Project Title	FAIR EArth Sciences & Environment services
Project Acronym	FAIR-EASE
Grant Agreement No.	1010587
Start Date of Project	1/09/2022
Duration of Project	36 Months
Project Website	www.fairease.eu

1 D5.1 Report on Key Requirement from Use Case Pilots

Work Package	**WP5**
Lead Author (Org)	**Maria Luisa Chiusano (UNINA)**
Contributing Author(s) (Org)	**Pilot description:** - **Pilot 5.1.1: Reiner Schlitzer (AWI), Simona Simoncelli (INGV), Charles Troupin (ULiege)** - **Pilot 5.1.2: Giuliano Langella (UNINA), Fabio Terribile (UNINA)** - **Pilot 5.1.3: Nicolas Pascal (CNRS), Marie Boichu (CNRS), Raphaël Grandin (IPGP, Université Paris Cité)** - **Pilot 5.2.1: Virginie Racapé (Pokapok), Catherine Schmechtig (CNRS), Raphaëlle Sauzede (CNRS), Alban Sizun (Pokapok), Allessandra Giorgetti (OGS), Catalina Reyes (OGS)** - **Pilot 5.3.1: Cymon J. Cox (CCMAR), Katrina Exter (VLIZ), Marc Portier (VLIZ), Maria Luisa Chiusano (UNINA), Stelios Ninidakis (HCMR), Ioulia Santi (EMBRC)** **Other:** **- Luciano Bosso (SZN): Organization of Annexes; Formatting and Editing.**

Other Contributions (Org)	**- Luca Ambrosino (SZN); Marco Miralto (SZN): Text Review.** **Discussions with Pilot Members and with all members of the Technical Board contributed to conceive this document.**
Internal Reviewer(s) (Org)	**Vincent Breton (CNRS), Katrina Exter (VLIZ), Christelle Pierkot (CNRS), Alessandro Rizzo (IRD), Charles Troupin (ULiege)**
Due Date	**31.12.2022**
Date	**31.01.2023**
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)

2 Versioning and Contribution History

Version	Date	Author	Notes
0.1	15.11.2022	Maria Luisa Chiusano (UNINA)	Table of Content defined
0.2	10.12.2022	Maria Luisa Chiusano (UNINA) and Pilot reference authors	Partial written draft
0.3	20.01.2023	Maria Luisa Chiusano (UNINA) and Pilot reference authors	Largely complete draft
0.4	23.01.2023	Maria Luisa Chiusano (UNINA) and all co-authors	Complete draft ready
0.5	03.01.2023	Maria Luisa Chiusano (UNINA) and Pilot reference authors	Comments from Internal review added
0.6	16.01.2023	Maria Luisa Chiusano (UNINA) and Pilot reference authors	Updates with answers to internal reviewers complete
0.7	20.01.2023	Maria Luisa Chiusano (UNINA), Luciano Bosso (SZN).	Final revision, editing, cleaning and formatting complete
0.8	23.01.2023	All authors and contributors	Approval of the final document complete
0.9	27.01.2023	Maria Luisa Chiusano (UNINA) and the Technical Board	Approval of the document complete
1.0	31.01.2023	Corentin Lefevre (Neovia)	Final Edition

3 Table of Contents

Funded by the European Union

4 List of Figures

5 List of Tables

Funded by
the European Union

6 Terminology

Terminology /Acronym	Description
AAI	Authentication and Authorization Infrastructure
CF	Climate and Forecasting set of conventions (https://cfconventions.org), developed as a community standard for encoding data in netCDF. They were initially developed for model-generated data, but have been extended in Version 1.6 to cover observational data. The standard was initially developed by atmospheric modelers studying climate to ensure the data from different models were interoperable and intercomparable. CF was subsequently developed to cover oceanography and has been widely adopted by many oceanographic projects. CF provides guidelines and recommendations on how to place usage metadata information about data, within the netCDF file. This provides a definitive description of what the data in each variable represents, the spatial and temporal properties of the data and how a data value is representative of an interval or cell. As climate model data are often not simply representative of points in time and space, the standard allows describing coordinate intervals, multidimensional cells and climatological time coordinates.
CMS	Content Management System
DIVAnd	Data Interpolating Variational Analysis in n dimensions (https://github.com/gher-uliege/DIVAnd.jl)
EDMI	EOSC Dataset Minimum Information (https://eosc-edmi.github.io)
Earth Analytic Laboratory	Earth Analytic Laboratory to be implemented in FAIR-EASE
EOSC	European Open Science Cloud (https://eosc.eu)
EOV	Essential Ocean Variable
EMO BON	European Marine Omics Biodiversity Observation Network (https://www.embrc.eu/emo-bon)
ESDAC	European Soil Data Centre
EUSO	EU Soil Observatory
FAIR	Findable, Accessible, Interoperable and Reusable (https://www.go-fair.org/fair-principles)
GRIB	GRIdded Binary
HPC	High-Performance Computing
HTC	High-Throughput Computing
IaaS	Infrastructure as a Service
JRC	Joint Research Centre

Funded by the European Union

Terminology /Acronym	Description
JSON-LD	JavaScript Object Notation for Linked Data
LIMS	Laboratory Information Management System
NetCDF	Network Common Data Form. Software and self-describing, machine-independent data formats (https://www.unidata.ucar.edu/software/netcdf/)
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting (https://www.openarchives.org/pmh)
ODV	Ocean Data View (https://odv.awi.de/)
ODV Collection	Binary data storage format optimized for dense storage and fast access (Schlitzer, 2002)
RDA	Research Data Alliance (https://www.rd-alliance.org)
RI	Research infrastructure
SDN	Software Defined Networking
SAR	Synthetic Aperture Radar
inSAR	Interferometric Synthetic Aperture Radar
SENTINEL	Earth observation missions, on behalf of the joint ESA/European Commission initiative Copernicus, that covers different aspect of Earth observation: Atmospheric, Oceanic, and Land monitoring https://sentinel.esa.int
SOURCE	Sea Observations Utility for Reprocessing, Calibration and Evaluation (https://zenodo.org/record/6319836)
UC	Use Case
webODV	Online data analysis and visualization service based on ODV (https://webodv.awi.de/)
WP	Work Package

Funded by the European Union

7 Executive Summary

WP5 (Work Package) is in charge of coordinating the activities tackling three different use cases (UCs): 1) the Earth and Environment dynamics; 2) The Environmental Biogeochemical Asset; 3) The Biodiversity Observations. As documented in the FAIR-EASE work plan, all the use cases represent “Real Life-Science” challenges. During the progress of the project, specific Pilots will be considered per UC, each of them addressing specific objectives.

Specifically, this deliverable reports on the assessment of the key requirements and the mapping of main resources to be considered in the project to fulfill the aims and scope of the Pilots included in each use case.

The first use case (UC5.1), Earth and Environment Dynamics, involves three different Pilots: 1.1) the Coastal Waters Dynamics, 1.2) the Earth Critical Zones, and 1.3) the Volcano Space Observatory. Each of the Pilot focuses on different terrestrial systems, involving relevant scientific communities, each one proposes specific demonstrators and objectives in FAIR-EASE. General focusing is on the implementation/improvement of tools and web-based portals to solve relevant community specific needs for the observations and the analysis of different Earth systems. Specifically, the focus will be on the marine coastal environment near the river estuaries (Pilot 5.3.1) and on the volcano space (5.3.3). In addition, the improvement of methodologies for the assessment of land and soil degradation, such as erosion, loss of organic matter and biodiversity, contamination, and fulfill the Sustainable Development Goal (SDG) 15.3 (“combat desertification, restore degraded land and soil, and achieve a land degradation-neutral world”), will be considered too (Pilot 5.3.2).

The second use case (UC5.2), The Environmental Biogeochemical Asset, focuses on crucial challenges in different environmental frameworks and, moreover, involves methodological issues that are of general interest in the project. Indeed, the definition of standards, technologies and quality control approaches for Biogeochemical (BGC) data acquisition is a relevant issue to the aim of adjustment, validations and organization of reference collections and methodologies that can be exploited to support and improve our understanding of current, past and future conditions of earth and environment. The Pilot will focus on data and methods from marine environments.

The third use case (UC5.3), Biodiversity Observations, faces the challenge to characterize the biodiversity asset on Earth, monitoring and predicting changes that will impact health and socio-economic interests. Specifically, the Pilot here will focus on the definition of products and services for omics-oriented biodiversity observations in the marine environment, exploiting the ongoing effort in the European Marine Omics Biodiversity Observation Network (EMOBON).

Several actions were deployed to produce this deliverable, organized as follows:

- the main efforts were focused on the tuning among all partners in WP5, and with the other five WPs included in the project.

- The first goal was to define the most appropriate way to present the multifaceted aspects embedded in WP5 UCs, describing the different Pilots with the aim to highlight their objectives and priorities on technological needs, together with their requirements and expectations for the technical implementations of demonstrators in FAIR-EASE. The D5.1 schema is intended to fulfill this necessary uniformity through the different actions that are herein described (D5.1 Main text).
- The mapping of the major data resources and methods was started by setting up a shared template in table format (see Annex A and Tables 1 and 2 as a summary), that was filled by the Pilot reference people to report information from each community. The template is planned with enough depth of information to allow a general understanding even by non domain experts. The Annex A aims to provide a preliminary list of data resources and methods to be considered in FAIR-EASE and the listing in the shared tables will be considered a work in progress throughout all the project. Furthermore, these information will be reported into the FAIR-EASE DMP (D1.2; D1.3; and D1.6) whose purpose is to ensure the availability and utility of the project's data, re-used as well as created, based on FAIR policies throughout the entire lifecycle of the project and beyond.
- Driving all the Pilots through this general introductory effort was also useful to start to shape a common language and favor a transdisciplinary interaction paving the way to a more interactive exchange among partners with expertises from different domains. Indeed, while Pilot specific communities with highly specialized expertises needed to remain focused on their specific plans and on the design of their needs, establishing their interaction with the technical WPs, FAIR-EASE is also aiming to focus on the trans-domain dialog and needs. This requires planning on how to exploit useful exchanges and share of knowledge and experiences among different Pilots, to look for possible emerging goals, based on the progressing maturation of the whole FAIR-EASE community towards a common understanding, with the support and collaboration of the technical WPs.

This document should be considered as the result of a crucial preliminary step that will also contribute to the goal described in WP5 task 5.4, i.e. the establishment of trans-domain interoperability and possible synergies from transdisciplinary efforts. These aspects are expected to evolve during the progress of the entire project thanks to a prolific interaction among all partners.

Funded by
the European Union

8 Introduction

The study of Earth complexity and all related aspects is a huge complex effort. The observation and the understanding of natural or man forged assets and the associated changes in space and time at different scales require large, multifaceted datasets covering different Earth systems and different topics. These analyses are supported by different data, coming from different streams and observing technologies and generally from multidisciplinary contexts. Often researchers focused on a single aspect make use of multiple datasets simultaneously, and scientists focused on different domains exploit similar or the same resources with different methodologies.

In the scope of FAIR-EASE, five different Pilots, embedded in three use cases representing "Real Life-Science" challenges, aim to focus on different topics dealing with different Earth systems and tackling different technological and methodological issues. They all take into account the entire chain of data management, from acquisition to exploitation and the final delivery to end users. This permits to address current limits and opportunities in dataset production and accessibility, sharing and analytics, from different domains while improving workflows and establishing cross-domain agreements on reference data collections, standards and analytical strategies that will satisfy both single Pilots from specialized communities in FAIR-EASE, and provide useful operational frameworks for similar needs to external stake-holders.

Pilots in WP5 are expected to:

- **<u>define and agree on:</u>**

- resources and standards needed to face the challenges in the UCs;

- the computational strategies and requirements to feed WP2, WP3 and WP4;

- **<u>test and validate:</u>**

- the service implemented by WP2,WP3 and WP4;

- FAIRness of the data resources and methodologies (WP6);

- **<u>contribute to:</u>**

- dissemination, external user engagement (WP6);

- long term sustainability (WP1);

- **<u>exploit:</u>**

- trans-domain interoperability;

- synergies emerging from a transdisciplinary effort (T5.4);

In particular, to find an agreement on resources, standards and needs to face the specific challenges proposed by the different Pilots per Use Case, it was necessary to fully engage all members of each Pilot from the early stages of the project, driving them to provide an exhaustive description of the domain that characterizes each Pilot and of their specific objectives in the project. This last action included the need for a clear description for non-experts of the state of the art in each field, and the mapping of main resources and methodologies currently used among those available, describing also novel opportunities and expectations for improvements to be fulfilled by each Pilot during the progress of the project. This was the first step to allow the tuning among all partners involved in WP5 and to start a fruitful interaction with other WPs in the project.

The initial four months from the start of the project were therefore mainly focused on the start up of the activities with a series of meetings to favor the crosstalk among Pilots and with the technical WPs (WP2, 3 and 4), while planning the expected design of the computational strategies necessary for the implementation of best solutions for useful demonstrators during the progress of the project.

The document presented here is organized in 5 main Chapters (Chapter 9, 10, 11, 12, 13) that describe the background, the aims and the technical requirements and needs from each Pilot per use case.

Each chapter was contributed by the Pilot reference people and organized with the same schema. Chapter 14 recapitulates the whole context and poses the main aspects that need to be considered to plan the future actions in FAIR-EASE. It highlights possible cross domain needs that can be fulfilled by joint efforts under the FAIR-EASE umbrella. This is considered the start to fulfill the major expected goal in FAIR-EASE: the exchange of experience and expertises for the analysis of novel opportunities to improve data and methodologies usage, interoperability, and reusability in the framework of relevant "Real Life-Science" challenges.

9 USE CASE 5.1: Earth and Environmental Dynamics – Pilot 5.1.1: Coastal Waters Dynamics

9.1 GENERAL DESCRIPTION AND BACKGROUND

The Earth and Environmental Dynamics UC is a huge complex topic, according to the Earth system under analysis and the specific objectives under consideration. FAIR-EASE planned to include in this effort three different Pilots, each one focusing on different Earth systems, each one having specific goals to achieve.

The first Pilot is described in this chapter. It deals with coastal water dynamics, and focuses on the coastal marine environment near river estuaries, where important processes, such as the evolution of plankton blooms or the transport and fate of nutrients, carbon and contaminants, critically depend on many factors, including river discharge, ocean circulation, meteorological conditions as well as biogeochemical processes. Specifically, the proposed demonstrator will focus on the Po river discharging into the Northern Adriatic.

9.1.1 State of the Art in the Specific Domain

The Northern Adriatic has been monitored well over decades and numerous research activities covering all fields of environmental science are ongoing (Umgiesser et al, 2022; Ricci et al, 2022; Bonaldo et al, 2019).

9.1.2 Domain Specific Data Resources

Due to economic (fisheries, tourism), environmental (carbon cycle, pollutants, plastic) and scientific reasons, coastal regions, such as the Northern Adriatic, are usually well monitored. As a result, many diverse datasets, sometimes reaching back more than 100 years, exist and are publicly available. This includes data, sometimes daily observations, from dense station networks recording river discharge, as well as meteorological and oceanographic conditions. In addition, there are now more than three decades of data from satellite sensors measuring physical and biological properties of ocean surface waters, and a wide range of physical and biogeochemical numerical models, re-analysis and gridding systems producing nearly continuous sea state estimates in space and time.

The collections are accessible from different public resources, of which the main references are:

- https://data.marine.copernicus.eu/products
- https://explore.webodv.awi.de/
- https://emodnet-chemistry.webodv.awi.de/
- https://webodv-egi-ace.cloud.ba.infn.it/
- https://www.seadatanet.org/Products/Aggregated-datasets

Funded by the European Union

In all environmental disciplines, large community datasets of given types are being created by collecting, harmonizing and aggregating the individual observations. This work is conducted either by dedicated agencies or by groups of specialized scientists.

9.1.3 Main Standards

Community datasets are distributed in a variety of formats, including ASCII spreadsheets, netCDF or ODV Collection for use with the popular *Ocean Data View* (ODV) software (Schlitzer, 2015; Schlitzer, 2002; Schlitzer, 2000; Brown, 1998). The datasets used in this Pilot are either Climate and Forecast (CF) compliant netCDF or ODV Collection.

9.1.4 Software, Workflows, Containers

While the existence of a wide range of large aggregated datasets is definitely positive, the large variety of data types, formats and metadata standards makes data usage overly complicated and hinders the combined evaluation and the correlation of data from different disciplines and sources. Finding, retrieving and maintaining up-to-date copies of the many large datasets on end user's computers is labor intensive and ineffective. Present analysis and visualization software systems, such as ncview (http://meteora.ucsd.edu/~pierce/ncview_home_page.html), xarray (https://docs.xarray.dev) or Holoviews (https://holoviews.org), mostly deal with single datasets at a time or require programming skills to use them. Users are typically left with devising their own data workflows, making analysis of data connections and correlations between multidisciplinary datasets a time-consuming and inefficient task.

The webODV (https://webodv.awi.de/), DIVAnd (Barth et al., 2014 - https://github.com/gher-uliege/DIVAnd.jl) and SOURCE (Oliveri et al., 2022 - https://doi.org/10.3389/fmars.2021.750387) tools are provided as online services and can be used in web browsers or JupyterLab environments.

9.1.5 Types of Applications

The aim is to develop new and innovative web services and online tools that enable and facilitate seamless access to multidisciplinary environmental datasets and to provide new ways of connecting and correlating data. The new services will be applied to science questions, and their utility and relevance will be evaluated by project partners and external users.

9.1.6 Types of Expected Users

Earth system scientists across all disciplines and intermediate and advanced students. The webODV interactive analysis and visualization tool is easy to use and may also appeal to the scientifically literate public.

9.2 PECULIARITY OF THE PILOT AND AIMS

9.2.1 Core Mission Statement in FAIR-EASE

The Pilot here proposed focuses on the coastal marine environment near river estuaries, in particular the Po river discharging into the Northern Adriatic.

With the aim of supporting users in the exploitation of multiple, heterogeneous datasets, the Coastal Dynamics Pilot will significantly extend and enhance the capabilities of two existing well-proven, operational services: (1) webODV for on-line interactive analysis and visualization of environmental data; (2) DIVAnd for intelligent spatial and temporal gridding of oceanographic data; and (3) a new proposed web service based on the SOURCE software tool, for the intercomparison of model results with time series observations of essential ocean variables (EOV).

9.2.2 Proposed Applications and Demonstrators

webODV, DIVAnd and SOURCE are components of a common workflow to process multidisciplinary datasets: in addition to the data analysis and visualization services, webODV will also be used to prepare, combine and subset datasets, creating as output a set of netCDF files. Those netCDF files will then be read by DIVAnd for the preparation of gridded products, also written as netCDF files. SOURCE will also create outputs in netCDF format that can be used in both DIVAnd and webODV for the multidisciplinary analysis of different data types. webODV and DIVAnd are designed to work with any type of geo-referenced data, not only oceanography, hence these tools are expected to benefit environmental and climate research in general.

9.2.3 Aims and Scope of the Applications and Demonstrators

The aim is to facilitate seamless access to multidisciplinary environmental datasets and to provide new ways of connecting and correlating data with specific attention to the Po river estuary in the Northern Adriatic sea.

9.2.4 Main Targets

The services developed by Pilot 5.1.1 are targeted towards earth system scientists across all disciplines and to intermediate and advanced students. The webODV interactive analysis and visualization tool is easy to use and may also appeal to the scientifically literate public. The online services are targeted to the research community as well as to decision makers and members of the general public familiar with environmental systems and their data. DIVAnd has been designed for gridding of oceanographic observations but can be used without any adaptation to other types of geo-referenced data, for instance land temperature.

SOURCE has been designed for the research community to facilitate the intercomparison of observations with freely available ocean models' data and their correct successive usage for the ocean state assessment. This will boost the access, extraction and integration of model data into the multidisciplinary environmental assessment in the coastal region.

9.2.5 Data Flow Diagram and Description

Data flow diagram of the PILOT 5.1.1: COASTAL WATERS DYNAMICS (Fig. 1).

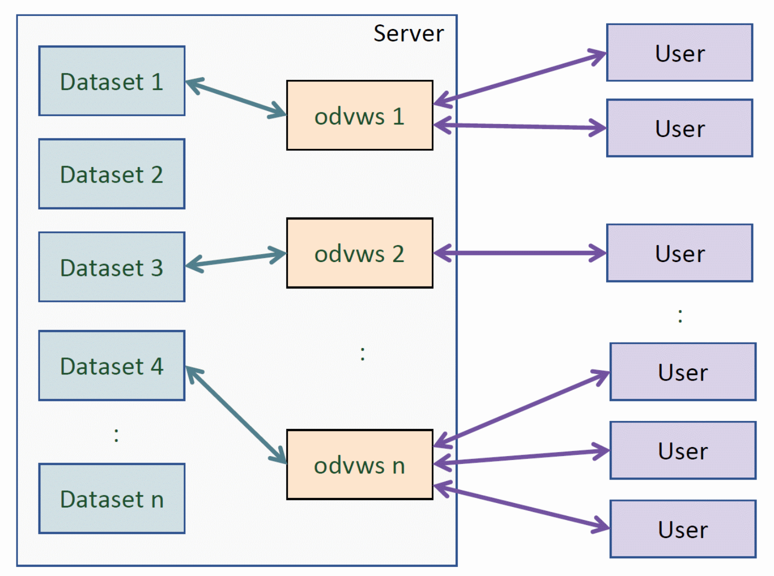

Figure 1 - webODV DATA HUB Architecture

9.2.6 Key Resource Data and Datasets

In all three cases, emphasis is on facilitated use of multidisciplinary datasets, including rivers, meteorology, physical and biogeochemical ocean data for the entire water column, satellite sensor data and output from physical and biogeochemical models. Example datasets are listed in the Annex A.

9.2.7 Key Metadata Descriptors

CF convention (https://cfconventions.org/) for satellite data and model output; variety of metadata definitions used by the different observational datasets.

9.2.8 Expected Services/Applications in the FAIR-EASE

The major technical/development activities promoted by the Pilot will be i) the implementation of a webODV data hub; ii) the implementation/expansion of DIVAnd and iii) the improvement of the software SOURCE. All these components are deployed as part of the Earth Analytic Lab within WP3.

Implementation of a webODV data hub

This Pilot aims to establish a webODV data hub as part of the WP3 Visualization services task. Similar to other operational webODV data hubs, such as at https://explore.webodv.awi.de/, https://emodnet-chemistry.webodv.awi.de/ and https://webodv-egi-ace.cloud.ba.infn.it/, this webODV deployment will consist of a specialized ODV server software (*odvws*) that not only provides the full ODV functionality, but also contains a WebSocket server used for communication with the user's browser window. In addition, this webODV instance will also consist of community datasets needed for the Coastal Dynamics Pilot not already served by other, existing webODV data hubs. This includes surface ocean temperature, chlorophyll and plankton data from satellite sensors as well as biogeochemical reanalysis and model forecast data. All these dataset are distributed as netCDF files.

Implementation/expansion of DIVAnd

This Pilot aims to offer users an access to DIVAnd configured to process observations in the coastal region. The formulation of DIVAnd, based on the minimisation of a cost function (Barth et al., 2014), allows one to add other constraints to the gridding process, for instance the advection. This means that the spatial correlation will be enhanced along streamlines if the user provides a velocity field covering the region of interest. Such an advection is particularly relevant in the case of a coastal application, where water properties near a river estuary are advected according to the currents

This is necessary because the distribution of in situ measurements is often scarce and mostly concentrated near the coast, while they are less available offshore. Using the information from a numerical model, in particular the current velocity at different time instances and vertical levels, would improve the quality of the gridded fields.

Improvement of the software SOURCE

The Pilot also aims to implement the SOURCE code (Oliveri et al., 2022 - https://doi.org/10.3389/fmars.2021.750387) in a Jupyter Lab in order to: access observations, process them (getting the main statistics), extract model data collocated in time and space with observations (synthetic data) and, intercompare them through simple metrics, such as bias, root mean square difference, correlation. This will provide contemporarily collocated observation and model based data for coastal ocean assessment and the estimate of model performance. This is necessary to facilitate the correct usage and integration of model data to the community not expert in ocean modeling.

SOURCE is designed for web applications/services that permits users to intercompare ocean models with in-situ observations in a selected spatial domain. Its present implementation covers the Mediterranean Basin and considers Near Real Time moored temperature and salinity observations downloaded from the Copernicus Marine Service (CMS) and two model products from the Istituto Nazionale di Geofisica e Vulcanologia (INGV, Italy). As an improvement, FAIR-EASE SOURCE service will access/extract/subset the data directly from the Copernicus Marine Service (CMS) using one of their recommended downloading techniques (see https://data.marine.copernicus.eu/product/MEDSEA_ANALYSISFORECAST_PHY_006_013/services) or through the ERDDAP service of EMODnet Physics (https://erddap.emodnet-physics.eu/erddap/index.html).

The code works in three modules: 1) the observation module that analyzes and reprocess the in situ data; 2) the model module to extract the synthetic data (model data at the exact time and location of observations) and; 3) the Calibration / Validation (CalVal) module that computes model skill scores to evaluate the model performance.

9.3 REPORT ON KEY REQUIREMENTS FROM THE PILOT

webODV data hubs (see Figure 1) provide online access to large numbers of large aggregated community datasets and allow interactive analysis and visualization of the data without the need to download datasets or install software locally. A webODV data hub server deployment consists of the community datasets to be served (in netCDF or ODV Collection formats) and an installation of a special variant of the ODV software (*odvws*) containing a Websocket server for communication with the user. *odvws* instances are started automatically whenever a user requests a given dataset. Access to the dataset (green arrows) can be in ReadOnly mode, suitable for data analysis and visualization, or in ReadWrite mode, which also allows data editing, quality control and dataset expansion.

Because of the interactive nature of webODV and due to the fact that most datasets are too large to be kept in memory as a whole, *odvws* is frequently reading (or writing in case of data editing) small parts of the data, typically as response to a user action, such as selection of a specific station or sample or when selecting options from context sensitive menus. In order to minimize latency and in the interest of a responsive, interactive user environment, it is critical to provide the fastest possible data access by *odvws*. This normally requires storing datasets on SSD disks connected in the fastest possible way to the processors running the *odvws* application. Placing the datasets on a slowly connected network reduces data access speeds and provokes the risk of frustrating end users. Bidirectional communication with the user (violet arrows) is via low latency Websocket messages triggered by client actions. Typically, Websocket traffic volume is small, and even standard DSL connections suffice.

The datasets used by webODV are large aggregated datasets created by the respective research communities. They will be used as distributed by the creators; any subsetting and filtering will be performed dynamically by *odvws* on request by the end user. The datasets come in two well established formats: (1) netCDF, optimized for gridded data, and (2) ODV Collection, optimized for heterogeneous, variable length environmental data.

Initially, the FAIR-EASE webODV data hub will allow users to analyze and visualize the datasets residing on this server. Access to datasets on other webODV data hubs and the combined usage of multiple, multidisciplinary datasets from the different distributed webODV hubs is a planned expansion of the system that will be added during the project lifetime.

Implementation/expansion of DIVAnd

DIVAnd is a software tool designed for the gridding of in situ measurements. The source code is available from GitHub at https://github.com/gher-uliege/DIVAnd.jl. The code is written in the Julia language (https://julialang.org/) and therefore a requirement is the installation of this language, at least the version 1.8.3, released in November 2022.

Along with DIVAnd, other Julia modules have to be installed, all of them can be installed via the Julia package manager:

- Julia, the Julia kernel for Jupyter;
- NCDatasets, DataStructures, CSV: modules to read and write different file formats;
- PhysOcean, a module with different utilities related to Physical Oceanography.
- PyPlot, a module for plotting, based on *Matplotlib (Python module).*
- Pluto, a module providing simple and reactive notebooks in Julia.

Other widespread Linux utilities are also needed: git, emacs-nox (terminal version of emacs is sufficient), libnetcdf-dev, netcdf-bin (with the tool ncdump), unzip, make, gcc, curl and nco.

In terms of hardware (to provide orders of magnitude), for large datasets a typical configuration that allows us to run out analysis in reasonable times: would be: for a 1/4° X 1/4° resolution on a domain covering all the European Seas, the analysis requires about 84 GB of RAM, 10 CPUs and 50 GB of disk space. If we focus on the Northern Adriatic Sea these numbers can obviously be reduced.

Following is a list of requirements for the applications to be run with DIVAnd:

(1) We need to ensure a fast connection to the datasets, possibly this can be achieved by the organization of an ad-hoc D*ata Lake*, whatever the number of datasets and their complexity.

(2) User input: ideally the user will request the datasets by specifying one or several variables (using the NODC vocabulary (https://vocab.nerc.ac.uk/; Vinci et al, 2011) or the CF standard names), a geographic bounding box and a time period.

(3) Output: one or several CF-compliant netCDF files, possibly following the contiguous ragged array representation (https://cfconventions.org/cf-conventions/cf-conventions.html# contiguous ragged array representation). In addition the netCDF files should use quality flags of the data with the relevant CF semantics (using a flag variable the attribute ancillary variables). Finally, they should also include information of the data originator (for traceability) in form of EDMO code (https://edmo.seadatanet.org/) of the data provider and local CDI ID (https://www.seadatanet.org/Metadata/CDI-Common-Data-Index).

(4) Content: we expect the content of the data lake not to be static, but to have the capability to evolve following our need, for instance if additional datasets have to be accessed (as this is typically the case during scientific applications).

(5) Accessing the data lake: a possible solution would be a RESTful API using the parameters specified in (2).

DIVAnd can be deployed in different ways: as a package in a Julia session; as a Docker container for JupyterHub (https://github.com/gher-uliege/DIVAnd-jupyterhub) or as a Singularity container (https://github.com/gher-uliege/DIVAnd-singularity).

Improvement of the software SOURCE

SOURCE is designed for web applications/services that permit to intercompare ocean models with in-situ observations in a selected spatial domain. Its present implementation covers the Mediterranean Basin and considers Near Real Time moored temperature and salinity observations downloaded from the Copernicus Marine Service (CMS) and two model products from INGV, but FAIR-EASE SOURCE cloud service will access/extract/subset the model data directly from CMS using one of their recommended services (see https://data.marine.copernicus.eu/product/MEDSEA_ANALYSISFORECAST_PHY_006_013/services), avoiding the downloading of huge amount of useless data.

A diagram of the proposed SOURCE workflow in the project Virtual Research Environment (VRE) is presented in Fig. 2 to exemplify the workflow in terms of data access and data storage.

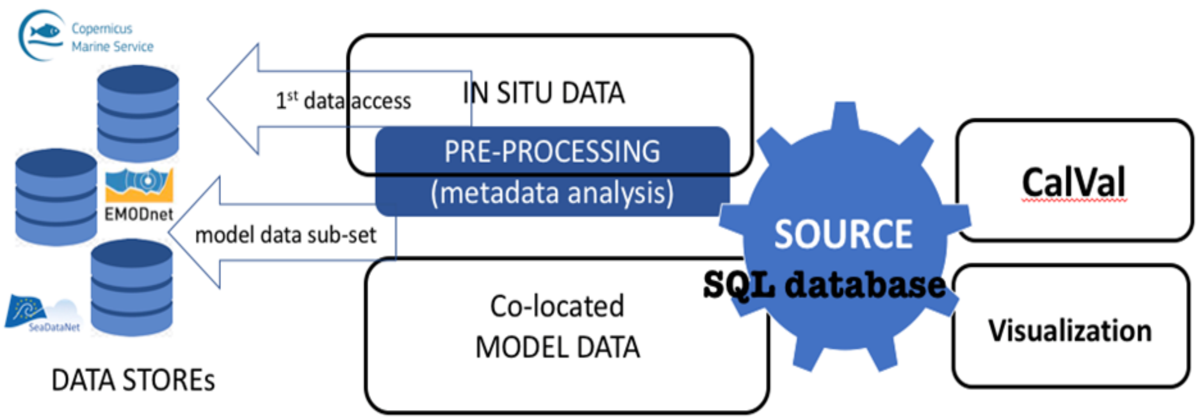

Figure 2 - Diagram of the expected VRE Source

The first data access should make available the desired in situ observations from the principal marine data infrastructures (Copernicus Marine Service, EMODnet Physics, SeaDataNet), then a data pre-processing should instruct the model data subsetting to get co-located synthetic (model) data to be inter-compared with observations in the SOURCE CalVal module. Data will be also visualized.

The SOURCE utility is written in Python and it uses several Python modules, such as:

– Vectorized numerical data analysis (numPy, sciPy, ObsPy, and pandas);
– Machine learning tools (scikit-learn);
– Hierarchical data storage (NetCDF-4) (HDF-5 extension);
– Relational metadata storage using Structured Query Language (SQL) as management system.

SOURCE is released for public use in the ZENODO platform (Oliveri and Simoncelli, 2022 - https://doi.org/10.5281/zenodo.6319836) with a Creative Commons CC-BY-SA-NC license.

The data needed to provide a first SOURCE release are listed in the Annex A.

9.3.1 Summary of Requirements

- webODV deployment serving satellite data as well as model and reanalysis output;
- Fast access to the datasets. The required community datasets are listed in the Appendix;
- NCDatasets, DataStructures, CSV: Julia modules to read and write different file formats;
- PhysOcean, a Julia module with different utilities related to Physical Oceanography;
- PyPlot, a Julia module for plotting, based on *Matplotlib (Python module*);
- Pluto, a Julia module providing simple and reactive notebooks;
- IJulia, the Julia kernel for Jupyter;
- Linux web server environment with nodejs;
- Standard Linux utilities: git, emacs-nox (terminal version of emacs is sufficient), libnetcdf-dev;
- netcdf-bin (with the tool ncdump), unzip, make, gcc, curl and nco;
- Python environment with numPy, sciPy, ObsPy, scikit-learn and pandas modules;
- Up-to-date server hardware, 32GB RAM, 8 or more vCPUs, ca. 200GB internal SSD storage.

9.4 POSSIBLE CROSS LINKS WITH OTHER PILOTS

The Coastal Dynamics and Ocean Bio-geochemical Observations Pilots are strongly linked, and extensive collaboration is planned in four main areas.

(1) A number of biogeochemical datasets, direct observations as well as reanalysis products, are used by both Pilots, and the workload for integrating these datasets into the FAIR-EASE data lake will be shared. (2) A webODV data hub for import, editing and quality control of biogeochemical data will be established collaboratively between the two Pilots. This hub will contain existing, validated biogeochemical data to facilitate the evaluation and quality control of new data from autonomous floats and gliders. Extensive use will be made of the wide range of manual quality control tools and the easy-to-use data flagging procedures provided by webODV. The webODV support for concurrent data editing, allowing more than one expert to work on a given dataset at the same time, will be exploited to speed-up the data evaluation process. (3) A webODV data hub for an adapted visualization framework for the validation and evaluation of the adjustment of the new quality controlled and adjusted biogeochemical data will be established. webODV will allow visualization of very heterogeneous biogeochemical data (i.e. satellite surface data vs. BGC-Argo in situ profiles data). This is an important point for the Ocean Bio-geochemical Observations Pilot, where one of the main goals is to qualify and validate the newly adjusted data by comparing it with all available nearby biogeochemical qualified data (for a given time/area).

(4) The newly quality controlled biogeochemical datasets can become datasets delivered by the FAIR-EASE project, presumably in the form of a data lake and then become available as

inputs for DIVAnd gridded products and for the webODV data analysis and visualization service. These datasets would be tagged as modified from the original data as the Argo submission of delayed mode data is Argo-specific, and while the FAIR-EASE project would facilitate the delayed mode submission, it could not bypass this workflow. The new biogeochemical data will also be used by SOURCE for improved model data comparison.

10 USE CASE 5.1: Earth and Environmental Dynamics – PILOT 5.1.2: Earth Critical Zone

10.1 GENERAL DESCRIPTION AND BACKGROUND

Still in the UC 5.1, Earth and Environmental Dynamics, this Pilot will focus on analyzing and visualizing the Earth Critical Zone (ECZ) system. The specific aim in the Pilot is to improve the assessment of land and soil degradation, such as erosion, loss of organic matter and soil biodiversity, compaction, salinization, landslides, contamination, sealing, desertification and climate change.

10.1.1 State of the Art in the Specific Domain

Over the last few decades, land and soil degradation processes have increased significantly. A recent review by the EU Soil Health and Food Mission Board and JRC (Wilkki & Reeve, 2020) showed that in 2020, about 60-70% of the soils were in an unhealthy state.

To challenge this major problem, the United Nations 2030 agenda, developed the Sustainable Development Goals (SDG) 15.3 target to combat desertification, restore degraded land and soil, and achieve a land degradation-neutral world. To achieve this goal, SDG indicator 15.3.1 was established by the United Nations Convention to Combat Desertification (UNCCD) to monitor the "proportion of land that is degraded over total land area" and as a basis to plan actions and investments to reverse land degradation. To enable uniform SDG indicator 15.3.1 assessment, the UNCCD has defined the following three sub-indicators: vegetation productivity, land cover changes, and soil organic carbon changes. They are calculated and combined using the "one-out-all-out" methodology (1OAO) proposed by the Trends.Earth software (Conservation International, 2018) which process raster dataset (e.g. MODIS, AVHRR, CLC) and allows a pixel-based qualitative index of potential degradation. Indeed these indicators are useful tools to have an explicit spatial picture of the soil and land conditions over time. Nevertheless, there is increasing evidence that the SDG indicator 15.3.1 approach has some embedded reliability problems (e.g. resolution, procedures), even raising potential controversies and misleading policy guidance.

10.1.2 Domain Specific Data Resources

The primary satellite data (mainly Sentinel 1/Sentinel 2) come from ESA.

In addition to satellite data, the following datasets are recommended for the implementation of the main Pilot output:

- soil databases (e.g. EUSO);
- EU geospatial elaborations such as those provided by ESDAC (e.g. erosion by JRC);
- national and regional ECZ datasets (e.g. Regione Campania and ISPRA ECZ datasets).

10.1.3 Main Standards

SENTINEL products are made available to users in SENTINEL-SAFE format, including image data in JPEG2000 format, quality indicators (e.g. defective pixels mask), auxiliary data and metadata (https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/data-formats).

For soil and ECZ datasets there could not be a unique standard with which global (FAO, ISRIC), national (ISPRA) and local (Regione Campania) datasets are stored. We will deal with both vector and raster data models having different formats but necessarily including one of the formats handled using GDAL (raster driver documentation at https://gdal.org/drivers/raster/index.html; vector driver documentation at https://gdal.org/drivers/vector/index.html).

10.1.4 Software, Workflows, Containers

The current approach is based on the calculation of sub-indicators for land degradation (productivity, land cover and soil organic carbon) which are typically performed in Trends.Earth software (http://trends.earth/docs/en/).

10.1.5 Types of Applications

Deploy an operational platform with easy-to-use, parametric and possibly real-time tools enabling the analysis of the different SDG 15.3.1 sub-indicators on the base of country needs.

10.1.6 Types of Expected Users

Users can be:

- EU, national and local policy makers, in particular EU member States are obliged to provide their own accountancy of states and dynamics of land degradation;
- any scientist, stakeholder or citizen dealing with soil degradation issues.

10.2 PECULIARITY OF THE PILOT AND AIMS

10.2.1 Core Mission Statement in FAIR-EASE

In FAIR-EASE we plan to address the main reliability problems in order to find solutions and produce an operational system enabling analysis over enhanced versions of the SDG 15.3.1 implementation.

10.2.2 Proposed Applications and Demonstrator/s

EU member States are obliged to provide their own accountancy about the state and the dynamics of land degradation, according to SDG15.3.1. Therefore, the improvement of the methodologies in the aims of this Pilot provides a useful demonstrator for the implementation of SDG 15.3.1.

The pilot will exploit the LandSupport environment for testing and validations of the methodologies. LandSupport (H2020-RUR-2017-2, grant agreement No. 774234, https://www.landsupport.eu/) is a web-based, free and open-access GeoSpatial Decision Support System (S-DSS) devoted to reconciling agriculture, environmental sustainability and policy implementation. Overall, the S-DSS contributed to the development and implementation of land use policies in Europe, and it can promote an integrated and participatory approach towards rural development and environmental policies allowing, among others, evaluation of trade-offs between different land uses.

10.2.3 Aims and Scope of the Applications and Demonstrators

Demonstrate that enhancements of the thematic accuracy of the SDG15.3.1 indicators are feasible while exploiting approaches from technical WPs in FAIR-EASE that could facilitate the production and deployment of land degradation workflows and analysis.

10.2.4 Main Targets

The main goals of the Pilot will be achieved through:

- Evaluation of the usefulness of ECZ data and services in EU infrastructures and networks to both assess status and keep monitoring of land degradation (SDG 15.3). This activity will be performed by testing Sentinel data (e.g. HRL Soil Sealing) with soil databases (e.g. EUSO) and currently available geospatial elaborations provided at EU level for instance by ESDAC (e.g. erosion by JRC). Testing will also be performed over national and regional ECZ dataset (e.g. Regione Campania and ISPRA ECZ dataset);
- Implementation of operational and preoperational web-mapping tools to assess land degradation at different NUTS levels (e.g. NUT2-NUTS3, see https://ec.europa.eu/eurostat/web/nuts/background) in real-time. Tools can deal with erosion, loss of soil biodiversity, salinization, sealing and climate change.

10.2.5 Data Flow Diagram and Description

Data flow diagram of the PILOT 5.1.2: EARTH CRITICAL ZONE is described in Figure 3.

Funded by the European Union

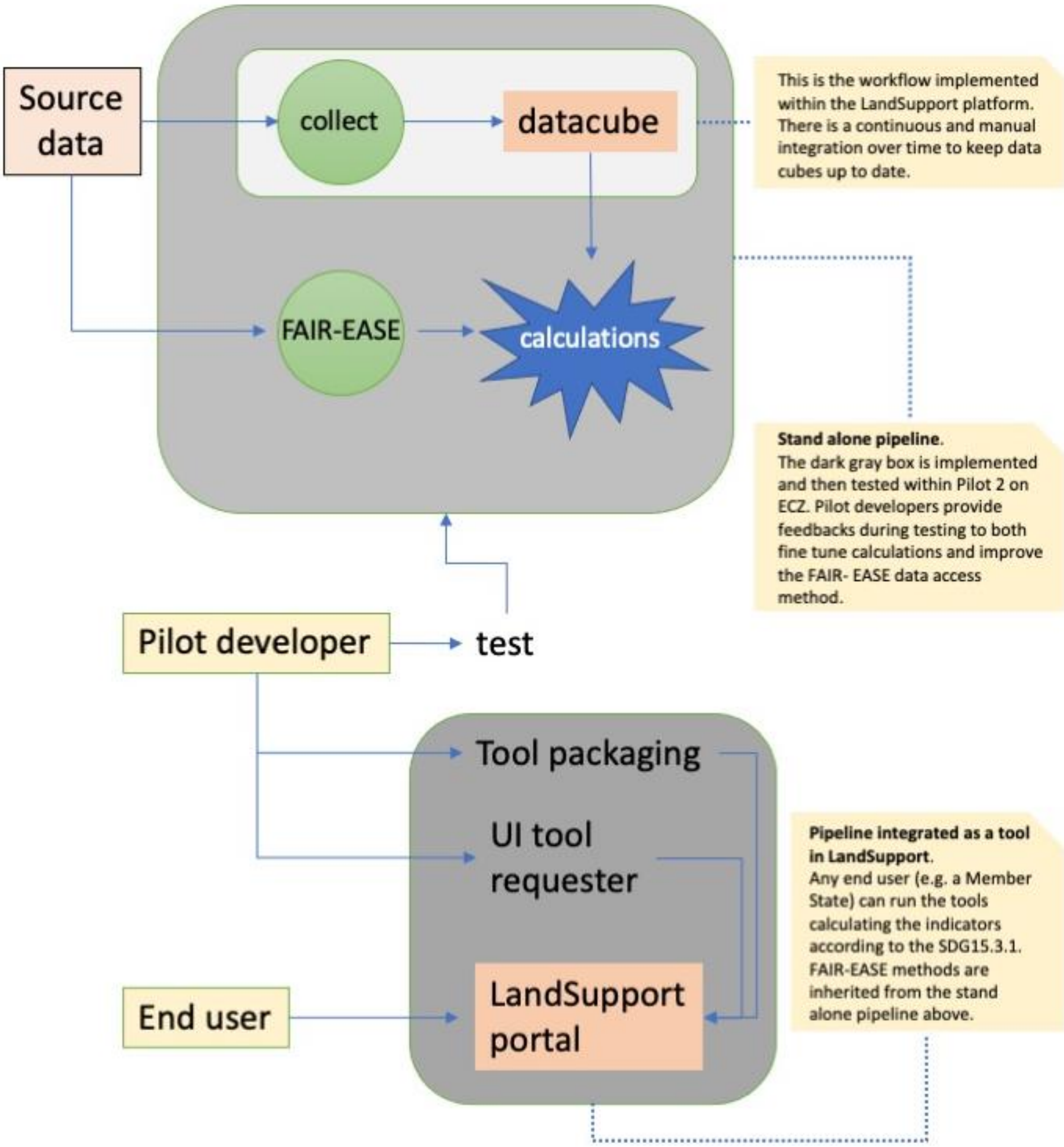

Figure 3 - Diagram of the Land Support portal

10.2.6 Key Resource Data and Datasets

Key data resources that could be useful to fulfill the effort are reported in Annex A. Nevertheless, one of the main activities during the progress of the project will be to assess what else and in which format could be appropriate to define the best procedures to meet the Pilot aims.

10.2.7 Key Metadata Descriptors

Two different retrieval protocols exist: OpenHub and HTTPS API (https://scihub.copernicus.eu/dhus/#/home). The data products are mostly in JP2000 format. A user guide can be found at https://scihub.copernicus.eu/userguide/.

10.2.8 Expected Services/Applications in the FAIR-EASE

The operational tool demonstrator will be included in the LandSupport platform as an end-user, easy-to-use and real-time parametric tool(s) and will be designed as a FAIR methodology to be exploited in the FAIR-EASE Earth analytic lab and by other stakeholders as well.

The new tool(s) will be in charge of demonstrating that the aims and the scope of the Pilot will be fulfilled. Links and connection with the Earth Analytic Lab will be defined during the progress of the project.

10.3 REPORT ON KEY REQUIREMENTS FROM THE PILOT

The key requirements for the specific aim of this Pilot will be defined during the first year implementation of the project. Meanwhile, the sharing of information related to the current technological asset in LandSupport and those from other Pilots with similar workflows and needs will permit the proper validation and plan of improvements in FAIR-EASE.

Methods for data access and subsetting are already implemented within the LandSupport platform using a raster datacube software manager (RASDAMAN, see Annex A) covered by an enterprise license (the license extends the use to 8 CPU cores and it covers updates and configuration support) Currently, the use of a licenced software is a limit in the context of free and open-access software. Possibly, a useful goal in FAIR-EASE would be to develop a technology capable of substituting the licensed raster datacube software manager with a free and open solution.

10.3.1 Summary of Requirements

- Fast access to the datasets, in particular satellite data;
- Fast and on-the-fly raster algebra – possibly executed remotely – while downloading only the final result. This could be achieved by moving the data query near the data location, instead of moving the data near to the calling procedure;
- Possibility to shortcut the request procedure mediated by the Fair-EASE tool/package by means of an URL et similia that can be easily embedded in the "calculation" pipeline;
- Optionally, data and metadata virtualization;
- Optionally, possibility to build data and metadata cache, according to user preference or frequency or elapsed time of execution (i.e. most recent N requests);
- Preference for Python environment, e.g. Jupyter(lab);
- Preference for R/MatLab environments;
- Flexibility in the creation of the working ENV: provide complete documentation about how to embed the Fair-EASE tool/package within R/Python also including the option for the creation of a docker container;

- Provide the minimum and the optimal hardware requirements in order to properly run the Fair-EASE tool/package also in a scalable way.

10.4 POSSIBLE CROSS LINKS WITH OTHER PILOTS

The strategies implemented in this Pilot will represent a reference for additional implementations in the Earth Analytic Lab. In addition, methodologies for efficient data gathering are already in place in the LandSupport Platform and evaluation of possible shared technological advancements will be considered within FAIR-EASE.

11 USE CASE 5.1: Earth and Environmental Dynamics – PILOT 5.1.3: Volcano Space Observatory

11.1 GENERAL DESCRIPTION AND BACKGROUND

The aim of this Pilot in UC5.1 will tackle another relevant Earth system, focusing on the volcano space. The aim is to build a web platform, named“Volcano Space Observatory Portal”, that will be capable of interactively aggregating and jointly analyzing remote sensing observations from Solid Earth and Atmospheric Science communities for the near-real-time monitoring of volcanic activity. Such data will include complementary observations of ground deformation using SAR (synthetic aperture radar) interferometry, topographic changes from optical stereophotogrammetry and 4D physico-chemical properties of volcanic emissions and plumes from a variety of polar orbiting and geostationary satellites as well as ground-based remote sensing observations.

11.1.1 State of the Art in the Specific Domain

Volcanic eruptions pose substantial hazards to communities at local and regional scales, due mainly to explosions, pyroclastic flows, ash falls and mud flows. In addition, large volcanic eruptions can impact climate, air quality and disrupt air traffic. When available, ground instrumentation is essential for volcano observatories to detect in advance any hazardous excursion away from baseline seismic, ground deformation and/or gas emission activities, and warn decision makers of potential threats. Unfortunately, many dangerous volcanoes in the world are poorly monitored, if at all, due to lack of resources and logistic difficulties posed by challenging environmental conditions. Furthermore, ground-based data can be subject to potential failure or destruction in the event of a powerful eruption, as exemplified by the recent Saint Vincent eruption in April 2021 (Boichu and Grandin, 2021). In this situation, volcanologists and atmospheric scientists have to rely on space-based observations to track the evolution of eruptive activity in near real-time (Shreve et al., 2020). A common indicator that is sought for is the volumetric flux of volcanic material as a function of time, because it conditions the load and atmospheric trajectory of volcanic gas and particles, hence potential hazards for air traffic and air quality (Boichu et al., 2015). Lava, ash and gas discharge result from a complex interaction of internal processes (magma/gas ascent rate, magma composition, storage and transport geometry) and superficial processes (volcanic edifice damage, conduit dynamics, interaction with the hydrothermal system), hence representing a key observation to interpret and potentially forecast eruptive dynamics, in conjunction, when available, with ground-based observations.

Currently, no thematic platform allows for seamlessly jointly exploring remote sensing data from the Solid earth and Atmospheric Science communities, although they provide complementary observations to constrain first the eruptive dynamics and subsequently various atmospheric impacts. Existing platforms rely only on satellite observations and are either specifically targeted for the long-range tracking of volcanic plumes (eg. SACS, Volcanic

Cloud Monitoring- NOAA/CIMSS), but generally focus on one indicator (e.g. ash or SO2) lacking local indicators of volcanic activity, or, to the contrary, are focused on the small-scale changes affecting the volcano using infrequent SAR or optical observations, while missing the broader picture and high temporal resolution provided by large-scale atmospheric observations (e.g COMET Volcanic and Magmatic Deformation Portal, MOUNTS platform). Multi-thematic platforms, such as the GeoHazard Exploration Platform, provide a range of observations that are currently too generic to satisfy the needs of the volcano community, and lack the possibility to interact with the products.

11.1.2 Domain Specific Data Resources

The Volcano Space Observatory Portal will benefit from the experience gained in building the VOLCPLUME portal (portal access, portal website), hosted at AERIS/ICARE (Lille), that already allows for exploring a large archive of atmospheric data stored in the same place for the monitoring of volcanic emissions and plumes. Currently, the datasets are physically stored at AERIS/ICARE (Lille).

Ground-based remote sensing or in-situ measurements are also analyzed, in combination with satellite observations. Such data originate from different ground-based networks (AERONET https://aeronet.gsfc.nasa.gov/ ; GEODAIR https://www.geodair.fr ; European Environment Agency https://www.eea.europa.eu; LIDAR networks: https://www.actris.fr/thematiques/gt1-aerosol-remote-sensing/). These data are downloaded through HTTPS API, and are usually csv ASCII files, but unfortunately are not in a normalized format.

Note that all the data needed for the atmospheric domain are already available online at AERIS/ICARE and work has been done during the PHIDIAS European project to bring them into an iRODS system (https://irods.org). We strongly recommend exploring that solution to share data in the so-called data lake.

To handle the data from the Solid Earth community, the Volcano Space Observatory Portal will need to access very heavy/voluminous SAR data (Sentinel-1 data, SLC format). These data are stored in CNES/PEPS (Toulouse). The PEPS facility is a mirror of the European COPERNICUS Sentinel Open Data Hub.

11.1.3 Main Standards

The primary satellite data for the Atmospheric community come from space agencies (ESA, NASA). Different retrieval protocol exist (FTP, OpenDAP, HTTPS APIs for NASA), HTTPS API for ESA/Copernicus https://s5phub.copernicus.eu/dhus/#/home. The data products are mostly in NetCDF or HDF5 formats. When NetCDF is used, the CF convention (http://cfconventions.org/) is widely used for metadata.

The commonly used vocabulary is based on the GCMD (https://gcmd.earthdata.nasa.gov/KeywordViewer). Note that at European level, a common harmonization effort is in progress in the framework of the ACTRIS Research Infrastructure (https://www.actris.eu/). A first version should be released in 2023. It is based on the Skosmos technology.

11.1.4 Software, Workflows, Containers

The interactive analysis and visualization interfaces are based on the Python stack: numpy/scipy/pandas/scikit-image/gdal… These libraries are embedded in the Pangeo (https://pangeo.io/) toolkit that would be a good baseline. The interactive visualization stack is also based on Python libraries like Bokeh/Panel/Holoviews/Datashader, embedded in the holoviz (https://holoviz.org/) stack.

At the moment for the VolcPlume Portal, the processing is done directly on the bare metal server hosting the application, but aiming at a better scalability, the idea is to run them on a computing cluster using Singularity containers. Again, work has been done in this direction during the PHIDIAS (https://www.phidias-hpc.eu/) project to trigger processes in an HPC infrastructure.

11.1.5 Types of Applications

Basically, the purpose of this Pilot is to propose a user-friendly, interactive and efficient way to explore and jointly analyze a broad panel of heterogeneous remote sensing observations, from both the atmosphere and solid earth communities, for the near-real time monitoring of the volcanic activity in case of eruption and the multi-scale impact of volcanic emissions on the atmosphere.

11.1.6 Types of Expected Users

The "Volcano Space Observatory" portal will be of interest to a science-trained audience from Solid Earth and Atmospheric Science scientific communities but more broadly also for volcanic observatories worldwide to help hazard assessment, especially during explosive eruptions that may destroy ground instruments, as recently exemplified by the 2021 eruption of La Soufriere Saint Vincent (Antilles). The knowledge of volcanic activity at a high temporal resolution is also crucial for robustly determining large-scale impacts of volcanoes on atmosphere (air quality, air traffic) and climate. As such, this platform will be also of interest for scientists involved in the field of volcanic impacts at large, including institutions in charge of the monitoring of air quality and aviation safety.

11.2 PECULIARITY OF THE PILOT AND AIMS

11.2.1 Core Mission Statement in FAIR-EASE

In FAIR-EASE we plan to address the main reliability problems in order to find solutions and produce an operational system enabling analysis over enhanced versions of the SDG 15.3.1 implementation.

11.2.2 Proposed Applications and Demonstrators

There are several prerequisites to allow for creating the interactive visualization of the Volcano Space Observatory Portal proposed in this Pilot. The development of this new web service portal will benefit from the experience gained in building the VOLCPLUME portal (portal access, portal website), hosted at AERIS/ICARE (Lille), that already allows for exploring a large archive of atmospheric data stored in the same place. Currently, the datasets are physically stored at AERIS/ICARE (Lille), and the VOLCPLUME web portal also operates from within the computer facilities of ICARE (left part of Fig. 4.1). In the project, additional atmospheric data from European databases, that are not existing in the VOLCPLUME portal, will be added to the Volcano Space Observatory Portal. Specific analytic tools will also be developed allowing on-the-fly interactive processing of remote sensing data with visualization of resulting products in order to monitor the evolution of the volcanic activity.

To handle the data from the Solid Earth community, the Volcano Space Observatory Portal will need to access very heavy/voluminous SAR data from the Solid Earth community (Sentinel-1 data, SLC format). These data are stored in CNES/PEPS (Toulouse) (right of Fig. 4). The PEPS facility is a mirror of the European COPERNICUS Sentinel Open Data Hub. To alleviate the burden involved in downloading / processing these heavy datasets, it would be needed to be able to get access to spatial subsets inside the Sentinel-1 SLC products. This makes sense for volcanic targets because they are small compared to the cov*erage of the images.* It is relevant to underline that the processing of Sentinel-1 InSAR is expected to be "on-demand" when accessing the portal, since a pre-computation of interferograms would represent an excessive burden and delay. Indeed, interferometric processing of these SAR data is, by nature, combinatory. In other words, pairs of images, acquired at successive overpasses, are combined into interferograms. Processing of individual interferograms is computationally intensive, requires rapid disk access (read/write for temporary files), but most steps can be run independently in parallel. The larger the number of available threads/workers, the faster the result is. Access to computational resources will determine the physical location of the processing performed in the backend part of the platform. Currently, only the CNES computation infrastructure allows for proposing, in the same physical site, (a) access to the complete Sentinel-1 archive, via the PEPS infrastructure and (b) massive processing capabilities, via the HAL cluster (CNES's HPC infrastructure). As a result, the FAIR-EASE Volcano Space Observatory will rely strongly on the work done in partnership with CNES.

Architecture of the InSAR processing engine for Sentinel-1 radar imagery

1) Subsetting from the CNES/PEPS database. The very first expectation is to be able to download spatial subsets of Sentinel-1 products, either from COPERNICUS Open Data Hub or from CNES / PEPS. Technically, we need to be allowed to make "range requests" into these databases. This feature is currently possible only from inside CNES HPC facility, where data is stored on disk (/gpfs). For data stored on band, a first request for staging on the disks is required. After a few seconds/minutes, the data is moved to disk, and range requests become possible. These mechanisms will remain in place in the future CNES DataLake (which should open in mid-2023). We foresee the development of a first brick of code that allows for (a) interrogating the PEPS archive, from within CNES, and moving data from band to disk when necessary, (b) applying spatial and temporal subsetting via range requests, and (c) building a cube of Sentinel-1 sub-images that can be used for the InSAR processing.

Interoperability / bidirectional visibility of the CNES data lake with the FAIR-EASE data lake would be highly required.

2) Development of a demonstrator on the CNES JupyterHub "Datalabs / GDH" system.

The second point to address is to have access to a processing facility that offers possibility for :

- deploying a processing software;
- getting access to sufficient disk space (several TB) with rapid read/write capability for temporary and cached files;
- running a large number of workers in parallel (the more the better).

These requirements are all met by the CNES HPC facility "HAL", coupled with the JupyterHub "Datalab" portal, within the framework of the GeoDataHub project. This JupyterHub will serve as a platform for developing a specific tool for fast InSAR processing of Sentinel-1 (to be done). Assistance from FAIR-EASE project experts for development of interactive Python-based code, packaging of codes and deployment on a Jupyter system hosted at CNES would be a useful achievement.

3) Interaction with the user.

Regarding the Solid Earth data, the user will be provided with the possibility to explore the Sentinel-1 archive and define his-her search parameters based on geographic area of interest, time frame and SAR parameters (polarization, sub-swath). This part of the service already exists as an API hosted at IPGP (Institut de Physique du Globe de Paris) through the GDM SAR service. Then, when data is selected, processing will be performed at CNES (see above). Finally, results should be exposed to the user via a download service to be developed (an option could be an iRODS storage solution on the FAIR-EASE datalake). Technical guidelines for the staging of products onto the FAIR-EASE data lake would be appreciated.

4) Interactive analysis of products.

A layer of interactive visualization should also be developed and integrated within the Volcano Space Observatory so as to allow the joint analysis of remote sensing data from both Solid Earth and Atmospheric Science disciplines. To guarantee homogeneity of the tool with the pre-existing "VolcPlume" portal, Python libraries of the Holoviz ecosystem will be used. New tools for extension of the VolcPlume portal are also planned. Mutualisation of efforts within FAIR-EASE in the development of tools relying on Python Holoviz libraries would be a bonus.

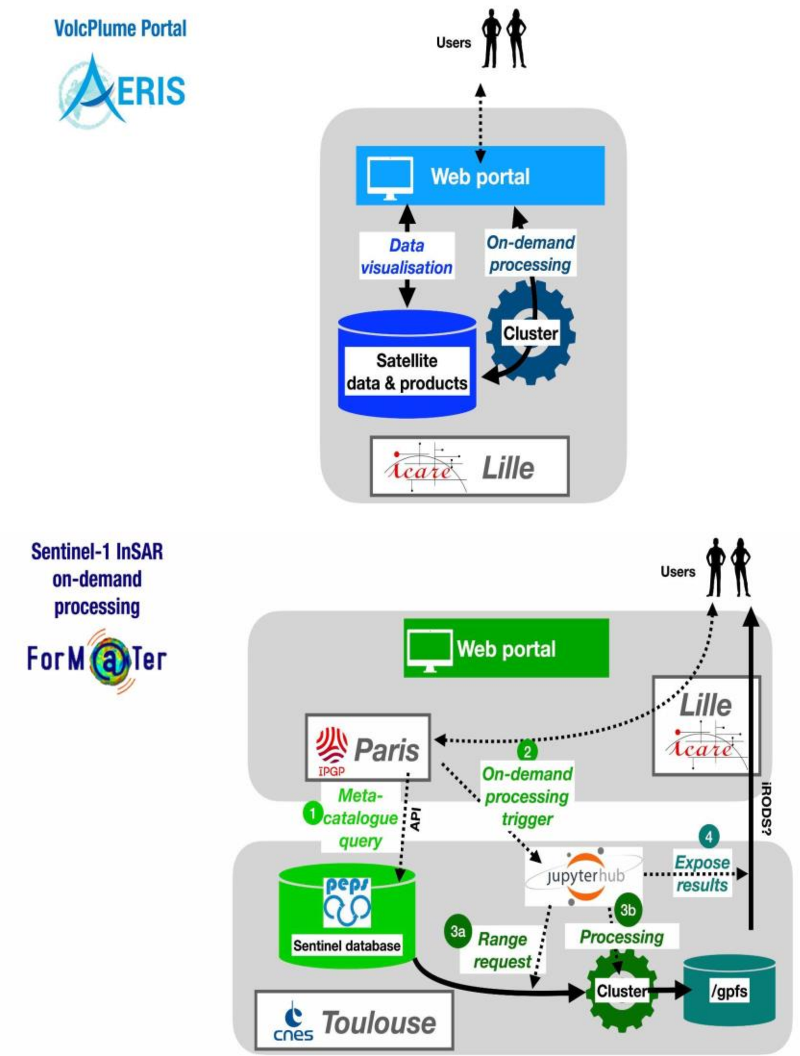

Figure 4 - Architecture of current services in the Pilot 5.1.3

The atmospheric AERIS communities (left) and the Solid Earth/ForM@Ter (right) , both involved in the Volcano Space Observatory Pilot.

11.2.3 Aims and Scope of the Applications and Demonstrators

The purpose of this Pilot is to propose a user-friendly, interactive and efficient way to explore and jointly analyse a broad panel of heterogeneous remote sensing observations, from both the atmosphere and solid earth communities, for the near-real time monitoring of the volcanic activity in case of eruption and the multi-scale impact of volcanic emissions on the atmosphere. As a concrete example, since the beginning of the FAIR-EASE project, the joint analysis of remote sensing data has already been conducted manually by the members of this Pilot for providing assistance to the Volcanological Observatory of Piton de La Fournaise (OVPF) of IPGP (Institut de Physique du Globe de Paris) during the eruption of Piton de La Fournaise volcano in September-October 2022. Interactions with the association in charge of air quality monitoring at La Reunion were also developed as several events of air pollution in sulfur dioxide of volcanic origin were highlighted in the southern part of the island during this eruption by members of this Pilot (https://www.data-terra.org/en/news/fairease-observe-volcanoes/; , https://www.aeris-data.fr/leruption-du-piton-de-la-fournaise-observee-par-la-plateforme-web-volcplume/)

11.2.4 Main Targets

The aim of this Pilot is to propose the joint analysis of heterogeneous remote sensing observations for the monitoring of global volcanic activity, allowing the focus on any major volcanic eruption worldwide.

11.2.5 Data Flow Diagram and Description

Data flow diagram of the PILOT 5.1.3: VOLCANO SPACE OBSERVATORY (Fig. 5).

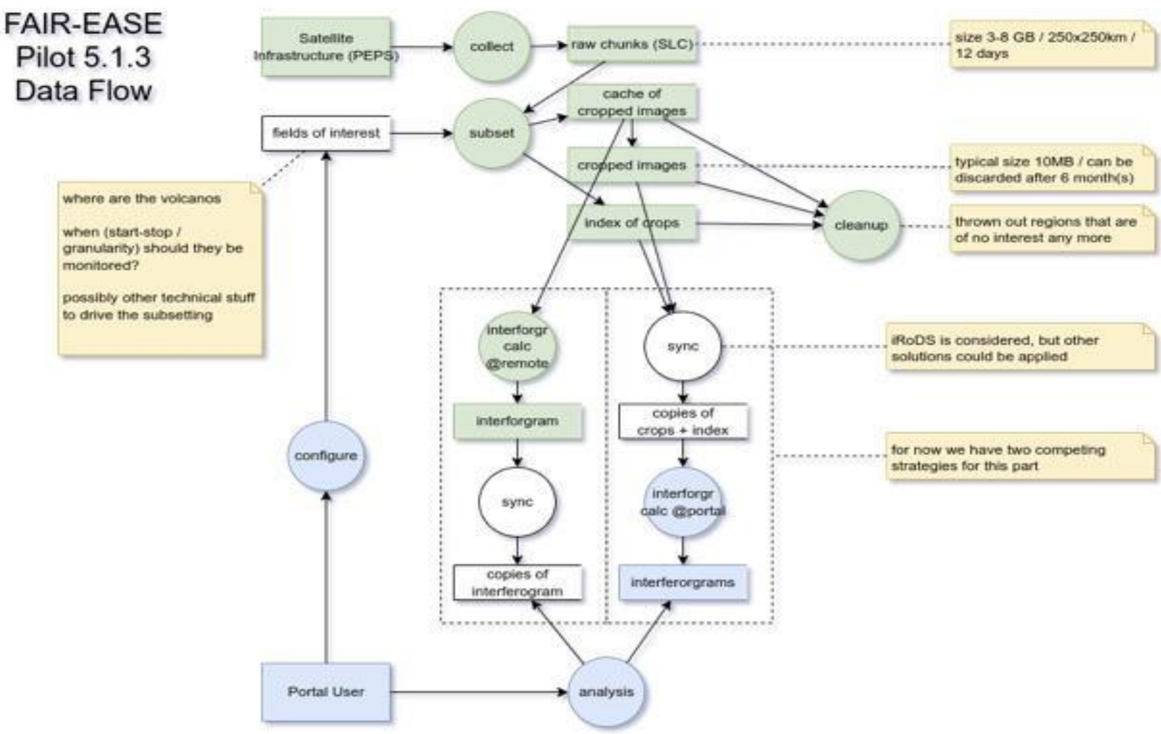

Figure 5 - Expected workflow in the Pilot 5.1.3

This diagram illustrates the challenge that will need to be overcome by the Volcano Space Observatory use case to propose to a user the efficient processing of InSAR data on small volcanic domains by applying spatial and temporal subsettings of heavy Sentinel-1 data via range requests. Such developments will allow a fast data processing in the back end and subsequently a fast visualization of observations on the portal so as to support a rapid analysis of the volcanic activity during a volcanic crisis. Note that the joint analysis of remote sensing data from both Solid Earth and Atmospheric Science disciplines is not shown in this diagram.

11.2.6 Key Resource Data and Datasets

See Annex A for details.

11.2.7 Key Metadata Descriptors

See Annex A for details.

11.2.8 Expected Services/Applications in the FAIR-EASE

We aim to provide access to the Volcano Space Observatory portal in the timeframe of the FAIR EASE project.

11.3 REPORT ON KEY REQUIREMENTS FROM THE PILOT

Summarized above.

11.3.1 Summary of Requirements

Summarized above.

11.4 POSSIBLE CROSS LINKS WITH OTHER PILOTS

We may envisage some cross links with the PILOT 5.1.1 Coastal Dynamics, to study for example variations in ocean color associated with phytoplankton blooms that may be produced by major volcanic eruptions.

12 USE CASE 5.2: The Environmental BioGeochemical Asset – PILOT 5.2.1 Ocean Bio-Geochemical Observations

12.1 GENERAL DESCRIPTION AND BACKGROUND

UC5.2, the Environmental Biogeochemical Asset, considers a crucial Real Life-Science challenge of interest for different environmental systems. It involves multifaceted aspects that are typical and peculiar of the domain under consideration, in relationships to the Earth system under analysis. In each domain, it requires dedicated expertises for the establishment of standards and the addressing of methodological issues according to the employed technologies and to their evolving. Specifically, the Pilot designed in FAIR-EASE tackles one of the relevant Earth systems to be considered to investigate bio-geochemical assets. The focus is on marine environments, for which the knowledge of the bio-geochemical asset is fundamental to understand processes involving marine ecosystems, their dynamics and the impact on the health status of this precious Earth environment.

12.1.1 State of the Art in the Specific Domain

The observation of marine biogeochemical processes (BGC) is useful to address fundamental scientific questions regarding the health of marine ecosystems (e.g. ocean acidification, oxygen minimum zone, biological carbon pump, phytoplankton communities...) and needs for ocean resource management (e.g. Johnson et al., 2009; Chai et al., 2020). Today, BGC sensors are deployed in the global ocean through various platforms (floats, gliders, sea mammals, moorings, etc.) by observing networks under GOOS-OCG international coordination (Global Ocean Observing System - Observations Coordination Group). We count more than 910 836 BGC profiles measured either by a BGC-Argo float, a glider or a sea-mammal throughout the World Ocean, including the European marginal seas (source: The Global Ocean In-situ near real time data set from Copernicus Marine Service - https://doi.org/10.48670/moi-00036, Dec. 2nd, 2022): 18% are already in delayed mode status (usable by scientists), 10% are adjusted automatically in real time whereas the remainder is being qualified as bad data that are potentially correctable. Biogeochemical data adjustment and validation represent thus a major challenge to significantly increase the volume of high quality BGC data made available to the scientific community.

12.1.2 Domain Specific Data Resources

The marine BGC observations are qualified, calibrated or validated from independent data collections, that are additional in situ observations, aggregated or gridded products, model outputs or satellite products of sea surface ocean. Moreover methodologies for the sensor adjustment sometimes require atmospheric information available in the near real time or reanalysis fields. Other methods need bathymetric fields to position autonomous platforms drifting under ice as well.

Funded by the European Union

12.1.3 Main Standards

Most of the data are available in netCDF (CF conventions).

12.1.4 Software, Workflows, Containers

Today, the BGC-Argo science team is the major contributor on an international initiative to calibrate, validate and trigger alerts on Ocean BGC in situ data. In recent years, BGC-Argo sensors have diversified (oxygen, nitrate, chlorophyll, suspended particles, pH) and methods of data Quality Assessment and Control, validation and adjustment have become more complex (Bellacicco et al., 2019; Estelmann et al., 2021; Jutard et al., 2021(a,b); Maurer et al., 2021, Schmechtig et al., 2021). Most of the methods are used in open source tools, available on various public github/gitlab and detailed in Annex A. These methods require an efficient access to external data: gridded climatologies, model outputs (meteorologic, oceanographic), discrete in-situ data, satellite data… applying the methods requires to combine/assemble the data in space (profiles, maps) and in time (month, year, time series) using extraction and colocation. To continue their development, this data scientist community aspires to a massive, high-performance, distributed data infrastructure to combine in situ, satellite and models data.

Today, software development and metadata standards are specific to the argo format and community. Methods are essentially sensor-dependent and not platform-dependent, that means applicable to the BGC sensor deployed on Glider or sea-mammals or other platforms including such sensor. Software standardizations and an easy cloud deployment would be welcome for data scientists of these communities.

12.1.5 Types of Applications

The aim of this Pilot is to provide a common platform to data scientists in order to qualify / calibrate / validate (BGC) data from sensors deployed on various platforms.

12.1.6 Types of Expected Users

The primary expected users are the data scientists of the BGC-Argo community, the BGC glider community and the sea-mammals community.

12.2 PECULIARITY OF THE PILOT AND AIMS

12.2.1 Core Mission Statement in FAIR-EASE

The aim of this Pilot in FAIR-EASE is to improve tools in support of data qualification/ /validation by expert teams, facilitating and harmonizing the data access and methodologies.

12.2.2 Proposed Applications and Demonstrators

The Pilot provides a common web platform to data scientists in order to qualify / calibrate / validate data from sensors deployed on various platforms :

- Qualification : Change or add data quality (visual treatment of flag);

- Calibration : Apply standard methodology to calibrate data (profiles and time series) measured by one sensor;
- Validation : Compare data (profiles or time series) with reference datasets;
- File export.

Services will be based on existing tools or code listed in the Annex A - Software/pipelines as detailed in section 5.2.8. This demonstrator will be designed to work on the original Argo/Glider or Sea mammals data files in order to update them with results from the qualification/calibration/validation before sending to the original source. This demonstrator will be designed to easily add specific standardized methodologies for future additional variables measured by a specific platform as well.

12.2.3 Aims and Scope of the Applications and Demonstrators

The aim of the demonstrator is to provide a single and efficient access to the three services (qualification/calibration/validation) through a web portal. Tools deployed for the calibration will be focused on ocean biogeochemical observations measured essentially by sensors deployed on Argo floats, Glider or sea mammals.

12.2.4 Main Targets

Main expected users will be experts or interested scientific users either for their own study or to adjust sensors before returning into the original data flow.

12.2.5 Data Flow Diagram and Description

The following fig. 6 shows the data flow diagram of the demonstrator. A user authentication to a web portal will get access to a query catalog to facilitate the navigation through the demonstrator services (gray boxes). Results of the queries are displayed in blue whereas data set format inside the demonstrator are colored in orange and green following the action required by the user (purple).

Data sources that feed the data-lake to adjust or validate by comparison the argo/glider BGC measurement (gridded climatologies, model outputs (meteorologic, oceanographic), discrete in-situ data, satellite data …) are listed in the annex A. A direct access to external sources will be possible through the web portal.

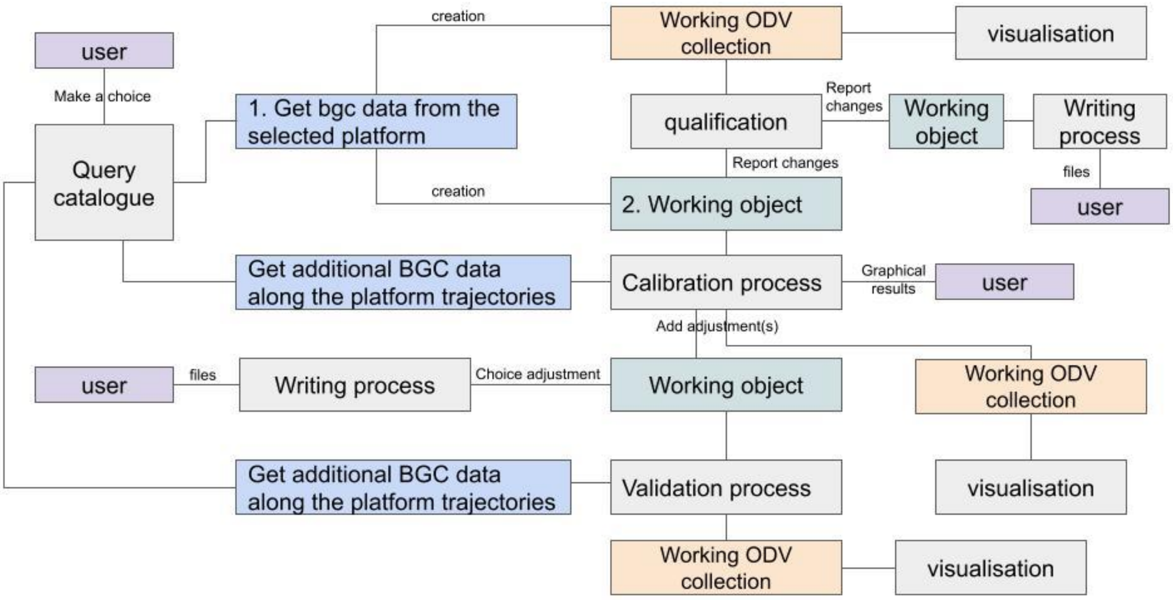

Figure 6 - Data flow diagram in the Pilot 5.2.1

Figure 6 shows an example of a set of consecutive actions that can be performed by a user working on a BGC set measured by a specific platform with three steps (qualification, calibration, validation) from the command panel :

Step 0: Initialization : pick the BGC set from your specific platform

– it creates your working object;

– it creates your working ODV collection 1.

Step 1 : Qualify the data

– visualize the data with web ODV;

– Quality Control your data (marking outliers as bad);

– report the QC changed in your working object.

Step 2. Calibrate the data

– pick one or several methods : e.g. method (a) and method (b) with the command panel;

– Apply the different methods in your working object;

– create a working ODV collection 2 for data adjusted with method (a) and with method(b).

Step 3. Validate the data

– Merge your working ODV collection 2 with reference data sets into a working ODV collection 3;

– Get statistics on your working object by comparison with reference data with the command panel;

– Visualize your results with your working ODV collection 3.

Step 4. Get outputs

– Pick the method that gives you the “best” results;

– Pick a format for the outputs (FAIR Format, original format).

All results and history of the actions are written in output files, either for the own users needs or if the user is an identified Delayed Mode operator of a specific datasets, he can push the outputs to the official workflow of the dataset. The demonstrator will not be allowed to directly feed the data-lake (see fig. 7) with the newly adjusted data.

12.2.6 Key Resource Data and Datasets

Key resources data and dataset are listed on the following table and on the left panel of the figure 5.2 in section 5.2.8 and in Annex A. As mentioned previously, the demonstrator will calibrate BGC sensors of in situ network/source Argo, Glider and sea-mammals. Aggregated in situ data set or products, model outputs, reanalysis and satellite products will be either used by method of calibration or for the data validation. Details regarding each data set (name, utility …) are listed in the Annex A.

12.2.7 Key Metadata Descriptors

Key metadata descriptors of each useful data set are detailed in Annex A. Nevertheless, most of the Key oceanic metadata follows the CF and Argo Common Vocabulary or SeaDataNet Common vocabulary available on the nerc web site (https://vocab.nerc.ac.uk/collection/). Atmospheric fields coming from the Copernicus Climate store follow the ECMWF convention (https://apps.ecmwf.int/codes/grib/param-db). Other products have their own vocabulary.

Funded by the European Union

12.2.8 Expected Services/Applications in FAIR-EASE

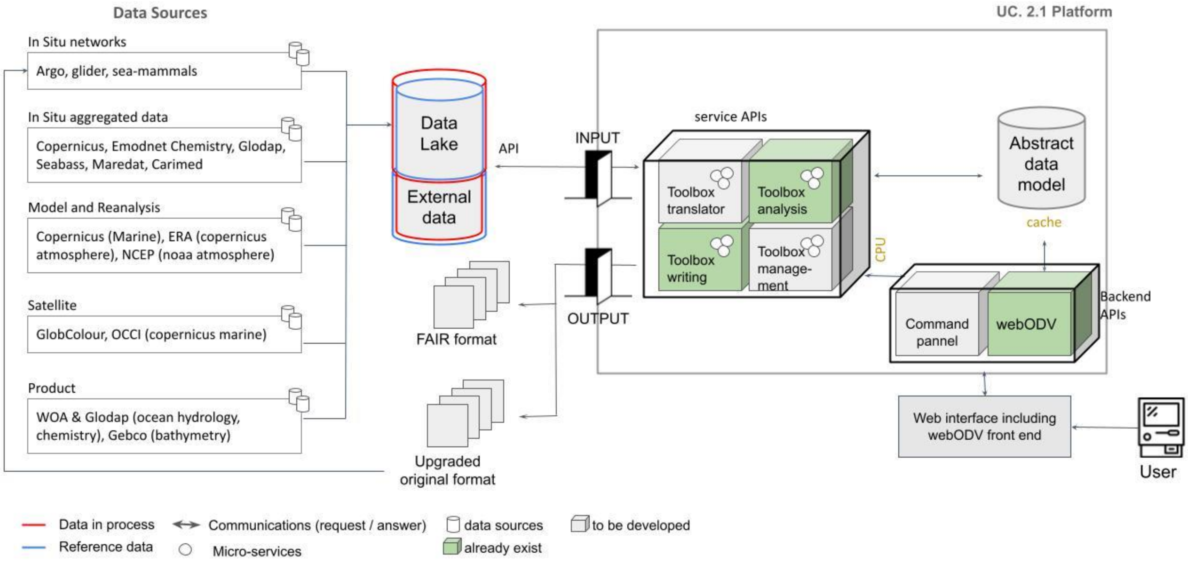

Figure 7 - Expected Workflow in the Pilot

The demonstrator will be composed by :

– Abstract data model

The abstract data model of the proposed demonstrator is like a storage for the toolboxes computations. It should store intermediate analyses, and consist of one or many database engines. These storage engines can be self deployed inside Docker or Singularity containers clusters or ideally be part of a higher service level of the platform resource provider. Both non-persistent and persistent volumes are required. A minimum amount of persistent volumes could be needed on these clusters for medium and long term intermediate results caching.

– 4 specific toolboxes called “translator”, “management”, “analysis” and “writing”.

Toolbox “translator” will fit the data from data-lake format to our platform internal data model. Cerbere, a tool to manipulate any type of spatio-temporal observations (https://cerbere.readthedocs.io/projects/cerbere/en/latest/working_with_data.html) can be a component of this toolbox. Toolbox “management” will organize predefined workflow, potentially through Galaxy software, according to user request. Toolbox “analysis” will include tools for calibration and validation. Most of the calibration methods listed in Annex A are mostly in open sources and already available on various github. Some of them need to be developed during the three years of the project. Toolbox “writing”, that already exists, will write results of qualification or calibration into the original netCDF files or in the fair format.

Depending on the execution pipelines, the application routines will use the abstract data model as their source and thus will share the same inner logic objects definitions. Toolboxes will be composed of heterogeneous micro-services encapsulating routines (R, Python, matlab, Octave). Currently, these micro services are independent scripts. The three years of the project will be used to evolve them into microservices communicating through APIs.

The system needs a security schema. If not zero trust architecture, a minimal segmentation should be fine : management + private + public zone (DMZ), access logs, etc. See web portal section.

– Command panel backend

The demonstrator will need a dedicated control panel to manage all tasks that the visualizer cannot achieve. Each end user (via web frontends) has instantaneous feedback via this command panel backend.

– webODV backend

The demonstrator will use webODV software (Schlitzer and Mieruch-Schnülle, 2023) to perform data qualification, visualization and validation. See section "possible cross links with other Pilots" into "2 USE CASE 5.1: Pilot 5.1.1 COASTAL DYNAMICS" for more details.

– Web portal including the webODV frontend

Users will have access to the demonstrator through a web GUI (graphical user interface) including the frontend of webODV (Schlitzer and Mieruch-Schnülle, 2023). The frontend of the "Command panel" will facilitate the choice between various actions (calibration, visualization, qualification, location, comparison, exportation) and between various data sets including in the data lake. This GUI will be MVC compliant (model view controller), which also means that a javascript frontend publication on a web https server communicates with an API backend. No direct access from the GUI interface to the Data Model will be granted. Load balancing and authentication layer should be applied on the hosting infrastructure's reverse proxy (…let's say featured firewall protecting the API servers).

12.3 REPORT ON KEY REQUIREMENTS FROM THE PILOT

Fast communication between the data-lake and the UC2.1 platform

The proposed implementation here expects:

- a communication like request/answer through an API with a common vocabulary.
- JSON objects are preferred for immediate integration into the platform data model (input workflow). Units conversion options and collections of all the associated metadata from the original source in the response object are useful. Requests should be either a complete data set or a subsetting data collection (x,y,z,t). For comparison or visualization that implies little queries, very quick answers are expected (less than 5 sec queries). All requests should reply with a response transfer estimation included in header or in 2 steps (pre-request, then request).

- Otherwise, a documented http websocket binary downloader will be needed at the time of building a delivery (whole RAW NetCDF Archive with injected modifications : output workflow). Please avoid, even ban aside FTP/SFTP transfers that need other services presentation and security management.
- In general, for long requests or encapsulated file transfers, a transfer or execution progress feedback, inside a download manager or via API queries on a user requests registry service (a data lake API service that monitors requests and allows you to check/cancel)

Unaltered data

The goal is to :

- Export qualified/calibrated/validated data fields (that can be defined like a FAIR-EASE format) containing all results (calibration result, methodology used, QC changed …).
- Add correction(s) and related information to the original file/entry file.

The data lake needs to preserve all the metadata from the data sources in order to rebuild (if necessary) the original netCDF files including adjustments and possible corrections (e.g. original argo netcdf file ⇒ original argo netcdf file + adjusted values and equation_calib informations = argo file output).

Storage capacity and virtual environment

We expect development/staging/production environments with possibly a continuous integration deployment. This would include interactive development IDEs like jupyter Notebooks and also fast data lake requests. For development / testing activity, data lake could produce on the fly parametric dummy content (containing on purpose 'out of range' data errors).

The storage engines like a database can be self deployed inside docker or singularity containers clusters or ideally be part of a higher (and clustered) service level of the platform resource provider.

We also expect extra volume to be used as a cache inside our platform. Cache is needed to store answers from queries addressed to the Datalake and intermediate computations. Computations and data are usually small sized . Nevertheless, it could sometimes require larger cpu to compute specific methods or larger space to store temporary computation results in order to continue the work before exportation.

User Authentication

We expect users or automations service accounts identification (AAI; oauth2 Single Sign On) with user ORCID. The authentication layer is giving and also limiting access via group management.

Two user modes for example : one expert mode and one "non-expert" mode. We expect a "checkbox" to choose:

- variable to be calibrated;
- method(s) to be used according to sensor or available observation (or to be compared in the "expert mode");
- Which results we want to apply.

The expert gives access to all methods and micro services and the non-expert mode gives access to limited methods.

Moreover, the user authentication will be useful to trace action and report information in the netCDF file before returning to their own data flow.

12.3.1 Summary of Requirements

Expectation 1 : A fast communication between the data-lake and the UC2.1 platform;

Expectation 2 : unaltered data;

Expectation 3 : Storage capacity and virtual environment;

Expectation 4 : User Authentication.

12.4 POSSIBLE CROSS LINKS WITH OTHER PILOTS

Reported in Pilot 5.1.1 COASTAL DYNAMICS.

Funded by the European Union

13 USE CASE 5.3: Biodiversity Observation – PILOT 5.3.1 Marine Omics Observation

13.1 GENERAL DESCRIPTION AND BACKGROUND

Biological observations are typically scattered across various initiatives worldwide. UC5.3 will focus on marine biodiversity, where multifaceted aspects and different levels of information can be taken into consideration although dealing with the same Earth system.

The Pilot designed in FAIR-EASE will focus on ocean observations. Specifically, the field of marine biodiversity observation lacks a structured and standardized approach for data and information sharing, in addition, the data depends on the specific level of information considered. The Pilot will focus on omics data.

13.1.1 State of the Art in the Specific Domain

Omics data are considered big data with an enormous potential, but, being the result of heterogeneous strategies for their collection, they need dedicated approaches for proper exploitation and knowledge extraction. The European Marine Omics Biodiversity Observation Network (EMO BON; https://www.embrc.eu/emo-bon) is a coordinated marine genomic biodiversity observation network that was created to reflect the need for a standardized approach to marine genomic long-term observations in Europe. EMO BON generates high-quality genomic biodiversity data that are made openly accessible and thereby available to support decision- and policy- making towards a holistic understanding and a sustainable management of the ocean.

The proposed Pilot aims to exploit the metagenomic data produced by EMO BON. After targeted analytical workflows and statistical analyses, the aim is to produce data that will allow 1) the exploration of biodiversity in space and time in European coastal habitats, and 2) the creation of targeted services to tackle major environmental issues related to the marine microbiome. All intermediate data products and analyses results will be made available through dedicated librairies/applications that will be developed and deployed in the framework of the Earth Analytic Laboratory.

Genome-based methods are an effective alternative to the traditional morphological species identification for biodiversity observation (Canonico et al., 2019; Cordier et al., 2020; Makiola et al., 2020). Genomic approaches can provide substantial amounts of relatively standardized and globally comparable data as they ‘read’ the universal code of life, i.e. the DNA sequences. Metagenomics is a set of methods that sequence the DNA present in a sample at random, thereby obtaining sequence data representing all, or most likely nearly all, of the organismal genomes present. Hence metagenomics can provide data not only on the biodiversity present in the environment, but also on the functional capacities of the species communities through gene function identification. Moreover, these methods allow the identification of species that remain unculturable under laboratory conditions, and therefore provide novel insights into elusive microbial community (Faust et al., 2015; Cordier et al., 2020; Abreu et al., 2022).

Genomics have been used extensively for the exploration of the marine microbiome and remarkable, previously unexplored, data have been collected by trans-Atlantic and trans-Pacific campaigns (de Vargas et al., 2015; Milici et al., 2016), circumnavigating expeditions in the Antarctic (Flaviani et al., 2018), and temporal studies on a single coastal area (Martin-Platero et al., 2018). Such works have expanded our understanding of the marine microbiome and provided a better overview of global biodiversity. Marine genomic data are now available in databases and the emergent results are analyzed and critically discussed in scientific publications.

The coupling of standardized biodiversity observation with genomic approaches generates the potential to use genomic data beyond strictly scientific discussions and towards applications, environmental management, and policymaking (Cordier et al., 2020; Rodríguez-Ezpeleta et al., 2021). However, the transformation of raw genomic data into meaningful information that can feed further scientific activities is not straightforward and requires experienced bioinformatics and statistical analyses (Mendoza et al., 2014). This work requires a structured, FAIRified, and openly available analytical workflow for genomic data, along with the ability to correlate the results with additional environmental data (e.g. the Blue Cloud Plankton Genomics VRE - https://blue-cloud.org/vlabs/plankton-genomics; Irisson et al., 2022)

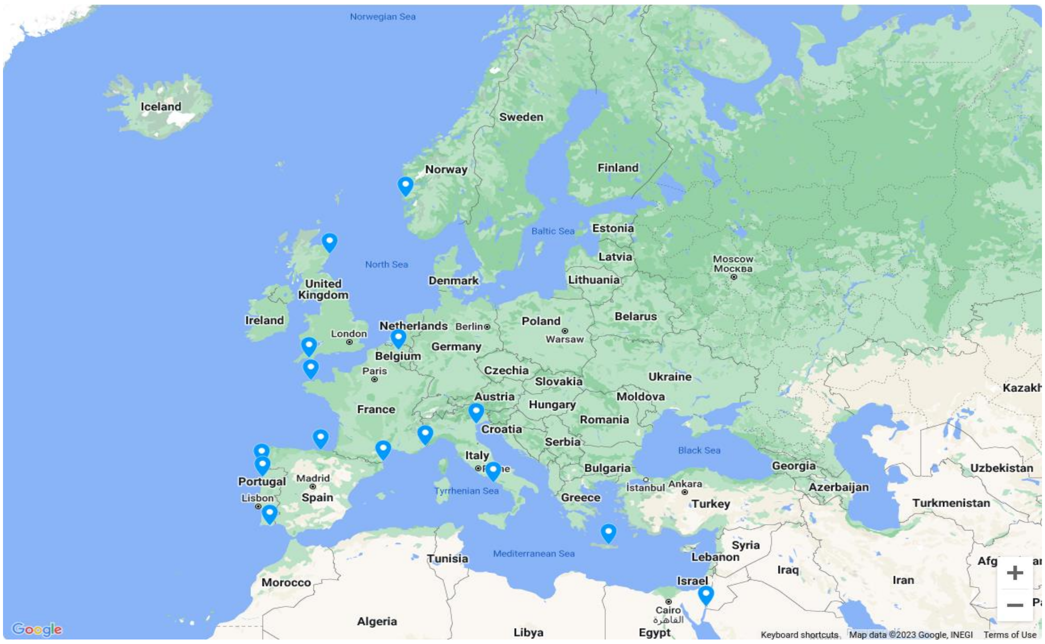

Figure 8 - EMBRC EMO BON metagenomic sampling sites

EMBRC's EMO BON is the first genomics observatory (GO) to be established for coastal European waters (Fig. 8). The data collected are compatible and comparable with genomic data produced from scientific studies across the globe. EMO BON collects samples from the water column, soft substrates, and hard substrates, aiming to allow researchers to deeply explore marine diversity in different habitats. It is based around a network of 16 sites from Northern Norway to the tropical Red Sea in Israel.

The Pilot needs the implementation of a VRE as the first point of access to the EMO BON project resources and metagenomic data for researchers. It will provide links to data in public repositories (e.g. EurOBIS, GBif) that overlap in time/location with EMO BON datasets, and will provide basic statistics and standard visualizations that will allow a comprehensive exploration of the data. In addition, the VRE will provide several services demonstrating the application of the data for molecular bioprospecting and the assessment of microbial spatial heterogeneity in connection to population ecology. The VRE will provide access to tools and analytical frameworks which will enable flexible exploitation of the EMO BON data for a variety of research questions related to the health and quality of coastal marine environments. Ultimately, the aim is for the Earth Analytic Laboratory to be integrated into the EOSC Blue Cloud and EOSC Service Catalog.

13.1.2 Domain Specific Data Resources

The metagenomic research domain has several well established sequence data resources, from a European perspective, the most important being the EMBL-EBI MGnify (https://www.ebi.ac.uk/metagenomics). Other sequence data are held in International Nucleotide Sequence Database Collaboration (INSDC; https://www.insdc.org) repositories in various formats, but not necessarily domain specific collections. Domain specific biodiversity data can be divided into taxonomic identification resources such as the World Register of Marine Species (WoRMs; https://www.marinespecies.org), and geographical distribution data held by repositories such as the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/).

13.1.3 Main Standards

The main standards for data and metadata in the field of metagenomics are being used by the Pilot, these are as follows:

- MIxS: Minimum Information about any (x) Sequence (https://genomicsstandardsconsortium.github.io/mixs/);
- EML: Ecological Metadata Language (https://eml.ecoinformatics.org/);
- Dublin Core DCMI Metadata Terms (https://eml.ecoinformatics.org/);
- Darwin Core RDF Vocabulary (http://rs.tdwg.org/dwc/terms/);
- Research Object Crates (RO-Crates) (https://www.researchobject.org/ro-crate/);
- Schema.org Vocabulary (https://schema.org/);
- JSON Linked Data Format Standard (JSON-LD) (https://json-ld.org/);

- Resource Description Framework (RDF) (https://www.w3.org/RDF/);
- MIBiG: Minimum Information about a Biosynthetic Gene cluster (https://mibig.secondarymetabolites.org/);
- BIOM: Biological Observation Matrix format (http://biom-format.org/);
- FASTA/FASTQ Sequence data exchange format (https://www.ncbi.nlm.nih.gov/genbank/fastaformat/);
- EMBL-EBI ENA API (https://www.ebi.ac.uk/ena/portal/api/);
- EMBL-EBI BioSamples API (https://www.ebi.ac.uk/biosamples/docs/references/api);
- World Register of Marine Species taxonomy (https://www.marinespecies.org/) (REST and SOAP services);
- Marine Regions gazetteer (https://www.marineregions.org/about.php) (REST services, LDE feed);
- NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi).

13.1.4 Software, Workflows, Containers

The VRE to be developed by this Pilot study ingests RO-Crate (https://www.researchobject.org/ro-crate; Soiland-Reyes et al., 2022) formatted data products as its primary data sources. RO-Crate specifications can be found in the EMO BON repository of Github (https://github.com/emo-bon). Each RO-Crate will contain the data products, data, and metadata links to the metagenomics analysis of one EMO BON sequence sample analyzed by the MetaGOflow workflow (https://github.com/emo-bon/MetaGOflow). One EMO BON sample represents a single metagenomic sample of either the water column or sediment taken at a particular site during a particular sampling campaign. There are currently 16 sites in the EMO BON project, and sampling campaigns will occur bi-monthly. Each sample is DNA sequenced using a shotgun methodology and the sequence reads are analyzed via the MetaGOflow workflow. The MetaGOflow workflow is a modified version of the EMBL-EBI MGnify pipeline (version 5; https://github.com/EBI-Metagenomics/pipeline-v5). The data products resulting from the workflow that are to be analyzed in the VRE consist of taxonomic inventories and gene function profiles (in ASCII CSV text format).

The VRE itself is expected to be written for compatibility with the Blue Cloud platform which is based on the Data4Science Framework (https://www.d4science.org/), and will include a Data Discovery and Access Service, visualization tools (written in R, e.g. BiomeShiny: https://bio.tools/BiomeShiny), and “glue” scripting in Python and/or Java.

The implementation of the VRE will include two main demonstrator services:

1) the Bioprospecting Service will utilize the data products resulting from the Systems analyses of the Assembly sub-workflow of MetaGOflow, and additional results of analyses performed using DeepBCG (https://github.com/Merck/deepbgc) software in the VRE;

2) the Ecological Strategies Service will identify the differentiation and composition of different ecological communities in samples and relate those to environmental factors. The software to perform the analysis for the latter service will be custom-built by the Pilot using Python/R libraries, in particular, the R packages in BioConductor (https://www.bioconductor.org/) and the ecological analysis package Vegan (https://cran.r-project.org/web/packages/vegan/index.html).

13.1.5 Types of Applications

TThe AAI and the Data Access and Discovery Service implemented by the Blue Cloud infrastructure will be included. The VRE will be built around a Content Management Framework (including its web-injection language) with real-time backend analysis of data needed to provide the VRE Demonstrator Services.

All frameworks and software employed by the Pilot are based on Open Source libraries and most are on Github (listed in methods in Annex A, table 2). Nevertheless, a number of custom made applications/software libraries will be built in Python/R to implement VRE Service 2: Ecological Strategies.

13.1.6 Types of Expected Users

The Data Discovery and Access service is aimed at providing data access to all potential researchers, including, Marine Biologists, Marine Ecologists, Molecular Ecologists, Microbiologists, Operational monitoring programmes, Modelers, Marine Monitoring, and Observation Stakeholders. In addition the Bioprospecting Service (Demonstrator service 1) will engage Molecular Biologists, Microbiologists, Biotechnology scientists, and Biotechnology companies/businesses, and Policymakers, while the Ecological Strategies service (Demonstrator service 2) will be aimed at providing resources to Marine Ecologists, Molecular Ecologists, Microbiologists, Marine Biologists, and Modelers.

13.2 PECULIARITY OF THE PILOT AND AIMS

The EMO BON project is the first European-wide coastal microbial biodiversity monitoring programme. This Pilot aims to provide access to comparable data products and a VRE platform for exploration of the EMO BON data.

13.2.1 Core Mission Statement in FAIR-EASE

European marine microbial biodiversity is a poorly understood and under-exploited component of continental biodiversity. This Pilot aims to provide a platform for the analysis of spatial- and time-comparable marine microbial metagenomics data sets for the exploration of biodiversity and its correlations with environmental quality using data from the EMBRC European Marine Omics Biodiversity Observation Network (EMO BON) project.

13.2.2 Proposed Applications and Demonstrators

The Marine Omic Observation Pilot will provide a flexible VRE for exploration of microbial marine biodiversity and habitat assessment. The main application to be provided by FE is a VRE for data discovery, visualization, and analyses. Two demonstration services will be provided by the VRE: 1) a Bioprospecting Service which provides real-time analyses of biosynthetic gene clusters (BCGs) using a modified version of the Systems workflow of the MGnify resources, and 2) an Ecological Strategies Services that characterizes the ecological components of a sample with respect to ecological strategy and habitat quality.

13.2.3 Aims and Scope of the Applications and Demonstrators

The aim of Pilot 3.1 is to provide infrastructure for access, data discovery, and bespoke services within a custom-made VRE for the EMBRC EMO BON metagenomic data, metadata, and data products and thereby enable novel metagenomic-based analyses of European marine coastal waters. The scope of the Pilot 3.1 is within the context of the EMBRC EMO BON project that provides metagenomic data and data products to the marine research community for analysis of the health and ecological quality of the marine habitats.

13.2.4 Main Targets

European marine molecular biologists and ecologists, specifically, marine biologists and ecologists working in the fields of coastal habitat ecology, restoration, and quality assessment based on molecular data.

13.2.5 Data Flow Diagram and Description

The data flow diagram of the PILOT 5.3.1 (Fig. 9) is here described:

- The FAIR-EASE will provide the “data access” and “analyses” components of the data flow presented in Fig. 10;
- Some central management decisions and reference concerning procedures, RO-crate profiles and actual planning are governing the overall collaboration and data management agreements;
- "Sampling" involves the field work producing physical samples and their direct/raw data representations;
- During the "Sequencing" step the physical samples are being DNA-sequenced, the data products of which to be linked to the data-representations;
- The process of "Analysis" derives information out of the sequencing data that is tied to the samples once more;
- In the diagram the "Live Aggregate" step functions as a placeholder for the actual work that is to a large extent to become the result of the FAIR-EASE work: a postprocessing of all the available data in such a way that it enables data scientists to explore the results as well as feed into classical publication (release) systems (e.g. GBIF and OBIS).

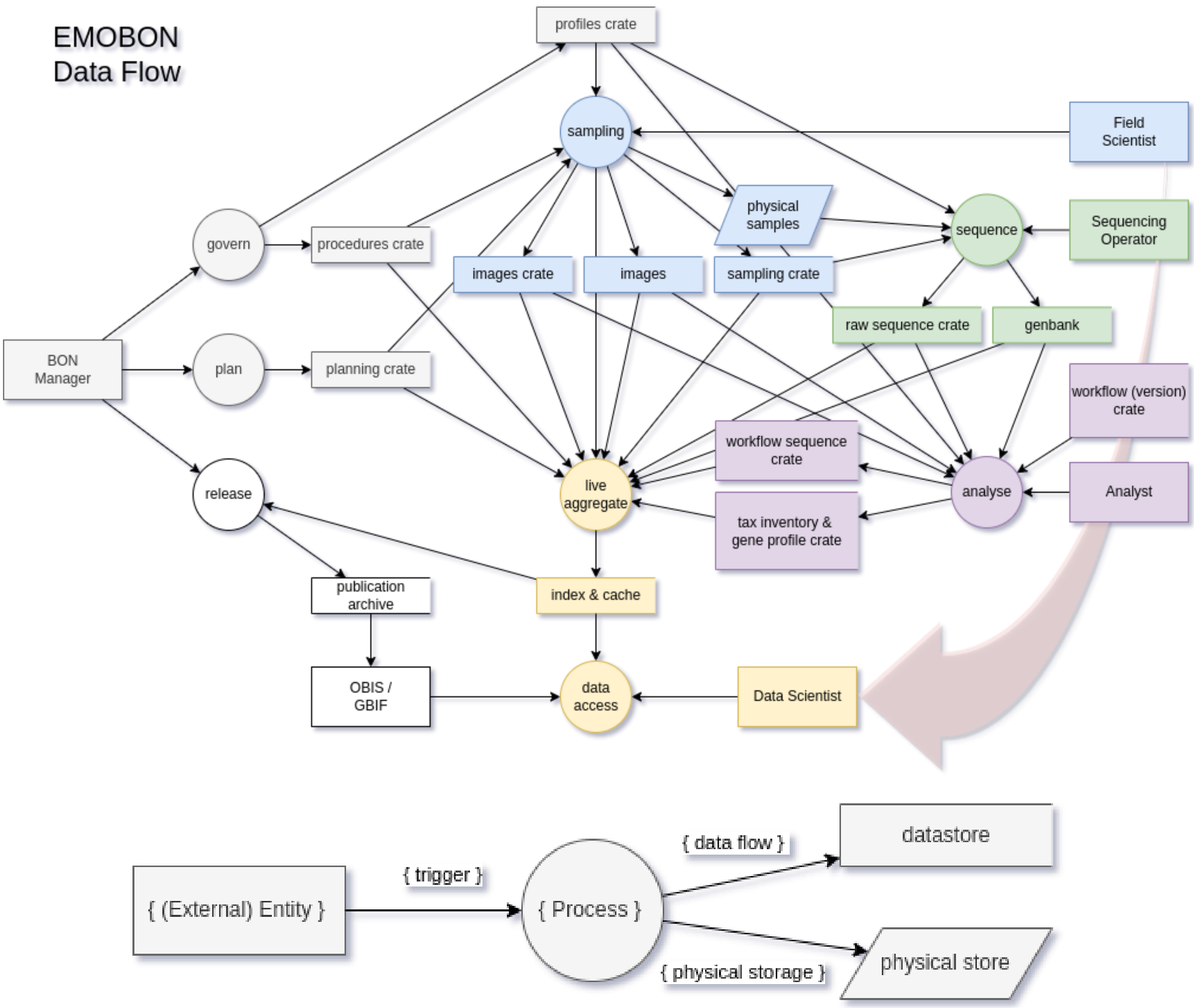

Figure 9 - Data Flow Diagram for the Marine Omics Observatory

The diagram legend is provided using a combination of basic UML generalization/specialization from Class Diagrams with a typical RDF visulalisation

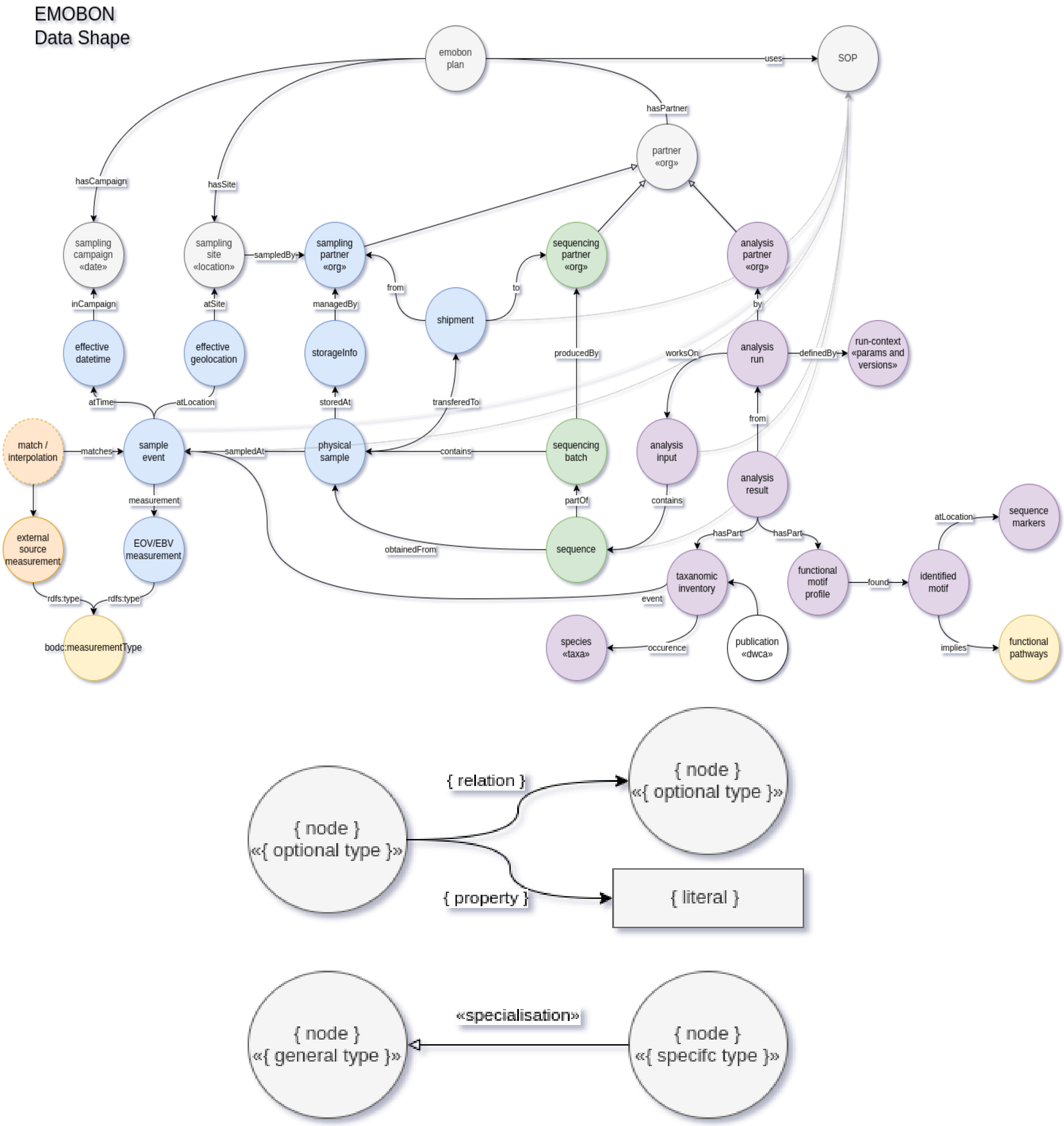

Figure 10 - DataShape diagram for the Marine Omics Observatory

Following the same color-coding as used in the Data-Flow Diagram, figure 10 gives a general view of the basic entities coming from the different steps in the flow, and how they are interconnected.

13.2.6 Key Resource Data and Datasets

The main taxonomic and biodiversity data repositories that will provide external (meta)data for the Pilot 3.1 VRE are the Global Biodiversity Information System (GBIF), the Ocean Biodiversity Information System (OBIS) and EurOBIS (its European partner), the World Register of Marine Species (WoRMS), and the NCBI GenBank Taxonomy. Ecological data, including access to long-term ecological observations data sets, is provided by the European Marine and Observation Network (EMODnet), while physical oceanographic data are available from SeaDataNet. (Meta)data taken from these sources will be/are used to annotate/identify the EMO BON data and will be used as complementary data as the user requires (based on time and location, and on species). The VRE bioprospecting service will rely on identifying biosynthetic gene clusters from the BiG-FAM repository. Details of these repositories can be found in the Template.xls document.

The EMO BON project utilizes RO-Crates for dissemination of raw data, metadata, and data products. The RO-Crates are provided by EMBRC and will be made available through the FAIR-EASE VRE/Blue Cloud infrastructure. The following are detailed descriptions of the EMO BON RO-Crate specifications:

–

ample Metadata RO-Crate: this RO-Crate contains the sample metadata collected during the sampling event and any other relevant provenance data for the sample – these Ro-Crates will be organized one per observatory and will be on (https://github.com/emo-bon/). These RO-Crates will include all metadata from sampling to shipping for analyses, and link to the EMBL Biosample entry. (Meta)data from here will be used to allow exploration of the EMO BON dataset (e.g. locations, dates, environmental observations);

–

etaGOflow Data Products RO-Crate: this RO-Crate contains the results and metadata for the analysis of 1 raw sequence data sample (i.e. 1 sequence result (shotgun sequencing reaction) from 1 DNA sample from 1 sampling event at 1 site) using the MetaGOflow workflow, and it's associated metadata. These will be primary data input to the Pilot VRE. Draft RO-C: https://github.com/emo-bon/MetaGOflow-Data-Products-RO-Crate. Note that the actual data files in these RO-Crates have various (ascii-based) formats, but these will additionally be complemented with the relevant (meta)data provided as RDF;

–

etaGOflow Sequence Data RO-Crate: this RO-Crate contains the Quality-Controlled sequence data (after adaptor removal and quality trimming of the raw data) and the assembled contigs from the MegaHit workflow assembly step. *These will be large files and we probably do not want these included* in 3) because it is the primary input to the FE VRE and therefore should be as "light" as possible. Draft RO-Crate: https://github.com/emo-bon/MetaGOflow-Sequence-Data-RO-Crate;

13.2.7 Key Metadata Descriptors

- RO-Crate Profiles (https://www.researchobject.org/ro-crate/profiles.html) to formally describe the content-structure of the datasources;
- Rules and software validators for agreed data formats (e.g. for csv delimiter, encoding, …);
- Vocabulary selection for semantic uplifting;
- Provenance vocabulary and model definition;

– Adoption of higher level feed vocabularies, e.g. dcat, ldes, DCMI, etc;

– RDF (turtle of JSON-LD).

Other technical details: as described above, access and authentication (AAI) to the VRE could be provided by the Blue Cloud infrastructure.

13.2.8 Expected Services/Applications in the FAIR-EASE

The EMO BON VRE intends to maintain compatibility with the Blue Cloud infrastructure and provide a flexible Data Access and Discovery and basic derived statistics and visualizations of the data that are common for metagenomic data analyses.

The EMO BON VRE should have 3 main functions:

1) as an interface to explore and again access to the EMO BON data using the Data Access and Discovery Service, as it will be implemented in FAIR-EASE (WP2),

2) to visualize and do simple statistical analyses of the data products that result from the metaGOflow workflow, good examples of these type of visualization of metagenomic data results are the Roscoff VRE (https://test.sb-roscoff.fr/mygod/) and the software BiomeShiny (https://github.com/BioData-PT/Biome-Shiny),

3) to provide analytical services, that run in real time, that use the EMO BON data products to answer novel marine research questions: a) the Bioprospecting Service, and b) the Ecological Strategies Service.

Expected Impact: The EMBRC EMO BON project is the first of its kind in European coastal waters, offering standardized, open, and FAIR metagenomic data and rich metadata from many sites in a long-term sampling time series. The intention is that the FAIR-EASE EMO BON VRE will provide the first point of access to the data products and services. It is expected that access to standardized, regular time-series marine community data will enhance Europe's ability to characterize, monitor, and improve the quality of marine coastal environments.

13.3 REPORT ON KEY REQUIREMENTS FROM THE PILOT

1. Data-driven requirements.

Firstly, we underline that the Sample Analysis data products to be ingested by the VRE are in the form of RO-Crate formatted data packets. Specifications of the various RO-Crates are provided above. The data payload of the RO-Crate to be ingested will be the data products resulting from the metaGOflow analyses, plus its associated metadata in the form of an additional linked RO-Crate. The data products themselves are CSV formatted ASCII text files. The fields within the data products, and the associated metadata for the sample, will be converted to RDF format and deployed to the "data lake" where it will be accessed by the VRE.

The following specific requirements are requested:

a. Design of a Sample Analysis RDF Schema (RDFS) defining a vocabulary for triples, including UML implementation diagram,

b. Software to convert Sample Analysis data to RDF (triples) and generation of the triples data sets,

c. Design of access API to the store of the triples,

d. Storage of triple in “data lake”.

2. VRE-driven requirements.

The technical synchronization with the Blue Cloud infrastructure will be an added value for the FAIR-EASE Earth Analytic Laboratory, for that reason its components should be compatible with the BC infrastructure.

The following specific requirements are requested:

- Specification of the software stack and software requirements needed for VRE deployment to Blue Cloud,
- Distributed Data as a Service (DDaaS) components for marine data external sources as currently implemented in the Blue Cloud Gateway (i.e. EMODNet, Eur/OBIS, GBiF, WoRMS, etc: API access software libraries,
- Additional DDaaS components not currently provided by Blue Cloud, to be determined in collaboration with WP2
- Data Discovery and Access service,
- Access and Authentication Service (AAI),
- Visualization components for metagenomics, new software visualization tools,
- Tools for calculation of basic statistical analyses and correlations among samples of the data,
- VRE Service 1 Bioprospecting analysis software and visual interpretation tools,
- VRE Service 2 Ecological Strategies Service, new software analyses and Interpretation.

13.3.1 Summary of Requirements

This Pilot 5.1.1 plans to use new and innovative services that are expected to be developed in WP3, and utilize multidisciplinary datasets and model outputs discovered using the services in WP2 and collated in WP4 to demonstrate seamless access to multiple datasets and to provide new ways for their dataset's connections and correlations. The primary data input are data products and their metadata (as RO-Crates) produced by the EMBRC EMO BON project that are private to the RI. All links to external data and metadata resources are in open public repositories. The data are to be triple-ized and stored in a data lake and made accessible through a VRE designed by the Earth Analytics Laboratory to provide public access to data, data visualization, and bespoke services.

Primary Data Sets (the metadata/data not publicly accessible from other repositories):

– EMBRC EMO BON Data Product RO-Crate (examples do not currently do not exist, only RO-Crate specification);
– URL - currently unavailable;
– Will be stored on EMBRC infrastructure;
– Private RI-only access;
– ~50GB per item; 100 items per year;
– Format: text string files, various standardized formats/vocabularies;
– Contains, and also links to, metadata.

13.4 POSSIBLE CROSS LINKS WITH OTHER PILOTS

Envisaged cross links with other FAIR-EASE Pilots are mainly with Pilot 5.1.1 Coastal Dynamics and Pilot 5.2.1 Ocean Bio-Geochemical Observations where some of the external repositories overlap - e.g. EMODNet. However, because the FAIR-EASE project adopts the Blue Cloud infrastructure, including its Distributed Data as a Service components that link to many external marine data resources through API's, many other cross links needed (e.g. OBIS, EMODNet, WorMS, etc) by the Pilot are already in-place.

Funded by the European Union

14 Analysis

14.1 MAIN OUTCOMES

D5.1 describes the scientific contexts and aims of each of the five Pilots representing the three different use cases in FAIR-EASE. It provides the description of main data resources and current workflows in each Pilot with the aim to highlight major bottlenecks and possible improvements that, despite fulfilling primary tasks and useful demonstrators per Pilot, can provide novel tools and strategies to external stakeholders that need to exploit similar resources.

The principal general aims per Pilot are here summarized:

Pilot 5.1.1: Coastal Waters Dynamics

- Develop and test innovative web services for scientific analysis and visualization of multidisciplinary datasets and model results;

- Combine new web technology with multidisciplinary datasets and establish the Pilot's website targeted for scientific users and the general public;

- Contribute a wide variety of community datasets (rivers, meteorology , satellites , water column physical and chemical data, models).

Pilot 5.1.2: Earth Critical Zone

- Set up methodologies to analyze the Earth Critical Zone (ECZ) information with the specific aim to improve the assessment of land and soil degradation to meet the sustainable goal SDG 15.3.

Pilot 5.1.3: Volcano Space Observatory

- Build a web interface capable of interactively aggregating and jointly analyzing satellite observations from Solid Earth and Atmospheric communities for the near-real-time monitoring of volcanic activity.

Pilot 5.2.1: Ocean Bio-Geochemical Observation

- Provide a common platform of services for sensor qualification, calibration and/or marine BGC data validation to meet the needs of the user community (for example the interoperability and the resources access).

Pilot 5.3.1: Marine Omics Observation

- Set up a web-based VRE to provide products and services targeted at non-specialist researchers interested in omics approaches to study marine biodiversity.

This document highlights that all Pilots are facing comparable challenges, i.e. the management of heterogeneous datasets obtained by multidisciplinary efforts from different reference resources, some of which are of common interest, and the modeling of relevant aspects related to different Earth environmental and biodiversity dynamics. This is one of the main reasons why FAIR-EASE was designed involving the current set of Pilots: beyond their diversification, they are all focused on Earth systems and on their related data, such as sea and coastal waters, soil, air, biotic resources. Therefore, crossdomain interactions in this multidisciplinary context will favor novel opportunities during the progress of the project, permitting the unfolding of more ambitious goals thanks to intrinsic overlapping interests, although apparently different specific aims. Indeed, the access to collections in common data repositories and the possible sharing of methodologies meet the challenges of developing gateways between metadata standards, contributing to an interoperable framework that will allow the matching of data representations designed for different domains across disciplines.

14.2 ACTIVITIES SET UP

To achieve this deliverable during the initial stage of the project, the principal aim in WP5 was to speed up the start up of the activities, with the help and the support of the Technical Board.

Main initial actions in WP5 were focused on the tuning and recapitulation of the objectives of the different Pilots in FAIR-EASE, while driving the start up of an operative framework and an integrative effort.

Bimonthly WP5 internal meetings and three main workshops (reported in the Annex B) were focused on the general presentation of the Pilots (Kick-off meeting), a more detailed presentation to the technical WPs (Workshop 1), and exchange of more technical details typical of each Pilot (one to one meetings, during which each Pilot confronts the technical WPs), respectively.

Finally, the organization of a uniform schema to present each Pilot was planned to provide a comprehensive document and establish the overall framework of targets, planned demonstrators and expected requirements from WP5 Pilots.

This document was prepared with the contribution of all the people representing the different scientific communities involved in FAIR-EASE. Specifically, the Pilot Reference Persons in WP5 took the responsibility to interact with all the partners in each Pilot and appropriately define the main expected technological improvements to meet the proposed goals, that are here reported.

This document was a helpful starting point for interactions within WP5, also favoring those with the technical WPs. Future efforts will be targeted at reinforcing the sharing and the technical dialogue, establishing a suitable framework to fix how to face common needs that can be of general interest. Because of the relevance of these aspects according to the call and the project, and the advances that will be acquired during its progress, this objective will be under constant attention during the entire course of the project.

14.3 CHALLENGES AND OPPORTUNITIES

The extensive effort required to face the first deliverable of this WP presented major predictable initial challenges that needed to be faced in the very early timing of the project, while all partners were tuning and focusing on their specific plans, starting their interaction within WP5 and with other WPs.

Specifically, in this initial stage, all the partners had to take into consideration that Pilots activities are embedded in the unique framework of WP5, favoring possible cross links emerging from specific needs. This was not a minor aspect, neither a minor effort in the work package and in the project, also considering that the deliverable was necessarily planned at the beginning of the project and defines the backbone to focus on the next ones.

It's evident also from this report that all the members of the different Pilots have expertises in different domains, all relevant for Life Sciences. They belong to reference scientific communities that rely on well-established workflows. They joined the project with clear aims, intending to offer new tools and/or to improve the exploitation of currently available open resources and of some of the conventional methodologies useful to their interest and in their community, related to the specific targets of the proposed Pilots in each use case.

The peculiarities of the Pilots make them appear almost heterogeneous. Nevertheless, this was exactly for the purpose of FAIR-EASE: all the pilots deal with similar data resources and often follow similar workflows, in terms of data type and their management, highlighting that the improvement of some technological aspects can cover multiple interests in the frame of a unique effort in the project. Favoring exchange of technical information paves the way to synergies that clearly emerge from these initial trans-domain cross-talk.

The assessment of the different strategies and the validation of the interoperability of the different data resources are the fundamental aims of the project. Therefore the scientific context in FAIR-EASE will determine the necessary cross-domain interaction among the different communities to meet these goals, while facing the proposed challenges and offering competitive and scalable tools during the implementation of the Earth Analytics Lab expected in the project. This is expected to support a FAIR perspective while fulfilling specific objectives and goals that arise from the needs and from the current states of the art in each community.

14.4 PRIORITY AND GENERAL REQUIREMENTS

The document clearly depicts the main tasks and requirements from each Pilot. They range from the implementation/improvement of computing resources to efficiently manage integrated complex collections (Pilot 5.1.1, 5.1.3, 5.2), to the implementation of indicators useful to address sustainable goals that can be provided as demonstrators to the entire European community (5.1.2), to the implementation of tools to investigate biodiversity assets (5.3) exploiting omics resources for bioprospecting and ecological strategies.

It emerges from this deliverable that each specific Pilot has its own challenges with specific needs and goals. However, despite the apparent high heterogeneity among Pilots, and the need to still evolve towards a common language within the project, it is nevertheless also evident that main priorities and possible cross links clearly emerge.

The priorities per all the Pilots and expectations can be summarized as follows:

1. develop and test innovative services for straightforward analyses and visualization of complex datasets;

2. efficient accessibility to data resources for modeling and near real time monitoring;

3. improve methods and technologies to contribute to sustainable goals and priorities;

4. contribute data and methods useful to different scientific communities with heterogeneous level of expertises (user friendly visualization or on demand processing by command line approaches for different activities such as validation, calibration, comparison, modeling, prediction, assessments);

5. combine web technology with multidisciplinary datasets and establish data access targeted both to expert users and to a more general audience.

14.5 CROSS LINKS

From these main expectations, it emerges the need to exploit and validate pre-operational existing services that appear heterogeneous among Pilots, aiming to improve a FAIR-EASE data access and to fix possible common goals when considering access to complex data or partitions of complex collections.

Interestingly, since the beginning of the project some common requirements soon emerged.

All Pilots claim for efficient data access to specific datasets or their partitions, mainly from well established repositories, and for workspace to set up appropriate analytical platforms or virtual environments, or to integrate with existing ones. Cross links with already available resources within EOSC need to be reinforced or established.

In addition, during the presentation of the Pilots several overlapping requirements on data resources emerged. Annex A clearly shows that main reference resources are common to several Pilots, while community specific technological requirements can be a starting point for further validations and possible improvements arising from an interdomain agreement on best solutions and standards, the first step towards improved interoperability and reusability.

Interestingly, strong interconnections can be foreseen and plannings are being finalized, while the interactions are occurring in the initial stage of the project.

15 Conclusions

D5.1 was a necessary effort to provide a first assessment of FAIR-EASE Pilots reference resources and main workflows for their implementation, together with the overview of the state of the art per each Pilot and a clear definition of their demonstrators to make them understandable in a multidisciplinary context. The effort highlights the key requirements and the bottlenecks and/or useful expected improvements to be implemented in FAIR-EASE, to increase capabilities and improve the expected demonstrators beyond the current landscape.

The mapping exercise that is reported in this deliverable (Annex A) helped to identify the relevant assets and state of development of data resources and methods exploited by the different Pilots in FAIR-EASE, and to establish focused collaborative activities in areas of common interests, oriented to the exchange of expertises and knowledge, and to the integration of resources supporting the offering of the EOSC Portal.

The activities that were organized to meet this deliverable were helpful in laying foundation to fulfill the FAIR-EASE goals, i.e. to provide researchers with a set of highly innovative new or improved services that would exploit, in a structural way, cloud-based EOSC technologies and European computing and data management capacities.

The priority of WP5 is to address the current needs of the Pilots to start the interaction with the technical WPs, to plan the improvements of their workflows and the FAIRification of the resources. This anticipates the needs of possible stakeholders and of the associated research communities, which is one of the goals of FAIR-EASE.

This preliminary overview also highlights the need to compare and tune the pre-operational services currently available, that can be customized and integrated in the existing workflows across different disciplines, facilitating cross-disciplinary collaborations, reducing the time to results and increasing productivity, but also establishing the appropriate interactions for emerging novel opportunities. Thus, WP5, with the help of the technical WPs, will address primarily EOSC resources at a mature level, also trying to facilitate the evolving and adoption of emerging ones. Building on this initial analysis, the technical WPs will shape best practices for processes, specifications, guidelines, tools and APIs necessary for:

a) opening the "market" (interoperability), thus offering the opportunity to register voluminous resources under EOSC Providers, aggregating them in the EOSC Portal with minimum disturbance for the EOSC Providers.

b) Deploying processes that allow for the provision of added value services by combining these processes with other resources.

c) Integrating processes that allow for an automated update of resources descriptions and their metadata without human intervention (e.g., open APIs, harvesting, etc.).

In addition, the work of these 4 months also aimed to favor a relevant aspect to be fulfilled in the project, i.e. the starting of cross domain interactions to investigate possible intersections. This is favoring discussions and the set up of common plannings to push forward the amelioration of the available strategies and also of the FAIRification of data resources and tools, and, possibly, it will pave the way to the definition of additional standards in data organization and analytical strategies in the EOSC framework.

Data FAIRification is one of the key actions in the project. The diversification of each scientific community involved in FAIR-EASE will enhance the impact of this effort on improving the TRL of the current available components and the framework of the European Open Science Cloud (EOSC) ecosystem of data and services. This will push far beyond the state-of-the-art, and fulfill the overall objective to customize and operate distributed and integrated services for observation and modeling of the Earth system, environment and biodiversity, which is the main objective of FAIR-EASE.

D5.1 reports the main technical issues that are expected to be solved by the Pilots in the project. Of course WP5 also expects an iterative process to improve the plannings, and this deliverable is the first step of a continuous and inclusive development in close cooperation/co-design with other WPs in FAIR-EASE and user communities. This is the suitable framework for the integration of the service-based architecture offered through the EOSC.

16 References

Abreu, A., Bourgois, E., Gristwood, A., Troublé, R., Acinas, S. G., Bork, P., et al. (2022). Priorities for ocean microbiome research. Nature Microbiology 7, 937–947. doi: 10.1038/s41564-022-01145-5

Bellacicco, M., Cornec, M., Organelli, E., Brewin, R.J.W., Neukermans, G., Volpe, G., et al. (2019). Global variability of optical backscattering by non-algal particles from a Biogeochemical-Argo dataset. Geophysical Research Letters 46, 9767-9767. doi: 10.1029/2019gl084078

Barth, A., Beckers, J.M., Troupin, C., Alvera-Azcárate, A., Vandenbulcke, L. (2014). DIVAnd-1.0: n-dimensional variational data analysis for ocean observations. Geoscientific Model Development 7, 225-241, doi: 10.5194/gmd-7-225-2014

Boichu, M., Grandin, R. (2021). La télédétection épaule la surveillance de l'éruption de la Soufrière de Saint-Vincent. La météorologie 114, 2-4. doi: 10.37053/lameteorologie-2021-0078

Boichu, M., Clarisse, L., Péré, J. C., Herbin, H., Goloub, P., Thieuleux, F., et al. (2015). Temporal variations of flux and altitude of sulfur dioxide emissions during volcanic eruptions: implications for long-range dispersal of volcanic clouds. Atmospheric Chemistry and Physics 15, 8381-8400. doi: https://doi.org/10.5194/acp-15-8381-2015

Bonaldo, D., Antonioli, F., Archetti, R., Bezzi, A., Correggiari, A., Davolio, S., ... & Carniel, S. (2019). Integrating multidisciplinary instruments for assessing coastal vulnerability to erosion and sea level rise: Lessons and challenges from the Adriatic Sea, Italy. Journal of coastal conservation, 23(1), 19-37.

Brown, M. (1998). Ocean Data View 4.0. Oceanography 11, 19–21. doi: 10.5670/oceanog.1998.04

Canonico, G., Buttigieg, P. L., Montes, E., Muller-Karger, F. E., Stepien, C., Wright, D., et al. (2019). Global Observational Needs and Resources for Marine Biodiversity. Frontiers in Marine Science 6. doi: 10.3389/fmars.2019.00367

Chai, F., Johnson, K.S., Claustre, H., Xing, X., Wang, Y., Boss, E., et al. (2020). Monitoring ocean biogeochemistry with autonomous platforms. Nature Reviews Earth & Environment 1, 315-326. doi: 10.1038/s43017-020-0053-y

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., et al. (2020). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. Molecular Ecology 30, 2937-2958. doi: 10.1111/mec.15472

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. Science 348, 1261605. doi: 10.1126/science.1261605

Estelmann, A., Bittig, H., Körtzinger, A., Paba, V., Donnelly, M. (2021). Euro-Argo RISE Deliverable 4.6: Recommendations for Enhancement of O2 QC Methods: Guide to Drift Correction Procedure for Oxygen Optodes on Argo Floats. Zenodo. doi : 10.5281/zenodo.7369128

Faust, K., Lahti, L., Gonze, D., de Vos, W. M., Raes, J. (2015). Metagenomics meets time series analysis: Unraveling microbial community dynamics. Current Opinion in Microbiology 25, 56–66. doi: 10.1016/j.mib.2015.04.004

Flaviani, F., Schroeder, D. C., Lebret, K., Balestreri, C., Highfield, A. C., Schroeder, J. L., et al. (2018). Distinct oceanic microbiomes from viruses to protists located near the Antarctic Circumpolar current. Frontiers in Microbiology 9, 1474. doi: 10.3389/fmicb.2018.01474

Irisson, J.O., Bittner, L., Schickele, A. (2022). Plankton Genomics. Zenodo. doi: 10.5281/zenodo.7277346

Johnson, K.S., W.M. Berelson, E.S. Boss, Z. Chase, H. Claustre, S.R. Emerson, N. et al. (2009). Observing biogeochemical cycles at global scales with profiling floats and gliders: Prospects for a global array. Oceanography 22, 216–225. doi:10.5670/oceanog.2009.81

Jutard, Q., Organelli, E., Briggs, N., Xing, X. G., Schmechtig, C., Boss, E., Poteau, A., Leymarie, E., Cornec, M., D'Ortenzio, F., Claustre, H. (2021)a. Correction of Biogeochemical-Argo Radiometry for Sensor Temperature-Dependence and Drift: Protocols for a Delayed-Mode Quality Control. Sensors 21. doi: 10.3390/s21186217

Jutard, Q., Organelli, E., D'Ortenzio, F., Claustre, H., Schmechtig, C. (2021)b. Euro-Argo RISE Deliverable 4.4: Recommendations for enhancement of Irradiance QC Methods. Zenodo. doi : 10.5281/zenodo.7369064

Makiola, A., Compson, Z. G., Baird, D. J., Barnes, M. A., Boerlijst, S. P., Bouchez, A., et al. (2020). Key Questions for Next-Generation Biomonitoring. Frontiers in Environmental Science 7, 197. doi: 10.3389/fenvs.2019.00197

Martin-Platero, A. M., Cleary, B., Kauffman, K., Preheim, S. P., McGillicuddy, D. J., Alm, E. J., et al. (2018). High resolution time series reveals cohesive but short-lived communities in coastal plankton. Nature communications 9, 266. doi: 10.1038/s41467-017-02571-4

Maurer, T.L., Plant, J.N., Johnson, K.S. (2021). Delayed-Mode Quality Control of Oxygen, Nitrate, and pH Data on SOCCOM Biogeochemical Profiling Floats. Frontiers in Marine Science 8. doi:10.3389/fmars.2021.683207

Mendoza, M.L.Z., Sicheritz-Pontén, T., Gilbert, T.M.P. (2014). Environmental genes and genomes: Understanding the differences and challenges in the approaches and software for their analyses. Briefings in Bioinformatics 16, 745–758. doi: 10.1093/bib/bbv001

Milici, M., Deng, Z. L., Tomasch, J., Decelle, J., Wos-Oxley, M. L., Wang, H., et al. (2016). Co-occurrence analysis of microbial taxa in the Atlantic ocean reveals high connectivity in the free-living bacterioplankton. Frontiers in Microbiology 7, 649. doi: 10.3389/fmicb.2016.00649

Oliveri, P., Simoncelli, S. (2022). SOURCE v1.4.0 (1.4.0). Zenodo. https://doi.org/10.5281/zenodo.6319836

Oliveri, P., Simoncelli, S., Di Pietro, P., Fratianni, C., Mattia, G., Delrosso, D., et al. (2022). SOURCE: Sea Observations Utility for Reprocessing, Calibration and Evaluation. Frontiers in Marine Science 8:750387. doi: 10.3389/fmars.2021.750387

Ricci, F., Capellacci, S., Campanelli, A., Grilli, F., Marini, M., & Penna, A. (2022). Long-term dynamics of annual and seasonal physical and biogeochemical properties: Role of minor river discharges in the North-western Adriatic coast. Estuarine, Coastal and Shelf Science, 107902.

Rodríguez-Ezpeleta, N., Zinger, L., Kinziger, A., Bik, Bonin, A., Coissac, E., et al. (2021). Biodiversity monitoring using environmental DNA. Molecular ecology resources 21. doi: 10.1111/1755-0998.13399

Schlitzer, R., Mieruch-Schnülle, S. (2023) webODV: Interactive online data analysis and visualization, https://webodv.awi.de.

Schlitzer, R. (2015). Data Analysis and Visualization with Ocean Data View. CMOS Bulletin SCMO 43, 9-13. doi: 10013/epic.45187

Schlitzer, R. (2002). Interactive analysis and visualization of geoscience data with Ocean Data View. Computers & Geosciences 28, 1211-1218. doi: https://doi.org/10.1016/S0098-3004(02)00040-7

Schlitzer, R. (2000). Electronic atlas of WOCE hydrographic and tracer data now available. Eos, Transactions American Geophysical Union 81, 45. doi: https://doi.org/10.1029/00EO00028

Schmechtig, C., D'Ortenzio, F., Jutard, Q., Sauzède, R., Renosh, P.R., Poteau, A., et al. (2021). Euro-Argo RISE Deliverable 4.2: Recommendations for enhancement of CHLA QC Methods. Zenodo. doi: 10.5281/zenodo.7369043

Shreve, T., Grandin, R., Boichu, M., Garaebiti, E., Moussallam, Y., Ballu, V., et al. (2019). From prodigious volcanic degassing to caldera subsidence and quiescence at Ambrym (Vanuatu): the influence of regional tectonics. Scientific reports 9, 1-13. doi: https://doi.org/10.1038/s41598-019-55141-7

Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L. J., Coppens, F., Fernández, J. M., et al. (2022). Packaging research artifacts with RO-Crate. Data Science 5. doi: 10.3233/DS-210053

Umgiesser, G., Ferrarin, C., Bajo, M., Bellafiore, D., Cucco, A., De Pascalis, F., ... & Arpaia, L. (2022). Hydrodynamic modelling in marginal and coastal seas—The case of the Adriatic Sea as a permanent laboratory for numerical approach. Ocean Modelling, 102123.

Vinci, M., Giorgetti, A., & Brosich, A. (2011). OGS NODC Meta-data standardization, discovery and reporting from a relational database. Mediterranean Marine Science, 39-44.

Funded by the European Union

17 Annexes

17.1 Annex A. Resource and Tools Mapping Table

The activities in WP5 aimed to organize a document summarizing the data resources and the methodologies that are expected to be employed to implement the demonstrators planned within each Pilot.

The document is released as Annex A and it is also expected to evolve during the course of the project to represent the reference resources of interest in FAIR-EASE. Finally, it will be included in the project DMP.

The link to the mapping table that will be maintained during the course of the whole project is available at:

https://docs.google.com/spreadsheets/d/18uRj4MFaCDJPcbcuscy1KdTG95h_imhj/edit#gid=1818952351

Table 1 and 2 reports a summary of the content in Annex A.

Table 1 - Summary list of the Data Resources

PILOT	AREA	THEME	REFERENCE RESOURCE	LINK	DATASET	PREFERRED LINK TO THE DATASET
Coastal Dynamics 5.1.1	Ocean	Ocean observations	Copernicus Marine and Environment Monitoring System (CMEMS)	https://data.marine.copernicus.eu/product/OCEANCOLOUR_MED_BGC_L4_MY_009_144/description	Surface Water Chlorophyll	
Coastal Dynamics 5.1.1	Ocean	Ocean observations	Copernicus Marine and Environment Monitoring System (CMEMS)	https://marine.copernicus.eu/about/producers/sst-tac	Surface Water Temperature	
Coastal Dynamics 5.1.1	Ocean	Model Reanalysis	Copernicus Marine and Environment Monitoring System (CMEMS)	https://data.marine.copernicus.eu/product/MEDSEA_MULTIYEAR_BGC_006_008/description?view=-&product_id=-&option=-	Ocean nutrients reanalysis	
Coastal Dynamics 5.1.1	Ocean	Ocean observations	Copernicus Marine and Environment Monitoring System (CMEMS)	https://data.marine.copernicus.eu/product/INSITU_GLO_PHY_TS_OA_MY_013_052/description	Original Temperature and Salinity observations	
Coastal Dynamics 5.1.1	Ocean	Eutrophication	European Marine Observation and Data Network (EMODnet) - Chemistry	https://emodnet.ec.europa.eu/en/checkpoint/black-sea/challenges/eutrophication	Original nutrient observations	https://emodnet-chemistry.webodv.awi.de/eutrophication%3EMediterranean%3EEutrophication_Med_profiles_2021/service/DataExploration
Coastal Dynamics 5.1.1	Rivers	River Runoff	Global Runoff Data Centre (GRDC)	https://explore.webodv.awi.de/	Original, daily river runoff data	https://explore.webodv.awi.de/rivers/discharge/grdc/daily/
Coastal Dynamics 5.1.1	Meteorology	Meteorological Observations	Global Historical Climatology Network - Daily (GHCN-Daily) v. 3,26	http://ec2-52-38-26-42.us-west-2.compute.amazonaws.com:8080/dataset/noaa-ncdc-c00861	Original, daily meteorological data	https://explore.webodv.awi.de/atmosphere/meteorology/ghcn/daily/

Coastal Dynamics 5.1.1	Ocean	Ocean observations	European Marine Observation and Data Network (EMODnet) - Physics	https://erddap.emodnet-physics.eu/erddap/tabledap/EP_ERD_L2A_TEMP_MO_PR_NT_GLO.html	Original, Sea Temperature (TEMP) Profiles from Mooring	
Coastal Dynamics 5.1.1	Ocean	Ocean observations	European Marine Observation and Data Network (EMODnet) - Physics	https://erddap.emodnet-physics.eu/erddap/tabledap/EP_ERD_L2A_PSAL_MO_PR_NT_GLO.html	Original, Practical Salinity (PSAL) Profiles from Mooring	
Earth Critical Zone 5.1.2	Soil	Soil database /Soil map	European Soil DAta Centre (ESDAC)	https://esdac.jrc.ec.europa.eu/	Soil Geographical Database of Europe (SGDBE) at scale 1:1.000.000	https://esdac.jrc.ec.europa.eu/tmp_dataset_access_req_19571
Earth Critical Zone 5.1.2	Soil	Soil database /Soil map	European Soil DAta Centre (ESDAC)	https://esdac.jrc.ec.europa.eu/	Global Soil Organic Carbon Map (GSPmap)	https://www.fao.org/global-soil-partnership/pillars-action/4-information-and-data-new/global-soil-organic-carbon-gsoc-map/en/
Earth Critical Zone 5.1.2	Soil	Soil database /Soil map	European Soil DAta Centre (ESDAC)	https://esdac.jrc.ec.europa.eu/	Top Soil Organic Carbon content of the Land Use/Land Cover Area Frame Survey (LUCAS)	https://esdac.jrc.ec.europa.eu/content/topsoil-soil-organic-carbon-lucas-eu25
Earth Critical Zone 5.1.2	Soil	Soil database /Soil map	European Soil DAta Centre (ESDAC)	https://esdac.jrc.ec.europa.eu/	Soil Organic Carbon Stocks by Climate_Scenarios	https://esdac.jrc.ec.europa.eu/content/soil-organic-carbon-soc-projections-europe
Earth Critical Zone 5.1.2	Soil	Soil database /Soil map	European Soil DAta Centre (ESDAC)	https://esdac.jrc.ec.europa.eu/	Soil Organic Carbon Saturation Capacity	https://esdac.jrc.ec.europa.eu/content/soil-organic-carbon-saturation-capacity
Earth Critical Zone 5.1.2	Soil	Soil database /Soil map	European Soil DAta Centre	https://esdac.jrc.ec.europa.eu/	Soil Hydrology grid 250m - EU-SoilHydroGrids	https://esdac.jrc.ec.europa.eu/content/3d-soil-hydraulic-

Funded by the European Union

			(ESDAC)		ver 1.0	database-europe-1-km-and-250-m-resolution
Earth Critical Zone 5.1.2	Soil	Soil Erosion	European Soil DAta Centre (ESDAC)	https://esdac.jrc.ec.europa.eu/	Soil Erodibility Factor 500m (K-Factor)	https://esdac.jrc.ec.europa.eu/themes/soil-erodibility-europe
Earth Critical Zone 5.1.2	Soil	Soil Erosion	European Soil DAta Centre (ESDAC)	https://esdac.jrc.ec.europa.eu/	Water erosion map 100m resolution by using WaTEM/SEDEM model	https://esdac.jrc.ec.europa.eu/themes/sediment-transport-using-watemsedem
Earth Critical Zone 5.1.2	Soil	Soil Erosion	European Soil DAta Centre (ESDAC)	https://esdac.jrc.ec.europa.eu/	Soil loss by wind erosion	https://esdac.jrc.ec.europa.eu/content/Soil_erosion_by_wind
Earth Critical Zone 5.1.2	Soil	Soil Erosion	European Soil DAta Centre (ESDAC)	https://esdac.jrc.ec.europa.eu/	Index of Land Susceptibility to Wind Erosion (ILSWE 1981-2010)	https://esdac.jrc.ec.europa.eu/wyz_856/_20_wer/ILSWE.zip
Earth Critical Zone 5.1.2	Soil	Other soil threats: compaction, pollution…	European Soil DAta Centre (ESDAC)	https://esdac.jrc.ec.europa.eu/	Natural susceptibility to soil compaction in Europe 1Km	https://esdac.jrc.ec.europa.eu/content/natural-susceptibility-soil-compaction-europe
Earth Critical Zone 5.1.2	Soil	Ecosystem Services	Joint Research Centre Data Catalog	https://data.jrc.ec.europa.eu/	Water Retention Index (WRI) 2010	https://data.jrc.ec.europa.eu/dataset/06c3f085-c1e3-4228-949d-82a0899b8d7d
Earth Critical Zone 5.1.2	Soil	Ecosystem Services	Joint Research Centre Data Catalog	https://data.jrc.ec.europa.eu/	Mapping and Assessment of Ecosystems and their Services (10 Km resolution)	https://data.jrc.ec.europa.eu/dataset/7e3f0681-5967-41f7-ae9b-87f1c3cfac4f
Earth Critical Zone 5.1.2	Soil	Ecosystem Services	Joint Research Centre Data Catalog	https://data.jrc.ec.europa.eu/	Forest Carbon Potential 2000-2010	https://data.jrc.ec.europa.eu/dataset/ae111193-b95e-40ac-901c-02290613bf0f
Earth Critical Zone 5.1.2	Soil	Ecosystem Services	Joint Research Centre Data Catalog	https://data.jrc.ec.europa.eu/	Soil Biomass Productivity maps of grasslands and pasture, of	https://esdac.jrc.ec.europa.eu/content/soil-biomass-productivity-maps-grasslands-and-

					croplands and of forest areas in the European Union	pasture-coplands-and-forest-areas-european
Earth Critical Zone 5.1.2	Soil	Land use/Land cover	Copernicus	https://land.copernicus.eu/	Corine Land Cover 1990, 2000, 2006, 2012, 2018 (100m resolution)	https://land.copernicus.eu/pan-european/corine-land-cover
Earth Critical Zone 5.1.2	Soil	Land use/Land cover	Copernicus	https://land.copernicus.eu/	European Settlement Map 2016, 2017, 2019	https://land.copernicus.eu/pan-european/GHSL/european-settlement-map
Earth Critical Zone 5.1.2	Soil	Land use/Land cover	Copernicus	https://land.copernicus.eu/	Imperviousness	https://land.copernicus.eu/pan-european/high-resolution-layers/imperviousness
Earth Critical Zone 5.1.2	Soil	Elevation and morphological features			Elevation (DEM) 25x25Km	https://data.jrc.ec.europa.eu/dataset/7e3f0681-5967-41f7-ae9b-87f1c3cfac4f
Earth Critical Zone 5.1.2	Soil	Elevation and morphological features	Copernicus	https://land.copernicus.eu/	EU-DEM v1.1 is a digital surface model (DSM)	https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1?tab=mapview
Earth Critical Zone 5.1.2	Soil	Elevation and morphological features	Joint Research Centre Data Catalog	https://data.jrc.ec.europa.eu/	Global Landform Classification	https://esdac.jrc.ec.europa.eu/content/global-landform-classification
Earth Critical Zone 5.1.2	Geology and hydrogeology	Hydrogeology	Data Europa	https://data.europa.eu	Internationale Hydrogeologische Karte von Europa 1:1.500.000 (IHME1500)	https://data.europa.eu/data/datasets/341255a9-180f-4bf9-b96f-d085339ea86d?locale=it
Earth Critical Zone 5.1.2	Geology and hydrogeology	Other geological themes	Joint Research Centre Data Catalog	https://data.jrc.ec.europa.eu/	European Landslide Susceptibility Map version 2 (ELSUS v2)	https://esdac.jrc.ec.europa.eu/content/european-landslide-susceptibility-map-elsus-v2

Funded by the European Union

Earth Critical Zone 5.1.2	Geology and hydrogeology	Other geological themes	Joint Research Centre Data Catalog	https://data.jrc.ec.europa.eu/	River flood hazard map for Europe and the Mediterranean Basin region - 20-year return period	https://data.jrc.ec.europa.eu/dataset/jrc-floods-floodmapeu_rp20y-tif
Earth Critical Zone 5.1.2	Water	Rivers and lakes maps	European Environment Agency	https://www.eea.europa.eu/	Water Information System for Europe (WISE)	https://www.eea.europa.eu/data-and-maps/data/wise-wfd-spatial-3
Earth Critical Zone 5.1.2	Nature and biodiversity	Soil Biodiversity	Joint Research Centre Data Catalog	https://data.jrc.ec.europa.eu/	Global Soil Biodiversity Atlas Maps	https://data.jrc.ec.europa.eu/dataset/jrc-esdac-104
Earth Critical Zone 5.1.2	Habitat	Habitat	European Environment Agency	https://www.eea.europa.eu/	European inventory of nationally designated protected areas	https://www.eea.europa.eu/data-and-maps/data/nationally-designated-areas-national-cdda-17
Earth Critical Zone 5.1.2	Habitat	Habitat	European Environment Agency	https://www.eea.europa.eu/	Ecosystem types of Europe	https://www.eea.europa.eu/data-and-maps/data/ecosystem-types-of-europe-1
Earth Critical Zone 5.1.2	Habitat	Habitat	European Environment Agency	https://www.eea.europa.eu/	Natura 2000 data - the European network of protected sites	https://www.eea.europa.eu/data-and-maps/data/natura-14
Earth Critical Zone 5.1.2	Green infrastructures	Landscape elements	Copernicus	https://land.copernicus.eu/	Green Linear Elements	https://land.copernicus.eu/local/riparian-zones/green-linear-elements-gle-image
Earth Critical Zone 5.1.2	Green infrastructures	Riparian zones	Copernicus	https://land.copernicus.eu/	Delineation of Riparian Zones	https://land.copernicus.eu/local/riparian-zones/riparian-zones-delineation/view
Earth Critical Zone 5.1.2	Eurostat	Population data, Agriculture data, IASK support			EU_2012-2015-2016-2017-2018 - LAU2 NUTS2	
Earth Critical Zone 5.1.2	Eurostat	Population data, Agriculture			EU_historical population (1961-2011)	

		data, IASK support			LAU2 LAU1	
Volcano Space Observatory 5.1.3	Products and Algorithms	Atmosphere	European Space Agency	https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-5p/products-algorithms	Sentinel S5P Sulfur dioxide, Aerosol Index and Aerosol layer height Level 2 data	https://s5phub.copernicus.eu/dhus/#/home
Volcano Space Observatory 5.1.3	Earth observation	Atmosphere	NASA	https://modaps.modaps.eosdis.nasa.gov/services/about/products/c6/MYD04_L2.html	MODIS Aerosol level 2 product	https://www.icare.univ-lille.fr/
Volcano Space Observatory 5.1.3	Earth observation	Model	Copernicus	https://cds.climate.copernicus.eu/cdsapp#!/dataset/10.24381/cds.bd0915c6?tab=overview	ECMWF ERA5 hourly data	https://www.icare.univ-lille.fr/
Volcano Space Observatory 5.1.3	Earth observation	Atmosphere	NASA	https://www.nasa.gov/mission_pages/calipso/mission/	CALIOP L1B data	https://www.icare.univ-lille.fr/
Volcano Space Observatory 5.1.3	Earth observation	Atmosphere	NASA	https://aeronet.gsfc.nasa.gov/	Aeronet L1, L15 and L2 data	https://aeronet.gsfc.nasa.gov/
Volcano Space Observatory 5.1.3	Earth observation	Atmosphere	GEOD'AIR	https://www.geodair.fr	GEODAIR data (SO2, PM, etc)	https://www.geodair.fr/donnees/api
Volcano Space Observatory 5.1.3	Earth observation	Atmosphere	Icare-Aeris	https://www.icare.univ-lille.fr/parasol/	PARASOL/POLDER Level 2 products	https://www.icare.univ-lille.fr/
Volcano Space Observatory 5.1.3	Earth observation	Atmosphere	NASA	https://www.earthdata.nasa.gov/learn/find-data/near-real-time/omps	SUOMI-NPP/OMPS L2 products	https://www.icare.univ-lille.fr/
Volcano Space Observatory 5.1.3	Earth observation	Atmosphere	European Space Agency	https://navigator.eumetsat.int/search?query=iasi	IASI L2 products - ESA	https://navigator.eumetsat.int/search?query=iasi
Volcano Space Observatory 5.1.3	Earth observation	Atmosphere	Aeris-Iasi	https://iasi.aeris-data.fr/	IASI L2 products - ULB-LATMOS	https://iasi.aeris-data.fr/
Volcano Space Observatory 5.1.3	Earth observation	Atmosphere	Copernicus	https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-1-sar	Sentinel-1 Single Look Complex products	https://peps.cnes.fr/rocket

Volcano Space Observatory 5.1.3	Earth observation	Atmosphere	CNES	https://pleiades.cnes.fr/fr/PLEIADES/Fr/GP_systeme.htm	Pléiades satellite optical imagery	
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	ARGO	http://www.argodatamgt.org/Access-to-data	Argo floats data and metadata from Global Data Assembly Centre (Argo GDAC)	https://www.seanoe.org/data/00311/42182/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	ARGO	http://www.argodatamgt.org/Access-to-data	ARGO GDAC (live)	ftp://ftp.ifremer.fr/ifremer/argo/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	ARGO	http://www.argodatamgt.org/Access-to-data	ARGO GDAC ERDDAP API	https://erddap.ifremer.fr/erddap/tabledap/ArgoFloats-synthetic-BGC.html
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	European Gliding Observatories (EGO)	https://www.ego-network.org/dokuwiki/doku.php	Ocean gliders : Data and metadata from Global Data Assembly Centre (OceanGliders GDAC)	https://www.seanoe.org/data/00453/56509/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	Marine Mammals Exploring the Oceans Pole to Pole (MEOP)	https://www.meop.net/database/meop-databases/	MEOP-CTD in-situ data collection: a Southern ocean Marine-mammals calibrated sea water temperatures and salinities observations	https://www.seanoe.org/data/00343/45461/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	Copernicus Marine and Environment Monitoring System (CMEMS)	https://marine.copernicus.eu/	Copernicus Marine In Situ - Global Ocean - Delayed Mode Biogeochemical product	https://data.marine.copernicus.eu/product/INSITU_GLO_BGC_DISCRETE_MY_013_046/description
Ocean Bio-Geochemical Observations	Ocean	Ocean observations	Copernicus Marine Service	https://marine.copernicus.eu/	Global Ocean Colour (Copernicus-GlobColour),	https://data.marine.copernicus.eu/product/OCEANCOLOUR_GLO_BGC_L3_MY_

5.2.1					Bio-Geo-Chemical, L3 (daily) from Satellite Observations (1997-ongoing)	009_103/description
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	Copernicus Marine Service	https://marine.copernicus.eu/	Global Ocean Colour Plankton and Reflectances MY L3 daily observations	https://data.marine.copernicus.eu/product/OCEANCOLOUR_GLO_BGC_L3_MY_009_107/description
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	Copernicus Marine Service	https://marine.copernicus.eu/	Global Ocean 3D Chlorophyll-a concentration, Particulate Backscattering coefficient and Particulate Organic Carbon	https://data.marine.copernicus.eu/product/MULTIOBS_GLO_BIO_BGC_3D_REP_015_010/description
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	Copernicus Marine Service	https://marine.copernicus.eu/	Nutrient profiles vertical distribution	https://data.marine.copernicus.eu/product/MULTIOBS_GLO_BIO_NUTRIENTS_PROFILES_REP_015_009/description
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	Copernicus Marine Service	https://marine.copernicus.eu/	Multi Observation Global Ocean 3D Temperature Salinity Height Geostrophic Current and MLD	https://data.marine.copernicus.eu/product/MULTIOBS_GLO_PHY_TSUV_3D_MYNRT_015_012/description
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	Copernicus Marine Service	https://marine.copernicus.eu/	Global Ocean Surface Carbon	https://data.marine.copernicus.eu/product/MULTIOBS_GLO_BIO_CARBON_SURFACE_REP_015_008/description
Ocean Bio-Geochemical Observations 5.2.1	Atmosphere/Land and Ocean	Meteorological fields	Copernicus Climate Data Store	https://cds.climate.copernicus.eu/cdsapp#!/home	ERA5 provides hourly estimates of a large number of atmospheric, land and oceanic climate variables.	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview

Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	EMODnet Chemistry - Black sea	https://www.emodnet-chemistry.eu/data	Black Sea - Eutrophication and Acidity aggregated datasets 1935/2020 v2021	https://sextant.ifremer.fr/record/b55f9e70-ce8e-4d7c-b6bf-bd0587e90bf9/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	EMODnet Chemistry - Arctic Ocean	https://www.emodnet-chemistry.eu/data	Arctic Ocean - Eutrophication and Acidity aggregated datasets 1923/2020 v2021	https://sextant.ifremer.fr/record/b57cd201-8fea-4caf-8efd-a9f6e76893c2/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	EMODnet Chemistry - North Sea	https://www.emodnet-chemistry.eu/data	North Sea - Eutrophication and Acidity aggregated datasets 1921/2020 v2021	https://sextant.ifremer.fr/record/62010039e-ed25-4cd9-92d1-ba7cc05256c8/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	EMODnet Chemistry - Mediterranean Sea	https://www.emodnet-chemistry.eu/data	Mediterranean Sea - Eutrophication and Acidity aggregated datasets 1911/2020 v2021	https://sextant.ifremer.fr/record/4e105ccd-46ea-48b3-b373-15028b677174/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	EMODnet Chemistry - North East Atlantic Ocean	https://www.emodnet-chemistry.eu/data	North East Atlantic Ocean - Eutrophication and Acidity aggregated datasets 1921/2020 v2021	https://sextant.ifremer.fr/record/a6d89ed2-17d0-4a8a-97fe-7e99d8e6520d/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	EMODnet Chemistry - Baltic Sea	https://www.emodnet-chemistry.eu/data	Baltic Sea - Eutrophication and Acidity aggregated datasets 1902/2020 v2021	https://sextant.ifremer.fr/record/eedd6334-dbc3-4621-82dc-1d6cb51f5c8c/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	Glodap (Global Ocean Data Analysis Project)	https://www.glodap.info	GLODAP / Biogeochemical parameters + carbonate	https://www.glodap.info/index.php/merged-and-adjusted-data-product-v2-

					variables	2022
Ocean Bio-Geochemical Observations 5.2.1	Atmosphere /Land and Ocean	Meteorological fields	NOAA - Physical Science Laboratory	https://psl.noaa.gov	NOAA-NCEP / Air.sig995, rhum.sig995, slp, sfc	https://psl.noaa.gov/thredds/catalog/Datasets/ncep.reanalysis/Dailies/surface/catalog.html
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	NOAA - Physical Science Laboratory	https://www.ncei.noaa.gov	NOAA-WOA / Temperature, Salinite, Percent Oxygen saturation, Nitrate, dissolved oxygen	https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Ocean observations	NASA - EarthData	https://www.earthdata.nasa.gov	SeaBass / HPLC, CHLA	https://seabass.gsfc.nasa.gov/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Bathymetry	General Bathymetric Chart of the Oceans (GEBCO)	https://www.gebco.net	GEBCO / Bathymetry	https://www.bodc.ac.uk/data/open_download/gebco/gebco_2022_sub_ice_topo/zip/
Marine Omics Observation 5.3.1	Nature and biodiversity	Biodiversity	World Register of Marine Species	https://www.marinespecies.org	Taxonomic classification data: species names and IDs, together with trait and other related information. We need access to the taxon match services: web services and/or the taxon match user-tool	https://www.marinespecies.org/aphia.php?p=webservice
Marine Omics Observation 5.3.1	Ocean	Ocean observations	European Marine Observation and Data Network (EMODnet)	https://emodnet.ec.europa.eu/en	Ecological time-series data: in first instance to just access all the metadata records for datasets that have a particular time	https://emodnet.ec.europa.eu/en/emodnet-web-service-documentation

					and place and provide those lists for users of the VRE to access themselves	
Marine Omics Observation 5.3.1	Ocean	Ocean observations	SeaDataNet	https://www.seadatanet.org	Physical oceanographic data measurements: in first instance to just access all the metadata records for datasets that have a particular time and place and provide those lists for users of the VRE to access themselves	https://www.seadatanet.org/Metadata/EDMED-Datasets
Marine Omics Observation 5.3.1	Marine Omics	Sequence data	National Center for Biotechnology Information (NCBI)	https://www.ncbi.nlm.nih.gov/taxonomy	NBCI taxonomic classification system:species names and their IDs	https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi
Marine Omics Observation 5.3.1	Nature and biodiversity	Biodiversity	Global Biodiversity Information Facility (GBIF)	https://www.gbif.org	Ecological time-series data: in first instance to just access all the metadata records for datasets that have a particular time and place and provide those lists for users of the VRE to access themselves	https://www.gbif.org/developer/summary
Marine Omics Observation 5.3.1	Marine omics	Sequence data	European Nucleotide Archive (ENA)	https://www.ebi.ac.uk/ena/browser/home	Raw sequence data (fastq files)	https://github.com/enasequence/enaBrowserTools
Marine Omics Observation 5.3.1	Marine omics	Biodiversity	EMBRC Marine Omics Biodiversity Observation	https://github.com/emo-bon	The field sampling data that are the source data for	

			Network		the metagoflow outputs and contain information that link species detections to a time and place (and parameters).	
Marine Omics Observation 5.3.1	Marine omics	Sequence data	European Marine Omics Biodiversity Observation Network (EMO BON)	https://github.com/emo-bon/sequencing-data	The links (via IDs) to the raw sequence data in the European Nucleotide Archive	
Marine Omics Observation 5.3.1	Marine omics	Sequence data	BiG-FAM database	https://bigfam.bioinformatics.nl/home	The Biosynthetic Gene Cluster Family (GCF) database	
Marine Omics Observation 5.3.1	Nature and biodiversity	Biodiversity	Ocean Biodiversity Information System (OBIS)	https://obis.org/	Ecological time-series data: in first instance to just access all the metadata records for datasets that have a particular time and place and provide those lists for users of the VRE to access themselves.	
Marine Omics Observation 5.3.1	Ontology	Ontology	MIxS: Minimum Information about any (x) Sequence	https://genomicsstandardsconsortium.github.io/mixs/	N/A	https://genomicsstandardsconsortium.github.io/mixs/
Marine Omics Observation 5.3.1	Ontology	Ontology	EML: Ecological Metadata Language	https://eml.ecoinformatics.org/	Metadata Vocabulary Specification	https://eml.ecoinformatics.org/
Marine Omics Observation 5.3.1	Ontology	Ontology	Dublin Core DCMI Metadata Terms	https://eml.ecoinformatics.org/	Metadata Vocabulary Specification	https://eml.ecoinformatics.org/
Marine Omics Observation	Ontology	Ontology	Darwin Core RDF	http://rs.tdwg.org/dwc/terms	Data Format Specification	http://rs.tdwg.org/dwc/terms

Funded by the European Union

5.3.1			Vocabulary			
Marine Omics Observation 5.3.1	Standard	Standard	Research Object Crates (RO-Crates)	https://www.researchobject.org/ro-crate	Data Format Specification	https://www.researchobject.org/ro-crate
Marine Omics Observation 5.3.1	Ontology	Ontology	Schema.org Vocabulary	https://schema.org/	Metadata Volcabulary Specification	https://schema.org/
Marine Omics Observation 5.3.1	Standard	Standard	JSON Linked Data Format Standard	https://json-ld.org/	Data Format Specification	https://json-ld.org/
Marine Omics Observation 5.3.1	Standard	Standard	Resource Description Framework (RDF)	https://www.w3.org/RDF	Data Format Specification	https://www.w3.org/RDF
Marine Omics Observation 5.3.1	Standard	Standard	MIBiG: Minimum Information about a Biosynthetic Gene cluster	https://mibig.secondarymetabolites.org	Data Format Specification	https://mibig.secondarymetabolites.org
Marine Omics Observation 5.3.1	Data Format	Data Format	BIOM: Biological Observation Matrix format	http://biom-format.org	Data Format Specification	http://biom-format.org
Marine Omics Observation 5.3.1	Data Format	Data Format	FASTA/FASTQ Sequence data exchange format	https://www.ncbi.nlm.nih.gov/genbank/fastaformat	Data Format Specification	https://www.ncbi.nlm.nih.gov/genbank/fastaformat
Marine Omics Observation 5.3.1	Repository	Repository	EMBL-EBI ENA API	https://www.ebi.ac.uk/ena/portal/api	Data Access API	https://www.ebi.ac.uk/ena/portal/api
Marine Omics Observation 5.3.1	Repository	Repository	EMBL-EBI BioSamples API	https://www.ebi.ac.uk/biosamples/docs/references/api	Data Access API	https://www.ebi.ac.uk/biosamples/docs/references/api
Marine Omics Observation 5.3.1	Repository	Repository	World Register of Marine Species taxonomy	https://www.marinespecies.org	Data Access API	https://www.marinespecies.org
Marine Omics Observation 5.3.1	Repository	Repository	Marine Regions gazetteer	https://www.marineregions.org/about.php	Data Access API	https://www.marineregions.org/about.php
Marine Omics Observation	Repository	Repository	NCBI taxonomy	https://www.ncbi.nlm.nih.gov/Taxonomy/Bro	Data Access API	https://www.ncbi.nlm.nih.gov/Taxonom

5.3.1			browser	wser/wwwtax.cgi		y/Browser/wwwtax.cgi

Table 2 - List of Software/Pipelines

PILOT	AREA	THEME	SOFTWARE	DESCRIPTION	ACCESSIBILITY	LINKS
Coastal Dynamics 5.1.1	Ocean	Data Analysis	DIVA ND	Interpolation of in situ measurements	downloadable	https://github.com/gher-uliege/DIVAnd.jl
Coastal Dynamics 5.1.1	Atmosphere/Ocean/Rivers	Multidisciplinary Data Analysis and Visualization with webODV	webODV	Web service for online multidisciplinary data analysis and visualization	web service	https://odv.awi.de/; https://explore.webodv.awi.de/
Coastal Dynamics 5.1.1	Ocean	Data Analysis	SOURCE	Calibration and validation of ocean models with in situ data - model vd obs intercomparison		https://doi.org/10.5281/zenodo.5008245
Earth Critical Zone 5.1.2	Land	Mapping Land Degradation by implementing SDG15.3	TRENDS EARTH	Quantification SDG15.3	web service	https://docs.trends.earth/en/latest/for_users/features/landdegradation.html
Earth Critical Zone 5.1.2	Land	Raster Algebra	rasdaman	Spatial domain and array operations on-the-fly within the query to get the input data for a model in a spatiotemporal subset.	web service	http://rasdev.landsupport.eu/rasdaman/ows#/services
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration, validation	CONTENT	Robust estimation of open ocean CO2 variables and nutrient concentrations from T, S and O2 data using Bayesian neural networks.	downloadable	https://github.com/HCBScienceProducts/CONTENT
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration, validation	CANYON-B	Robust estimation of open ocean CO2 variables and nutrient concentrations from T, S and O2 data using Bayesian neural networks.	downloadable	https://github.com/BRCScienceProducts/ESPER
Ocean Bio-Geochemical	Ocean	Calibration, validation	ESPER-NN	Empirical Seawater Property Estimation	downloadable	https://github.com/BRCScience

Funded by the European Union

Observations 5.2.1				Routines (ESPERs) capable of predicting seawater phosphate, nitrate, silicate, oxygen, total titration seawater alkalinity, total hydrogen scale pH (pHT)		Products/ESPER
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration, validation	LIR	Updated methods for global locally interpolated estimation of alkalinity, pH, and nitrate	downloadable	https://github.com/BRCScienceProducts/LIRs
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration, validation	SOCA-light	Routines capable to predict Ed490 and PAR profiles from T,S and satellite data	Publication in prep.	In preparation
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration	Xing2011.R	Function (R) to estimate F490, the ratio between Chlorophyll-A profile and Fluorescence profile as a function of the downwelling irradiance profile measured at 490nm	downloadable	https://doi.org/10.17882/86384
Ocean Bio-Geochemical Observations 5.2.1	Ocean	File export	DM-Filler	scripts and programs to fill BGC-Argo files in Delayed Mode	downloadable	https://github.com/catsch/DM_FILLER
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration	Bittig's tool	Estimate coefficients to adjust dissolved oxygen concentration		NA
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration	Jutard's tool	Estimate coefficients to adjust radiometry	downloadable	https://github.com/qjutard/radiometry_QC
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration, validation	CANYON-MED	A Regional Neural Network Approach to Estimate Water-Column Nutrient Concentrations and Carbonate System Variables in the Mediterranean Sea: CANYON-MED	downloadable	https://github.com/MarineFou/CANYON-MED/tree/master/v2

Funded by the European Union

Ocean Bio-Geochemical Observations 5.2.1		Calibration	Clim Tool	TO BE DEVELOP; ClimTool select colocated profiles and apply statistics in order to compare the work profile to already existing measures		To be develop
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration	xing2012	Quenching correction for in vivo chlorophyll fluorescence acquired by autonomous platforms	downloadable	https://github.com/catsch/STEP2_QUENCHING
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration	terrats2020	Revisiting the quenching correction for in vivo chlorophyll fluorescence acquired by autonomous platforms	downloadable	https://github.com/catsch/STEP2_QUENCHING
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration	errorTool	TO BE DEVELOPED - Routines to estimate error		To be develop
Ocean Bio-Geochemical Observations 5.2.1	Ocean	calibration, validation, qualification	statistic Tools	TO BE DEVELOPED - Routines to perform statistic evaluation		To be develop
Ocean Bio-Geochemical Observations 5.2.1	Ocean	Calibration	Locodox	oxygen measurement from in situ data or in air measurement	downloadable	https://github.com/euroargodev/LOCODOX
Ocean Bio-Geochemical Observations 5.2.1	Ocean	all services	ODV collection translator	need to be built in order to use webODV in the pilot		NA
Ocean Bio-Geochemical Observations 5.2.1	Ocean	mapping data set	cerbere	Python API for any type of spatio-temporal observations manipulation (load, normalize, write)	web service	https://cerbere.gitlab-pages.ifremer.fr/cerbere/
Ocean Bio-Geochemical Observations 5.2.1	Ocean	collocation	felyx	Python API for subsetting satellite/in-situ data		https://felyx.gitlab-pages.ifremer.fr/felyx_docs/

Funded by the European Union

Marine Omics Observation 5.3.2	Marine Omics	Data visualization	BiomeShiny	visualization of metagenomic data results; this is an R(shiny) application		https://github.com/BioData-PT/Biome-Shiny

Funded by the European Union

17.2 Annex B. Main Activities to Achieve the Deliverable

WORKSHOP 1. Kick off meeting Paris, 20 September 2022 - 22 September 2022

The workshop took place in September from 20 to 22 in Paris (France). The workshop brought together all the members of the WPs included in the FAIR-EASE project. WP5 members provided a description of the activities of each Pilot involved in the project. The workshop was structured into three days. In the first days, the project coordinator and manager and WP1 and WP6 leaders presented the general view of the project. In the second day, WP2, 3, 4, and 5 showed their main objectives. Specifically, WP5 presented the Pilots. Pilot reference persons were identified among members of the Pilots as the responsible for the reporting of each Pilot in WP5.

On the last day, all the participants carried out an open discussion on the project and the technical board was formed, including all the WP leaders plus the coordinators Alessandro Rizzo. More information about the workshop, including a summary, principles, and recommendations, is reported on the Confluence platform at the link https://FAIR-EASE.atlassian.net/wiki/spaces/FAIREASE/pages/2654317/220920-22+-+Kick-Off+Meeting+Paris.

The WP5 Pilot presentations are available at:

https://FAIR-EASE.atlassian.net/wiki/spaces/FAIREASE/pages/2654317/220920-22+-+Kick-Off+Meeting+Paris

and include:

- the presentation of the Use Case/Pilot (aims, scope and expected impact) suitable for a multidomain/interdisciplinary audience;

- description of the current state of the art: resources and data types employed, methods, expected or current results (with clear information on main standards and formats currently employed);

- opportunities, bottlenecks, expected improvements within the project;

- highlights on possible cross-domain links;

- issues to be considered for support to/ from the other WPs.

The goals of WP5 were presented as it follows:

Agreements on:

-resources and standards needed to face the challenges in the UCs;

-the computational strategies and requirements to feed WP2, WP3and WP4.

Testing and validating:

-The service implemented by WP2,WP3 andWP4;

-Data FAIRness (WP6).

Contribution to:

-dissemination, external user engagement (WP6);

-long term sustainability (WP1).

Exploitation of:

-trans-domain interoperability;

-synergies emerging from a transdisciplinary effort (T5.4).

WORKSHOP 2. Presentation of Pilot technical requirements. Ostende, 7 November 2022

WP5 Pilots demonstrated their expectations during the first day of the Ostende workshop, providing information about: aims and scope, development work, working flow, and datasets. The goal of the Ostend meeting was to create a shared understanding of the technical needs of the Pilots.

More information, including a summary of the workshop, recommendations, and the answers collected from the contributors, is reported on the Confluence platform.

WP5 internal meetings

19 October 2022, (4:00 - 6:30 pm, 2:30 hours), 19 participants

Main output: Pilots begin to complete the expectations documents before completing the template to clarify specifics regarding the resources and methods to be used, improved upon, or implemented. We presume that if expectations refer to resources or methods to be used, improved upon, or implemented, specifics will be included in the template accordingly.

03 November 2022, (14:00 - 16:00 pm, 2 hours), 15 participants

Main output: Organization of the Pilots' presentations for the Ostende workshop.

15 November 2022, (3:15 - 5:15 pm, 2 hours), 13 participants

Main output: internal brainstorming on key requirements tweaked after the Ostende meeting.

29 November 2022, (2:00 - 4:00 pm, 2 hours), 12 participants

Main output: WP5 leader proposed a precise scheme to present Pilot workflows and expectations, that was discussed during the meeting. The scheme required the following information: 1) data access; 2) data format; 3) M=Methods/S=Secondary resources; 4) type of data; and 5) primary data resource.

One to One Meetings (technical WPs 2, 3, 4 versus each Pilot)

The scope of the One-to- One meetings was to have a dedicated discussion per each Pilot with the technical WPs, with the following topics:

– Pilot needs:

– size of dataset (MB/GB/TB);

- data access requirements;
- expectation from the Earth Analytic Laboratory;
- expectation from the data lake;
- expectation from DDAS
- Resources and methodology mapping
- Pilot workflows and diagrams of the technical architecture

12 December 2022 Volcano Space Observatory Pilot

(link to the conference page: https://FAIR-EASE.atlassian.net/wiki/spaces/FAIREASE/pages/30146589/2022-12-12+Meeting+notes+-+FAIR-EASE+Technical+WPs+Individual+Pilots+1st+sessions+Volcano+Space+Observatory+Pilot).

15/12/2022 Coastal Water Dynamics Pilot

(link to the conference page: https://FAIR-EASE.atlassian.net/wiki/spaces/FAIREASE/pages/30310401/2022-12-15+Meeting+notes+-+FAIR-EASE+Technical+WPs+Individual+Pilots+1st+sessions+Coastal+Water+Dynamics+Pilot).

15 December 2022 Marine Omics Observation Pilot

(link to the conference page: https://FAIR-EASE.atlassian.net/wiki/spaces/FAIREASE/pages/30375937/2022-12-15+Meeting+notes+-+FAIR-EASE+Technical+WPs+Individual+Pilots+1st+sessions+Marine+Omics+Observation+Pilot).

16 December 2022 Earth Critical Zone Pilot

(link to the conference page: https://FAIR-EASE.atlassian.net/wiki/spaces/FAIREASE/pages/30310415/2022-12-16+Meeting+notes+-+FAIR-EASE+Technical+WPs+Individual+Pilots+1st+sessions+ECZ+planning+and+management+.Pilot).

16 December 2022 Ocean BGC Observation Pilot

(link to the conference page: https://FAIR-EASE.atlassian.net/wiki/spaces/FAIREASE/pages/30310429/2022-12-16+Meeting+notes+-+FAIR-EASE+Technical+WPs+Individual+Pilots+1st+sessions+Ocean+BGC+Observation+Pilot).