of Jupyter Notebook while introducing new features and enhancements.

JupyterLab supports the use of kernels to execute code in different programming languages (e.g. Python, R, Julia) and offers integration with popular libraries and frameworks. It provides a comprehensive and customisable environment. It also offers an intuitive and flexible workspace that empowers users to explore, analyse, and communicate data effectively.

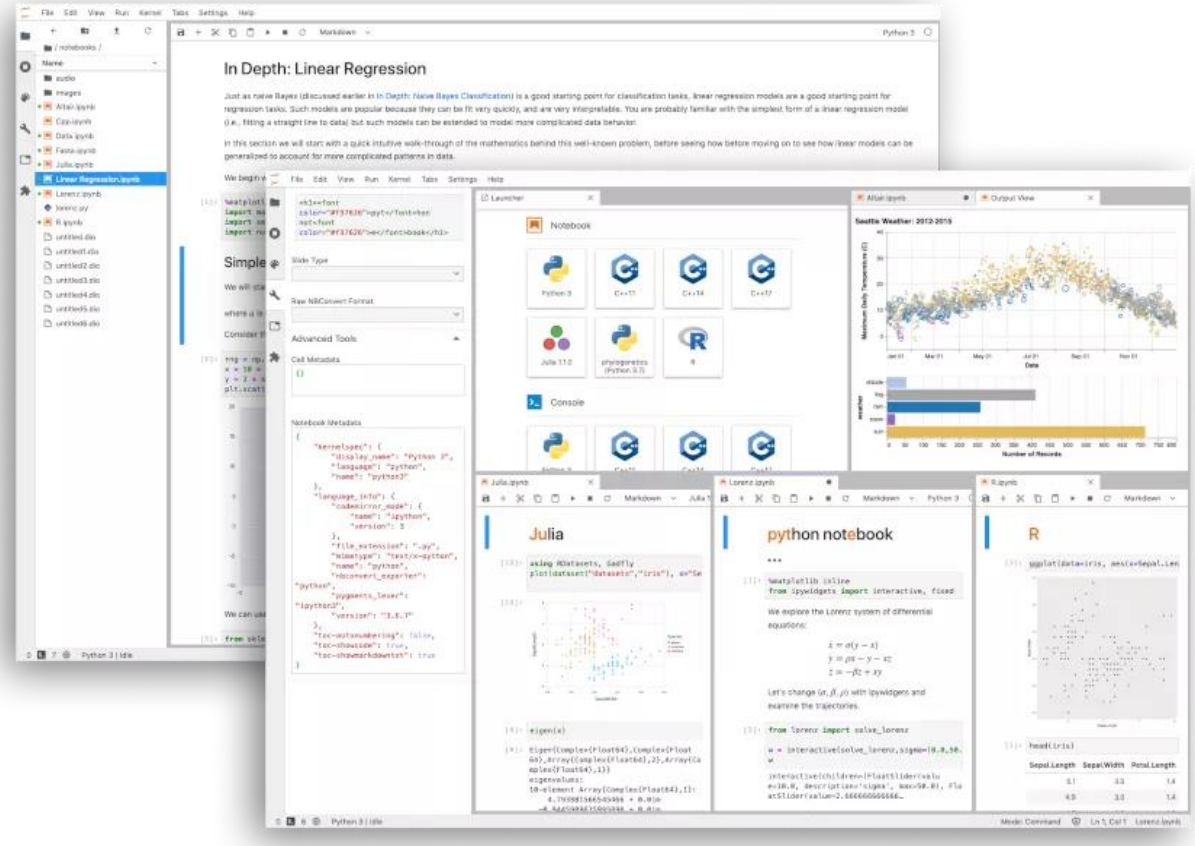


Figure 13 - JupyterLab screenshots

5.3.2. Pluto.jl

Pluto.jl (https://plutojl.org) is an interactive and reactive notebook environment for the Julia programming language. It aims to provide a flexible and dynamic platform for data exploration, analysis, and scientific computing. Pluto.jl introduces a new paradigm for notebooks, emphasising interactivity and reactivity, which allows for a more exploratory and interactive coding experience.

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, particularly for the Julia programming language community.

In this example, changing the parameter A and running the first cell will directly re-evaluate the second cell and display the new plot.

5.4. Data Exploration and Visualisation

5.4.1. WebODV

WebODV is the online version of the popular Ocean Data View (ODV) application (Schlitzer 2002), and allows interactive analysis and visualisation of a wide range of oceanographic and other environmental data in the user's web browser. In a typical deployment, a special version of the ODV software containing a websocket server for low-latency communication with the user is running on a dedicated server that is also hosting all the datasets to be served to users. Individual datasets are accessed via unique URLs in the user's web browser. No separate software needs to be installed on the user's side and no data needs to be downloaded.

The webODV user interface in the web browser is designed to mimic the user interface of ODV, thereby allowing the large number of previous ODV users to feel familiar and be productive from the start. Like the ODV software itself, webODV provides a rich interactive feature set via context sensitive menus and lets users create a wide variety of plot types as shown in the figure below.

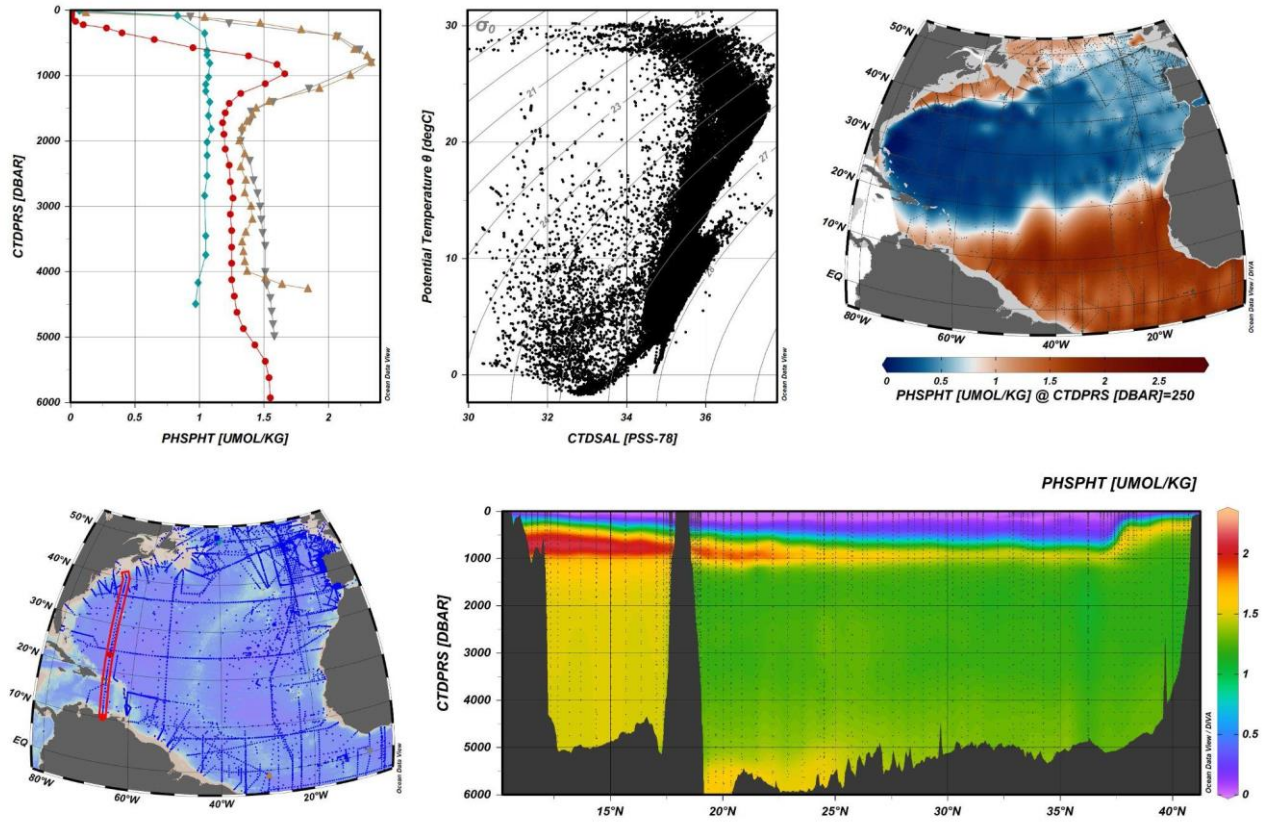


Figure 14 - WebODV screenshot

A number of operational webODV instances have already been established, serving large community datasets for a wide range of disciplines and projects. Examples are: explore.webodv.awi.de, geotraces.webodv.awi.de, emodnet-chemistry.webodv.awi.de, webodv-egi-ace.cloud.ba.infn.it, mvre.webodv.cloud.awi.de.

Within the FAIR-EASE project we will establish another webODV server that will focus on satellite data as well as on model and reanalysis output. These data are used by the Coastal Water Dynamics and other use cases such as the Marine Omics Observation pilot.

5.4.2. Examind - a GIS application

A lot of GIS tools exist for exploring geospatial data. These tools are agnostic of a specific thematic but are built for handling efficiently the geo-spatial dimension.

The majority of these tools implement OGC standards. Use of these standards allows users to easily combine different data sources coming from various communities. Of course, the price of this genericity is the lack of support for the specific needs of a given community.

The Examind Community software (https://community.examind.com/fr), supported by Geomatys, allows one to deliver a GIS data source by using standard OGC web services: WMS/WMTS for map viewing, WCS/WFS for subsetting downloading, SOS and SensorThings API for time series, WPS for on demand processing, CSW for metadata, and 3D Tiles for 3D dimensional visualisation.

All the visualisation clients that implement these standards can use the above data services without any additional development. For example, the OGC web services can be used by very common GIS software such as QGIS, or by classical cartographic viewers such as MapLibre, OpenLayers, and Ceisum.js.

Recently, the OGC published a major version of these APIs[6] in order to further implement the FAIR principles and to have a REST point more compatible with newer web technologies. It might be interesting to examine the contribution of these new APIs in the context of FAIR-Ease.

5.5. Galaxy

Even though the Galaxy workflow platform is a suitable solution for a lot of functions expected to be implemented in the EAL, it needs to be integrated into a wider system leveraging personal storage, vault management, data catalogue access, and data interfacing. From this perspective, multiple connections must be made, among them:

- Enabling Galaxy to tap into the EAL resources which necessitates authorisation mechanisms to authenticate EAL users
- Ensuring smooth data retrieval and transmission between Galaxy and the Data & Files Management module

This integration approach recognises the importance of not relying solely on a single tool for addressing various concepts, offering a spectrum of potential solutions.

Since 2005, the Galaxy project has fostered a global community focused on achieving accessible, reproducible, and collaborative research. This online platform designed for research facilitates the

[6] https://ogcapi.ogc.org/

sharing, development, and utilisation of diverse datasets and processing tools, all of which are freely accessible. Galaxy is deeply aligned with principles that closely relate to FAIR:

- Findable: all tools and workflows are available on web-based platforms, such as Galaxy ToolShed (https://toolshed.g2.bx.psu.edu/). Additionally, tools or scripts are shared on platforms like GitHub, ensuring their discoverability.

- Accessible: Galaxy enables the exploitation of tools and data without requiring programming expertise, making them easily usable and comprehensible for fellow researchers. The distribution of tools, workflows, and tutorials under open source licences, coupled with an intuitive interface, promotes straightforward access to analytical components and processes.

- Interoperable: Galaxy simplifies the reuse of analysis workflows, offering a comprehensive toolbox for data analysis. Users can perform intricate data analysis tasks in a reproducible and automated manner. The integration of the Conda package management system and the workflow-oriented functionality of Galaxy enables seamless interaction between different tools, regardless of the programming language they use. The platform aids users by suggesting commonly used tools based on the analysis history. Users can create and share their workflows adhering to standards like CWL (https://www.commonwl.org/) or RO-CRATE (https://www.researchobject.org/ro-crate/).

- Reusable: Galaxy aligns with FAIR principles by providing a cohesive system for the citation of both data and software. It assigns a unique identifier to each numeric object, ensuring proper attribution when reused. This guarantees acknowledgment and recognition for contributors. Furthermore, Galaxy enhances reusability by allowing tools to be re-executed when necessary. Tools and workflows are available for installation on any Galaxy instance.

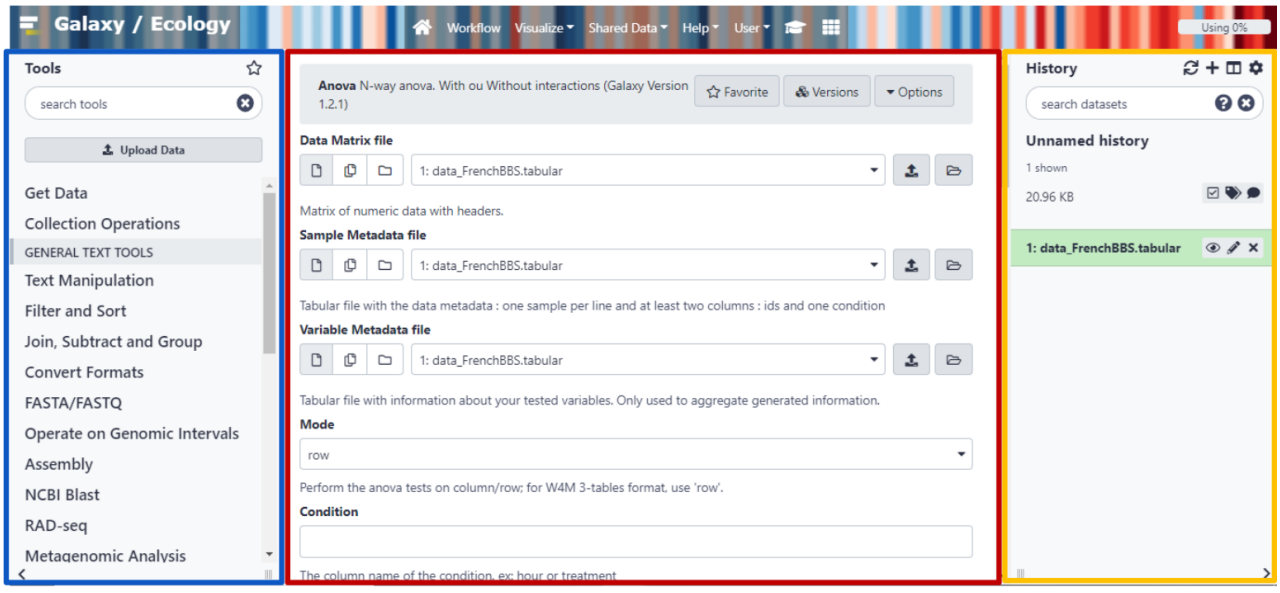


Figure 15 - Galaxy users' interface ecology.usegalaxy.eu.

In previous Figure, the blue square on the left corresponds to the list of tools, the red square in the middle corresponds to the interface of the current tool and the yellow square on the right corresponds to Galaxy history (sequence of tasks).

The Galaxy Training Network significantly contributes to enhancing the accessibility and reusability of tools and workflows. The Galaxy Training platform (https://training.galaxyproject.org/) hosts an extensive collection of tutorials. These tutorials serve as valuable resources for individuals seeking to learn how to navigate Galaxy, employ specific functionalities like Interactive Tools, or execute workflows for specific analyses.

6. Conclusion

The design of the EAL is ambitious, going beyond the project's intended lifespan. It represents not merely a collection of disconnected tools but rather a comprehensive and collaborative framework allowing setup of data workflows involving multiple tools. To draw a parallel with enterprise collaboration solutions like email, calendars, documentation, and video conferencing, true added value emerges when these components are seamlessly integrated.

Our primary focus will be on developing a MVP and initiating iterative cycles of improvement. This approach allows us to learn from practical implementation and user feedback, ensuring that the most critical features are prioritised and our project's scope is refined as needed.

Crucially, our efforts should revolve around data. This encompasses tasks such as ensuring smooth access to data sources, data transformation, data subsetting, data processing, and data visualisation. A robust data strategy will underpin the success of our project, as data is the foundation of many modern solutions.

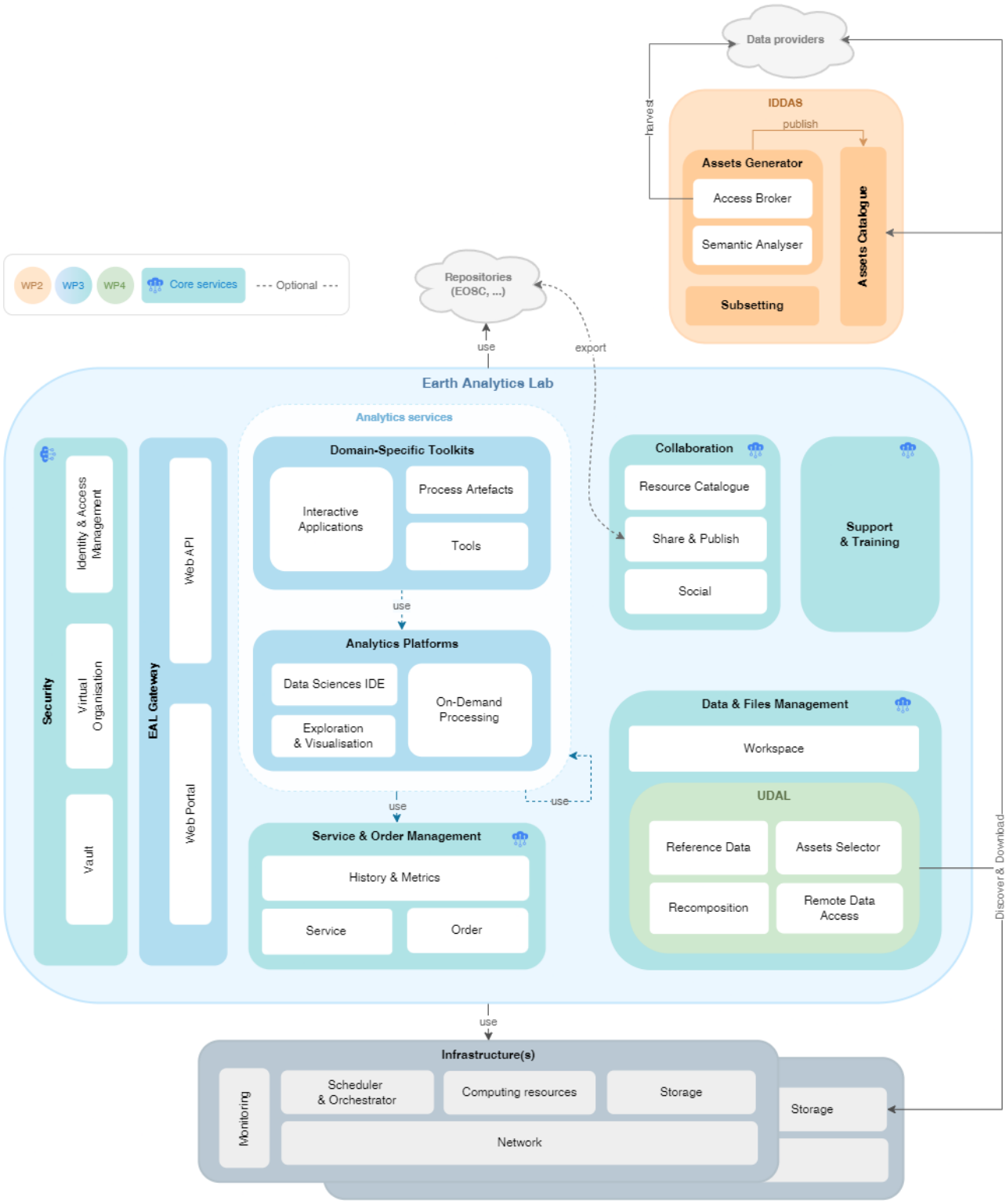
In addition to working on the technical aspects, dedicated time will be spent on establishing best practices and crafting effective training materials. These resources will not only facilitate smoother project development but also ensure that our team is well-equipped to handle the challenges that arise during implementation.

To optimise efficiency and resource allocation, existing tools and e-infrastructures should be leveraged wherever possible. This approach not only reduces development time but also capitalises on the expertise and resources already available in the ecosystem.

In summary, our project's success hinges on a balanced approach that prioritises an MVP, robust data management, best practices, and the utilisation of existing tools and resources. By following these principles, it is possible to meet the challenges and deliver a sustainable and effective solution.

7. Appendix

7.1. FAIR-EASE general architecture diagram



7.2. Example of research process stages

- Problem Definition: this stage involves clearly defining the research problem or objective. It includes identifying the specific question(s) to be answered, the scope of the analysis, and the desired outcomes.
- Data Discovery and Access: this stage involves identifying and obtaining relevant data for the study, ensuring its quality, and complying with ethical considerations. It plays a critical role in the research process, as they lay the foundation for generating meaningful and reliable insights. Properly identifying, accessing, and managing data ensures that research findings are based on sound data-driven evidence.
- Data Exploration and Preprocessing: this stage focuses on understanding and preparing the acquired data for analysis. It involves data cleaning, handling missing values, removing outliers, and transforming the data into a suitable format. Exploratory data analysis techniques are often applied to gain initial insights and identify patterns or issues in the data.
- Data Analysis: the data analysis stage involves applying various statistical, mathematical, and computational techniques to extract insights from the data. This can include methods such as regression analysis, spatial analysis, time series analysis, clustering, classification, or machine learning algorithms, depending on the specific research goals.
- Interpretation and Modeling: once the data analysis is performed, the results are interpreted and modelled to generate meaningful interpretations and conclusions. This stage involves drawing inferences, making predictions, or creating models to explain phenomena related to the Earth System.
- Visualisation and Communication: the results and findings from the analysis are visualised in this stage to effectively communicate the insights to stakeholders, researchers, or decision-makers. This can include creating maps, charts, graphs, or interactive visualisations that help convey the complex information in a clear and understandable manner.
- Reporting and Documentation: it is crucial to document the entire process, including the methodologies, assumptions, and limitations of the analysis. This stage involves preparing comprehensive reports, technical documentation, and research papers to share the findings with the scientific community and other interested parties.
- Iteration and Feedback: the stages mentioned above are often iterative in nature. Feedback from peers, subject matter experts, or collaborators is valuable for refining the analysis, revisiting certain steps, or incorporating additional data or techniques.

7.3. Common operations performed on scientific data

- Subsetting: creates a smaller subset of data by selecting specific variables or dimensions of interest while retaining all the data points or observations. It involves selecting a subset of variables or dimensions from a larger dataset while keeping the same set of observations. Subsetting is often used to simplify the dataset by focusing on a specific subset of variables or dimensions that are relevant to the analysis or research question. It does not involve

removing or filtering out data points; instead, it selects a reduced set of variables or dimensions.

- Filtering: involves selecting a subset of data based on specific criteria or conditions. It focuses on retaining only the data points or observations that meet certain predefined criteria. Filtering is commonly used to extract a subset of data that satisfies specific requirements, such as selecting data within a certain range of values, including specific categories, or meeting specific logical conditions. The filtering process removes data points that do not meet the specified criteria, resulting in a reduced dataset.

- Transformation: involves changing the representation or organisation of data while preserving its meaning and integrity. Data format transformation is commonly performed to make data compatible with different systems, software, or analytical tools, allowing seamless data integration and analysis. This transformation may include converting data between file formats (e.g., NetCDF to ODV), raw data files to datacube, and more. The goal of data transformation is to enable efficient data exchange, interoperability, and utilisation across various platforms or applications.

- Harmonisation: involves aligning variables, units, formats, and other data attributes to enable meaningful analysis and comparison. By harmonising data, researchers can combine datasets, identify patterns, and draw accurate conclusions. Data harmonisation is particularly crucial when working with diverse sources or heterogeneous data to ensure reliable and coherent analysis across the entire dataset.

- Colocation: analysis of spatial or temporal overlap between different datasets or variables. This involves identifying locations or time periods where two or more variables coincide or intersect. This analysis helps reveal relationships, correlations, or dependencies between the variables being studied. By identifying colocation patterns, researchers can gain insights into how different factors or variables interact or influence each other in specific regions or time frames. Data colocation is commonly used in fields such as remote sensing, climatology, ecology, and geospatial analysis to understand the interconnectedness of various phenomena and make informed decisions based on these relationships.

- Coregistration : aligns multiple images or datasets to a common coordinate system. This helps ensure that different images or data sources are properly matched and can be compared or combined for analysis.

- Reduction : reduces the volume, complexity, or dimensions of a dataset while preserving its key information and characteristics. It involves techniques that condense or summarise the data to make it more manageable and easier to analyse. Data reduction methods can include aggregating data points, selecting a subset of variables or dimensions, applying statistical techniques for summarisation (e.g., mean, median), or employing dimensionality reduction techniques. The goal of data reduction is to simplify and streamline the dataset while retaining the essential insights and patterns, enabling more efficient analysis and interpretation

- Aggregation: can be thought of as a grouped reduce. Aggregation typically involves applying statistical operations such as averaging, summing, counting, or taking the maximum or

minimum values across specific groups, time intervals, or spatial regions. This process helps in gaining a broader perspective on the data, identifying trends, patterns, and general characteristics, and reducing the complexity of the dataset.

- Resampling : alters the sample size or distribution of a dataset. This can involve reducing the number of data points (downsampling) or increasing the number of data points (upsampling) through various techniques. Resampling methods are used to adjust the data representation to address specific requirements, such as handling class imbalances, improving computational efficiency, or creating balanced datasets for analysis or modelling.

- Interpolation/Extrapolation: Interpolation and extrapolation techniques are used to estimate values between or beyond the observed data points, respectively. Interpolation predicts values within the observed range, while extrapolation extends the prediction outside the observed range. These techniques are valuable for filling in missing data or extending the analysis beyond the available data range.

- Integration/Merging: Integration or merging combines multiple datasets or variables into a single dataset, typically based on common identifiers or key fields. It enables the combination of complementary information and facilitates joint analysis.

7.4. Key functions and benefits of data visualisation

- Data Exploration: data visualisation allows researchers to explore large and complex datasets more effectively. By transforming raw data into visual representations such as charts, graphs, maps, and interactive visualisations, scientists can identify patterns, trends, and anomalies that may not be apparent in raw numerical or textual formats. It enables them to gain deeper insights into the data and uncover meaningful relationships.

- Communication and Presentation: data visualisation helps researchers communicate their findings and insights to a broader audience. Visual representations make it easier for non-experts, such as policymakers, stakeholders, and the general public, to understand complex scientific concepts and research outcomes. Well-designed visualisations can convey information more intuitively, effectively conveying key messages and supporting data-driven decision-making.

- Pattern Recognition: visualising data allows researchers to identify patterns and correlations that might not be immediately apparent through statistical analysis alone. By representing data visually, spatial and temporal patterns can be observed, enabling scientists to detect trends, cycles, and anomalies in Earth processes. This helps in understanding phenomena such as climate patterns, ocean currents, atmospheric conditions, and ecosystem dynamics.

- Geospatial Analysis: Earth science research often involves analysing geospatial data, such as satellite imagery, land cover maps, or climate models. Data visualisation enables the overlaying and analysis of geospatial data on maps, allowing scientists to identify spatial patterns, hotspots, and spatial relationships between different variables. It helps in studying phenomena such earth erosions, urbanisation, habitat fragmentation, nutrients or pollutants availability or distribution, or the impact of natural disasters.

- Interactive Exploration: interactive data visualisations allow users to interact with the data, drill down into specific details, and customise the visualisation parameters. Researchers can create interactive dashboards or web-based tools that enable users to explore the data themselves, select different variables, filter data based on specific criteria, and visualise results in real-time. This promotes a deeper understanding of the data and encourages user engagement.