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	fairease.eu

D4.2 - Landscaping exercise: the inclusion of special use-case datasets in the data lake

Work Package	**WP4, Interoperability, Integration, and Supporting Services**
Lead Author (Org)	**Nydia Catalina Reyes Suarez (OGS)** **Mark Portier (VLIZ)**
Contributing Author(s) (Org)	**Alessandra Giorgetti (OGS)** **Dataset selection** **- Pilot 5.1.1: Reiner Schlitzer (AWI)** **- Pilot 5.1.2: Giuliano Langella (UNINA)** **- Pilot 5.1.3: Marie Boichu (ULille) & Vincent Breton (CNRS)** **- Pilot 5.2.1: Virgine Racapé (POKAPOK)** **- Pilot 5.3.1: Cymon J. Cox (CCMAR)**
Due Date	**30.04.2023**
Date	**05.05.2023**
Version	**V1.0**

Dissemination Level

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

Funded by
the European Union

Versioning and contribution history

Version	Date	Author	Notes
0.1	21.12.2022	Catalina Reyes (OGS)	First draft
0.2	24.03.2023	Catalina Reyes (OGS)	Update after the plan proposed in the monthly meeting and the discussion with the coastal dynamics and the earth critical zone pilots.
0.3	06.04.2023	Catalina Reyes (OGS)	Draft version
0.4	11.04.2023	Catalina Reyes (OGS) Marc Portier (VLIZ)	Chapter 3 update
0.5	14.04.2023	Reiner Schlitzer (AWI) Catalina Reyes (OGS) Marc Portier (VLIZ)	Update after WP4 April meeting
0.6	24.04.2023	Virgine Racapé (POKAPOK) Reiner Schlitzer (AWI) Catalina Reyes (OGS) Marc Portier (VLIZ) Reviewers: Clément Weber (POKAPOK) Laurian Van Maldeghem (VLIZ) Potirakis Antonis (HCMR) Stelios Ninidakis (HCMR)	Revision by the pilots Final version for reviewers
1.0	05.05.2023	Catalina Reyes (OGS) Marc Portier (VLIZ)	Submission variant

Disclaimer

This document contains information which is proprietary to the FAIR-EASE Consortium. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to a third party, in whole or parts, except with the prior consent of the FAIR-EASE Consortium.

Funded by
the European Union

Table of Contents

Funded by
the European Union

Funded by
the European Union

List of Figures

List of Tables

TERMINOLOGY

Terminology/Acronym	Description
BGC	BioGeoChemical
BON	Biodiversity Observation Network
CC-BY	Creative Commons Attribution Licence (https://creativecommons.org/licenses/by/4.0/)
CDI	Common Data Index
CMEMS	Copernicus Marine Service
D	Deliverable
DCAT	Data Catalogue Vocabulary (https://www.w3.org/TR/vocab-dcat-2/)
DDAS	Data Discovery and Access Service
DI	Data Infrastructure
DoA	Description of Action
DOI	Digital Object Identifier
EBVs	Essential Biological Variables
EC	European Commission
ECMWF	European Centre for Medium-Range Weather Forecasts
ECZ	Earth Critical Zone
EMBRC	European Marine Biological Resource Centre
EOSC	European Open Science Cloud
EMO-BON	European Marine Omics Biodiversity Observation Network
EMODnet	European Marine Observation and Data Network
ENA	European Nucleotide Archive
EOV	Essential Ocean Variables
FAIR	Findable; Accessible; Interoperable; Reusable

Funded by
the European Union

Terminology/Acronym	Description
FE	FAIR EASE
GA	Grant Agreement to the project
GRDC	Global Runoff Data Centre
KPI	Key Performance Indicator
LDES	Linked Data Event Streams (https://w3id.org/ldes/specification)
LAI	Leaf Area Index
LC	Leaf Cover
LCC	Leaf Chlorophyll Content
M2M	Machine to machine
MSFD	Marine Strategy Framework Directive
MSP	Marine Spatial Planning
ODV	Ocean Data View
SAFE	Standard Archive Format for Europe
SPARQL	SPARQL Protocol and RDF Query Language
TROPOMI	Tropospheric Monitoring Instrument
UC	Use Case
VDAP	Virtual Data Analysis Platform
VRE	Virtual Research Environment
WP	Work Package

Funded by
the European Union

Executive Summary

This document describes the landscaping exercise proposed for deliverable 4.2 (D4.2) within Work Package (WP) 4 of the FAIR Earth Sciences & Environment services project (FAIR-EASE, FE). The goal of this exercise is to analyse different special use case (UC) datasets per pilot and the requirements they must meet to be included in the data lake infrastructure proposed in D4.1 (landscaping exercise: the (meta)data, software, and cloud needs for the data lake). The pilots per UC are:

- *UC1 - Earth and Environmental Dynamics: Coastal waters dynamics (Pilot 5.1.1), Earth Critical zones observatory (Pilot 5.1.2), and Volcano Space Observatory (Pilot 5.1.3),*
- *UC2 - Environmental Bio-geochemical Assets: Ocean Bio-Geo-Chemical Observatory (Pilot 5.2.1) and,*
- *UC3 - Biodiversity Observation: Marine Omics Observatory (Pilot 5.3.1).*

Datasets from each pilot were selected from Table 1 in Annex A of D5.1 (report on key requirements from Use Cases and Pilots, [1]) and with a prior selection from a representative from each pilot. These datasets were selected to cover as much diversity as possible and to reflect the multidisciplinary nature of each UC. The deliverable aims to analyse and highlight the criticalities of the selected datasets considering their current limitations and needs and how they could fit into the "data provider" view proposed for the "data lakes" architecture in D4.1.

Bear in mind that the datasets described should not be taken as the only source for the data lake ingestion. As stated, before a few special UCs datasets were selected using the minimum selection and maximum diversity criteria, to analyse the requirements for them to be ingested in the data lakes proposed.

Funded by
the European Union

1 Introduction

Within the FE project, WP4 is dedicated to setting up the “data lake” that will allow efficient access to local and distributed data by the Virtual Research Environments (VREs) and Virtual Data Analysis Platforms (VDAPs) developed in WP3. These instruments will be used and tested by the pilot UCs (WP5) and by the data access portal (WP2).

D4.2: *Landscaping exercise; the inclusion of special use-case datasets in the data lake* aims to describe and analyse a few selected UC datasets allowing for the minimum selection and maximum diversity criteria, the idea behind it was to cover as many possibilities of data access as possible (diversity) with 1 or 2 datasets listed by the pilots and that at the same time are of prior interest. Subsequently, it will be checked if these datasets meet the requirements necessary to be ingested into the architecture proposed in D4.1. The information about the pilots’ needs and the list of datasets per pilot has been gathered in D5.1 [1], while D2.1 provided the most important data infrastructures used by the pilots and their current state. These last two deliverables, together with D4.1 (where a description of the data lake architecture was proposed), were essential for the selection of the datasets and to assess the requirements which were already presented in D4.1. Thus this document is organised as follows:

- **Special Use Case datasets:** In this section a description of the selected UCs datasets and their data access will be given. This is a result of both D5.1 and D2.1, as well as internal meetings/email exchanges with a representative of each pilot.

- **Assessment of UC dataset requirements:** This section will focus on a thought exercise where the selected datasets are going to be analysed keeping into account the data lake architecture proposed. It will be aimed to list the requirements for the selected datasets in order to be included in conformity with the proposed architecture and their current limitations.

1.1 Dataset selection

It was asked the referents for each pilot to provide two datasets they consider essential and with a certain level of criticality in terms of data access. Based on these needs, and taking into account the data infrastructure the datasets belong to, as also reported in D2.1, the following “special” datasets were selected for each pilot amongst the ones mapped in Table 1 of Annex A of D5.1 [1] available at:
https://docs.google.com/spreadsheets/d/18uRj4MFaCDJPcbcuscy1KdTG95h_imhj/edit#gid=1818952351.

UC1 - Earth and Environmental Dynamics:

- *Coastal waters dynamics (Pilot 5.1.1):* daily river runoff and eutrophication and acidity aggregated datasets.
- *Earth Critical Zones observatory (Pilot 5.1.2):* satellite images from sentinel II

- *Volcano Space Observatory (Pilot 5.1.3):* Sentinel-5P/TROPOMI SO2 data and Sentinel-1 SLC data.

UC2 - Environmental Bio-geochemical Assets:

- Ocean Bio-Geo-Chemical Observatory (Pilot 5.2.1): BGC ARGO datasets and ERA5 reanalysis.

UC3 - Biodiversity Observation:

- Marine Omics Observatory (Pilot 5.3.1): Taxonomic inventories and community gene function profiles - Ro-create datasets and Triple Store / SPARQL endpoint.

2 Special Use Case datasets

The FE project seeks to engage the Earth and environmental science communities by developing three multidisciplinary UC that contribute to the component requirements of the FE system and help validate and demonstrate its capabilities to support open science by identifying the needs from real-world scientific challenges. These communities will be key players in the development of the FE tools and will contribute to the improvement of preoperational services. The following sections describe the special UC datasets selected for each UC and the pilots of the project.

2.1 UC1 - Earth and Environmental Dynamics

Three pilots cover this Use Case, each one set in a different terrestrial environment: Coastal Waters Dynamics, Earth Critical Zones, and Volcano Space Observatory. These pilots are based on integrated data from water, soil, air, and biotic resources and have different aims but comparable challenges, including multidisciplinary data integration, monitoring and modelling of dynamic spatio-temporal data [2].

2.1.1 Coastal waters dynamics (Pilot 5.1.1)

In the coastal marine environment, near river estuaries, important processes such as the evolution of plankton blooms or the transport and fate of nutrients, carbon, and contaminants take place and depend critically on many factors, including river discharge, ocean circulation, meteorological conditions and biogeochemical processes in the water column. This Pilot focuses on the region where the Po River discharges into the northern Adriatic Sea, and where these processes play a key role in fish stocks and have an effect on the quality of the ecosystem of coastal waters which in turn impacts socio-economic activities, for the whole North Adriatic Sea [2].

Funded by
the European Union

2.1.1.1 Daily River runoff

2.1.1.1.1 *Dataset description*

The Ocean Data View (ODV) *GRDC_Daily_River_Runoff_v2021* collection was produced from original daily runoff files downloaded from the Global Runoff Data Centre (GRDC, https://www.bafg.de/GRDC) in March 2021 by Reiner Schlitzer from the Alfred Wegener Institute. The GRDC is an international archive of river discharge data reaching back in some cases up to 200 years and fosters multinational and global long-term hydrological studies. The database of quality controlled "historical" mean daily and monthly discharge data grows steadily and currently comprises river discharge data from more than 9,900 monitoring stations from 159 countries (Figure 1).

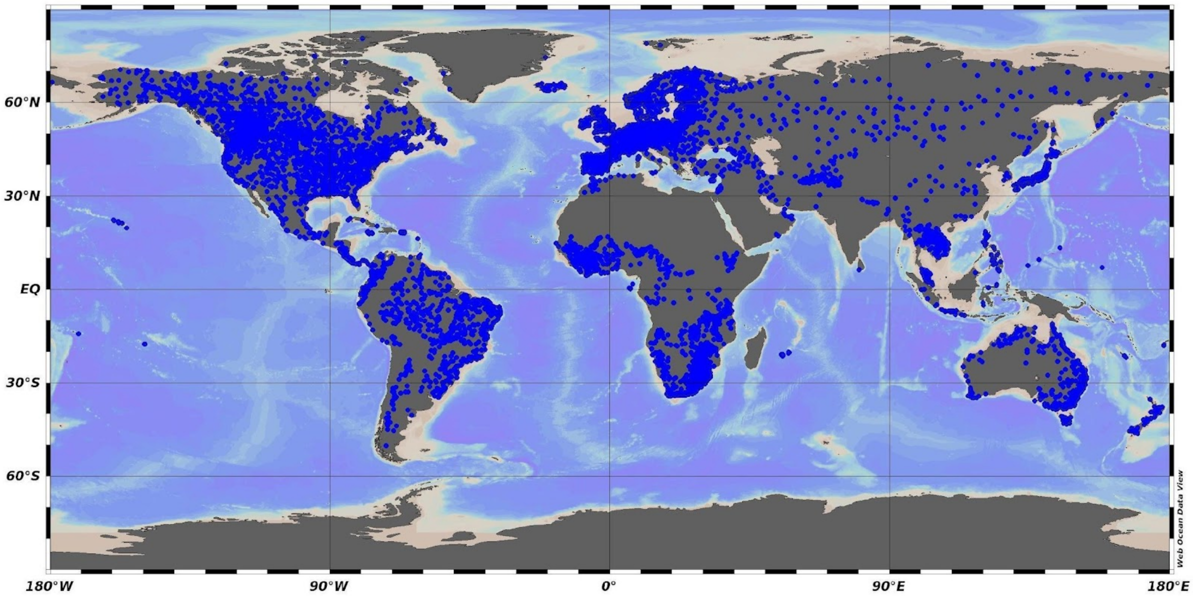


Figure 1 – *GRDC_Daily_River_Runoff_v2021* ODV collection [3]. Blue dots represent the stations containing river discharge data.

2.1.1.1.2 *Terms of use*

All hydrological data offered on the GRDC website are managed by GRDC with permission of the data owners, usually the National Hydrological Services who created the data. All hydrological data remains the property of the owner. The data can be downloaded from the GRDC website and may be used free of charge for research purposes. Commercial use and redistribution of the river discharge data to third parties, either in part or total, is not allowed without the written consent of the GRDC.

In the present case, GRDC granted permission to the Alfred Wegener Institute (AWI), allowing access to the entire GRDC data holding and allowing the creation of a harmonised, aggregated, global river discharge dataset as ODV collection. According to the agreement, this data collection is made available to users as an online data resource via the webODV Explore website https://explore.webodv.awi.de/. The direct link to the dataset is https://explore.webodv.awi.de/rivers/discharge/grdc/daily/grdc_daily_river_runoff_v2021/.

Funded by
the European Union

Users have to agree to the usage agreement and can then interactively analyse and visualise the runoff data in various ways using the ODV-online user interface in their web browser. Users may download and publish the created graphics, but export and direct download of the original data values is disabled, in compliance with the special agreement between GRDC and AWI. Access to the original runoff data is possible via the GRDC website https://www.bafg.de/GRDC/.

2.1.1.1.3 *Link to dataset and other relevant information*

- **Dataset name:** Daily River discharge data for 8079 stations worldwide from the Global Runoff Data Centre (GRDC) - version 2021.
- **URL provided:** https://explore.webodv.awi.de/rivers/discharge/grdc/daily/
- **Dataset type:** Time series
- **Data format:** ODV Collection[1].
- **Access:** Public.
- **M2M access:** No. The terms of use constrain the exploration of the data to actual end-user interactions. The procedure to ask for written consent is a mere formality but requires human intervention and dialogue.
- **Alternative access to data:** https://www.bafg.de/GRDC/EN/Home/homepage_node.html

2.1.1.2 Eutrophication and Acidity data collection

2.1.1.2.1 *Dataset description*

Eutrophication and acidity data collections are available through EMODnet Chemistry's data infrastructure (DI). This DI stores more than 600,000 datasets and common data index (CDI) metadata for eutrophication. The datasets are managed by tens of data centres, and a robot harvesting system allows all datasets within a configured query filter to be automatically retrieved from connected data centres [4]. The number of records per group of variables is shown in Table 1.

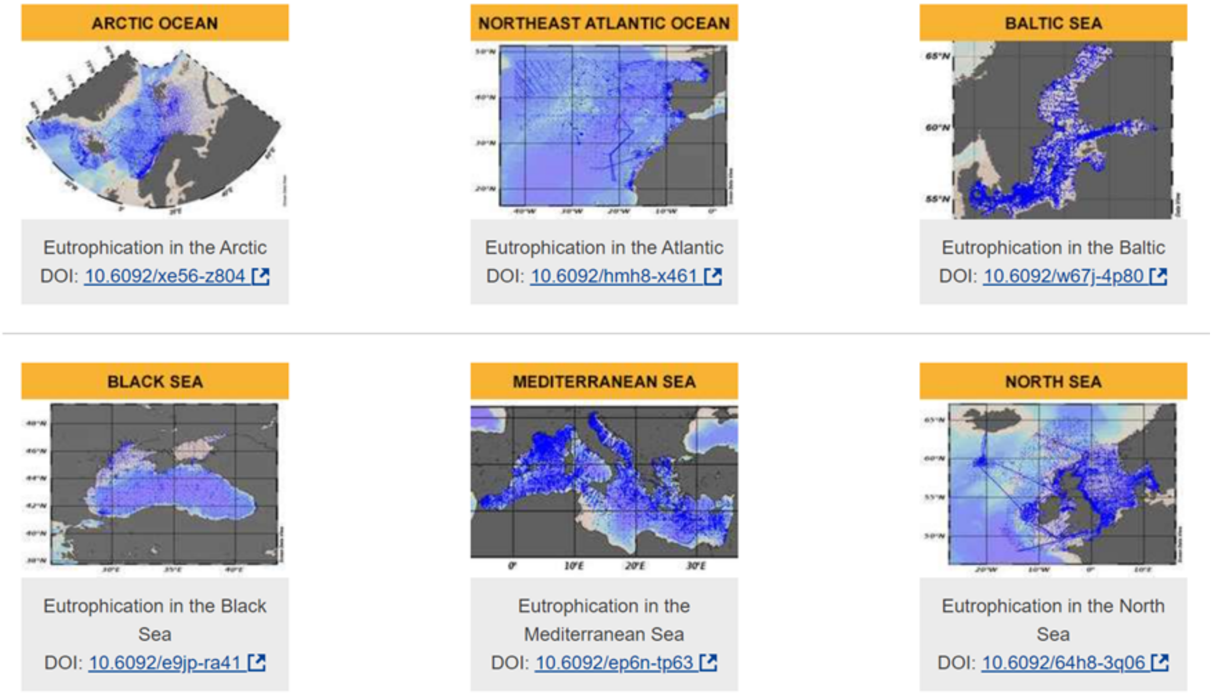
Table 1 - EMODnet Chemistry Eutrophication variables.

Group of variable	*No. of records*	*Examples of used parameters*
Chlorophyll	346712	Chlorophyll pigment concentration.
Dissolved gases	746722	Dissolved oxygen parameters.
Fertilisers	517036	Ammonium, phosphate, nutrient concentration parameters.
Organic matter	64767	Carbon concentration
Silicates	377457	Silicate concentration parameters in the water column.

[1] The ODV data format allows dense storage and very fast data access. Large data collections with millions of stations can easily be maintained and explored on inexpensive desktop and notebook computers. More information at: https://www.bodc.ac.uk/resources/delivery_formats/odv_format/

Funded by
the European Union

Data collections of standardised, harmonised, and validated datasets are available for six sea regions, as shown in Figure 2. For the purpose of D4.2, only Mediterranean Sea eutrophication and acidity aggregated datasets will be taken into account since the other sea regions have the same type of data access.



ARCTIC OCEAN

Eutrophication in the Arctic
DOI: 10.6092/xe56-z804

NORTHEAST ATLANTIC OCEAN

Eutrophication in the Atlantic
DOI: 10.6092/hmh8-x461

BALTIC SEA

Eutrophication in the Baltic
DOI: 10.6092/w67j-4p80

BLACK SEA

Eutrophication in the Black Sea
DOI: 10.6092/e9jp-ra41

MEDITERRANEAN SEA

Eutrophication in the Mediterranean Sea
DOI: 10.6092/ep6n-tp63

NORTH SEA

Eutrophication in the North Sea
DOI: 10.6092/64h8-3q06

Figure 2 - EMODnet chemistry eutrophication data collections.

Each regional data collection is accessible through a digital object identifier (DOI) or using webODV: https://emodnet-chemistry.webodv.awi.de/eutrophication%3EMediterranean. The aggregated data are available in the ODV format, which is composed of a metadata header followed by tab-separated values. The data collection can be easily handled and visualised using ODV Software (https://odv.awi.de/). Parameter names are based on P35, EMODnet Chemistry aggregated parameter names vocabulary, which is available at https://vocab.seadatanet.org/v_bodc_vocab_v2/search.asp?lib=P35.

Regional datasets concerning eutrophication and acidity are automatically harvested and resulting collections are aggregated and quality-controlled using ODV Software and following a common methodology for all Sea Regions [5]. The QC and data formats for this dataset is detailed in EMODnet chemistry eutrophication guidelines [6].

The original datasets can be searched and downloaded from EMODnet Chemistry CDI Data and Discovery Access Service: https://emodnet-chemistry.maris.nl/search. Data products are accessible on the Central Portal through the EMODnet Map viewer service and the EMODnet Catalogue Service, where users can also retrieve the full set of metadata, originators, and distributors.

Funded by
the European Union

2.1.1.2.2 *Terms of Use*

The products can be freely used for non-commercial and educational purposes. However, the data in the maps have been homogenised and filtered to allow comparisons between countries. Therefore, EMODnet Chemistry's visualisation products may not be comparable to source data accessible through other platforms.

EMODnet Chemistry is not responsible for the use of the data and products by third parties. Users are requested to give due credit to the EMODnet Chemistry project and the data originators.

2.1.1.2.3 *Link to dataset and other relevant information*

- **Dataset name:** Mediterranean Sea - Eutrophication and Acidity aggregated datasets 1911/2020 v2021.
- **URL provided:** https://sextant.ifremer.fr/documentation/emodnet_chemistry/api/catalogue.html#/metadata/4e105ccd-46ea-48b3-b373-15028b677174
- **Dataset type:** Time series and vertical profiles.
- **Data format:** ODV Collection
- **Access:** Public.
- **M2M access:** No
- **Alternative access to data:** https://emodnet-chemistry.webodv.awi.de/eutrophication%3EMediterranean

2.1.2 Earth Critical zones observatory (Pilot 5.1.2)

This Pilot focuses on analysing and visualising Earth Critical Zone (ECZ) information required for 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. The main interest of this Pilot will be to implement connections between data resources and analytical strategies in different environmental frameworks. This is the case for coastal erosion which affects the ECZ and coastal groundwater dynamics and soil salinization [2].

2.1.2.1 Satellite imagery from sentinel II

2.1.2.1.1 Dataset description

The Copernicus SENTINEL-2 mission comprises a constellation of two polar-orbiting satellites placed in the same sun-synchronous orbit, phased at 180° to each other. It aims at monitoring variability in land surface conditions, and its wide swath width (290 km) and high revisit time will support the monitoring of Earth's surface changes. Observation data acquired from the SENTINEL-2 mission are used by services such as

- Land monitoring: SENTINEL-2 makes a significant contribution to land monitoring services by providing input data for both land cover and land cover change mapping, and supporting the assessment of biogeophysical parameters such as Leaf Area Index (LAI), Leaf Chlorophyll Content (LCC) and Leaf Cover (LC) [12].

- Emergency management: The 10 m spatial resolution bands of SENTINEL-2 and the high revisit time of the mission supports the rapid acquisition and delivery of images to support disaster relief efforts. This includes mapping of urban areas, including at-threat buildings and complex structures, that have been previously identified as being at risk from natural hazards such as earthquakes and flooding [12].

- Security: In tandem with communication and global positioning (GPS) technologies, SENTINEL-2 supports the following security domains: border surveillance, maritime surveillance, and support to EU external actions [12].

- Climate change: SENTINEL-2 mission supports the attempts to mitigate deforestation by providing greater opportunities to acquire cloud-free image data. This will be of particular benefit in the tropical latitudes, where heavy cloud cover has previously delayed the potential acquisition of a complete catalogue of data.

- Marine: The Copernicus Marine Environment Monitoring Service (CMEMS) uses SENTINEL-2 data to provide high-resolution Ocean Colour products. The main goal is to support the Marine Strategy Framework Directive (MSFD) and Marine Spatial Planning (MSP) Directive, and Regional Seas Conventions (e.g. Barcelona Convention/Mediterranean). In addition, SENTINEL-2 imagery is used as a key source of information for the periodic validation of products derived from SENTINEL-1. In particular for
    - Iceberg detection and monitoring
    - Arctic sea routes monitoring

The SENTINEL-2 mission systematically acquires data over land and coastal areas in a band of latitude extending from 56° South (Isla Hornos, Cape Horn, South America) to 82.8° North (above Greenland):

- all coastal waters up to 20 km from the shore
- all islands greater than 100 km2
- all EU islands
- the Mediterranean Sea
- all closed seas (e.g. Caspian Sea).

Funded by
the European Union

SENTINEL-2 products available for Users are listed in the product types tables available at: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2/data-products [7].

2.1.2.1.2 Terms of use

The free, full and open data policy adopted for the Copernicus programme foresees access available to all users for the Sentinel data products. The Copernicus Data Hub distribution service will continue its full operations until the end of June 2023 to allow a smooth migration to the new Copernicus Data Space Ecosystem by all user communities. From July 2023 until September 2023, the Copernicus Data Hub distribution service will continue offering access to Sentinel data with a gradual ramp-down of the operations capacity and data offering.

The access and use of Copernicus Sentinel Data and Service Information are regulated under EU law.1 In particular, the law provides that users shall have free, full and open access to Copernicus Sentinel Data2 and Service Information without any express or implied warranty, including as regards quality and suitability for any purpose. EU law grants free access to Copernicus Sentinel Data and Service Information for the purpose of the following use in so far as it is lawful:

(a) reproduction;
(b) distribution;
(c) communication to the public;
(d) adaptation, modification and combination with other data and information;
(e) any combination of points (a) to (d).

EU law allows for specific limitations of access and use in the rare cases of security concerns, protection of third-party rights or risk of service disruption.

By using Sentinel Data or Service Information the users acknowledge that these conditions are applicable to them and that they renounce any claims for damages against the European Union and the providers of the said Data and Information. The scope of this waiver encompasses any dispute, including contracts and torts claims, that might be filed in court, in arbitration or in any other form of dispute settlement [8].

2.1.2.1.3 Link to dataset and other relevant information

- **Dataset name:** Satellite imagery from sentinel II
- **URL provided:** https://scihub.copernicus.eu/
- **Dataset type:** Images
- **Data format:** SENTINEL-SAFE format[2]

[2] The SENTINEL-SAFE format wraps a folder containing image data in a binary data format and product metadata in XML. This flexibility allows the format to be scalable enough to represent all levels of SENTINEL products.More information about the format here.

Funded by
the European Union

- **Access:** Public
- **M2M access:** Yes, with authentication.
- **Alternative access to data:**
    - https://scihub.copernicus.eu/dhus/#/home
    - API hub: https://scihub.copernicus.eu/twiki/do/view/SciHubWebPortal/APIHubDescription, https://documentation.dataspace.copernicus.eu/#/APIs/OpenSearch or https://documentation.dataspace.copernicus.eu/#/APIs/OData
    - https://scihub.copernicus.eu/gnss/#/home

2.1.3 Volcano Space Observatory (Pilot 5.1.3)

This Pilot focuses on the need of a web interface capable of interactively aggregating satellite observations from the Solid Earth and Atmospheric communities for the benefit of volcanic observatories and scientists. Supporting the implementation of innovative web services developed in the project, the interest is to design the implementation of an interface displaying a range of complementary data relevant to the near real-time monitoring of volcanic activity [2].

2.1.3.1 Sentinel-5P/TROPOMI SO2

2.1.3.1.1 Dataset description

The Tropospheric Monitoring Instrument (TROPOMI) onboard Sentinel-5 Precursor is a nadir-viewing, imaging spectrometer covering wavelength bands between the ultraviolet and the shortwave infrared. The instrument uses passive remote sensing techniques to attain its objective by measuring, at the Top Of Atmosphere (TOA), the solar radiation reflected by and radiated from the earth.

Sulphur dioxide (SO2) enters the Earth's atmosphere through both natural (~30%) and anthropogenic processes (~70%). It plays a role in chemistry on a local and global scale and its impact ranges from short-term pollution to effects on climate. Besides the total column of SO2, enhanced levels of SO2 are flagged within the products. The recognition of enhanced SO2 values is essential in order to detect and monitor volcanic eruptions and anthropogenic pollution sources. Volcanic SO2 emissions may also pose a threat to aviation, along with volcanic ash.

The S5p sensor TROPOMI samples the Earth’s surface with a revisit time of one day and with an unprecedented spatial resolution of 7.0x3.5 km2, respectively 5.5×3.5km2 (since 6th of August 2019). This allows the resolution of fine details and S5p to arguably be a valuable tool to better study anthropogenic SO2 emissions but also volcanic emissions, from degassing to eruptive processes (Figure 3).

The retrieval of SO2 vertical column is performed in near-real time (i.e. typically 3 hours after measurement) based on the DOAS technique, involving two main steps: First, the effective slant column amount Ns (corresponding to the integrated SO2 concentration along the mean atmospheric optical path) is derived through a least-squares fit of the measured Earth reflectance spectrum to laboratory absorption cross-sections. Second, slant columns are converted into vertical columns by means of air mass factors (AMF) obtained from suitable radiative transfer calculations, accounting for the presence of clouds and aerosols, surface properties and best-guess SO2 vertical profiles [9].

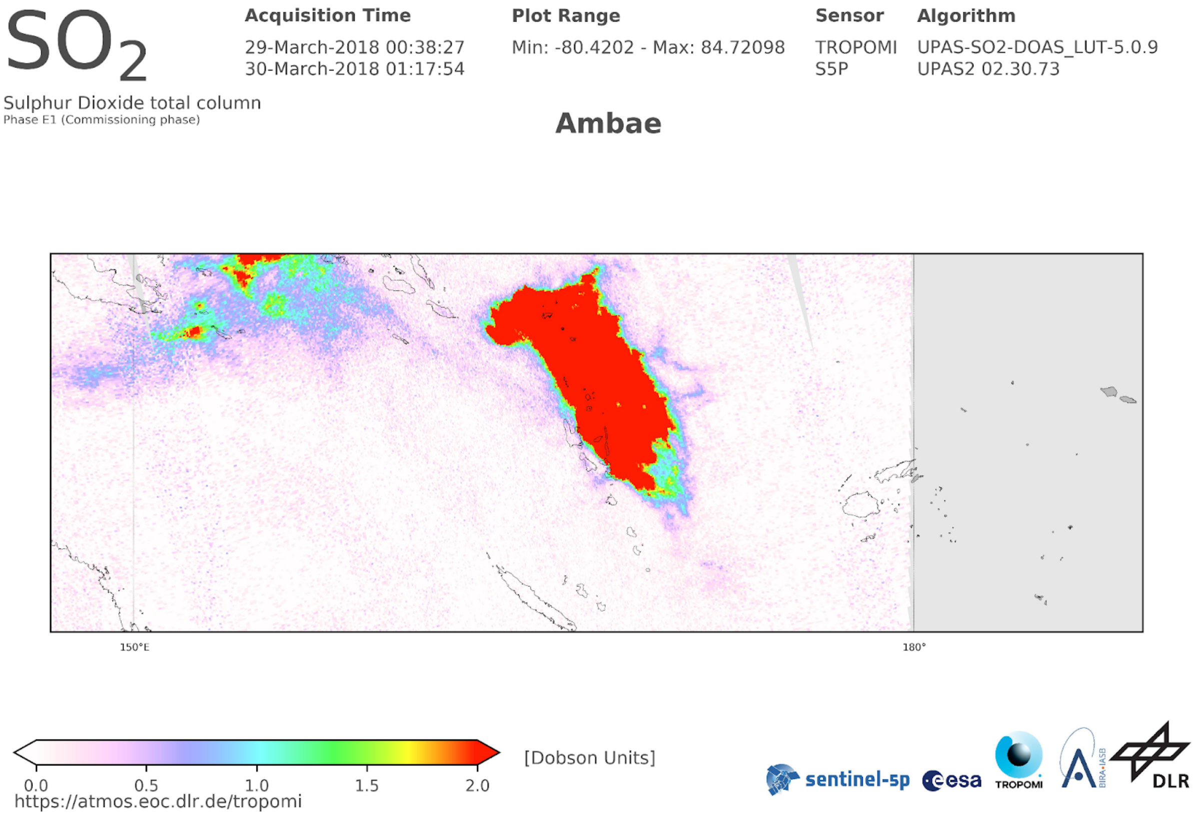


Figure 3 - The eruption of Ambae Volcano acquired on 29th March 2018.

2.1.3.1.2 Terms of Use

The TROPOMI Sulphur Dioxide SO2 product data are available from the Copernicus Open Data Hub https://scihub.copernicus.eu. The access and use of any Copernicus Sentinel data available through the Sentinel Data Hub are governed by the Legal Notice on the use of Copernicus Sentinel Data and Service Information and is given here: https://sentinels.copernicus.eu/documents/247904/690755/Sentinel_Data_Legal_Notice.

2.1.3.1.3 Link to dataset and other relevant information

- **Dataset name:** Sentinel-5 Precursor Level 2 Sulphur Dioxide (L2__SO2____)
- **URL provided:** https://s5phub.copernicus.eu/dhus/#/home
- **Dataset type:** Orbit data

Funded by
the European Union

- **Data format:** NetCDF
- **Access:** Public
- **M2M access:** Yes, with authentication.
- **Alternative access to data:**
  - https://download.geoservice.dlr.de/S5P_TROPOMI/files/
  - https://scihub.copernicus.eu/dhus/#/home
  - API hub: https://scihub.copernicus.eu/twiki/do/view/SciHubWebPortal/APIHubDescription, https://documentation.dataspace.copernicus.eu/#/APIs/OpenSearch or https://documentation.dataspace.copernicus.eu/#/APIs/OData

2.1.3.2 Sentinel-1 SLC data

Sentinel-1 is a Synthetic Aperture Radar (SAR) mission. The Sentinel-1 mission comprises a constellation of two polar-orbiting satellites, operating day and night and performing C-band synthetic aperture radar imaging, enabling them to acquire imagery regardless of the weather. Sentinel-1 will work in a pre-programmed operation mode to avoid conflicts and to produce a consistent long-term data archive built for applications based on long time series.

Main applications include: Monitoring sea ice and icebergs, monitoring land ice (glaciers, ice sheets, ice caps) river and lake ice, monitoring oil spills and ships, marine winds and waves, land-use change, agriculture, deforestation, land deformation and support to emergency management such as floods and earthquakes.

Level-1 Single Look Complex (SLC) products consist of focused SAR data geo-referenced using orbit and altitude data from the satellite and provided in zero-Doppler slant-range geometry. The products include a single look in each dimension using the full transmit signal bandwidth and consist of complex samples preserving the phase information.

2.1.3.2.1 Dataset description

Sentinel data products are made available systematically and free of charge to all data users including the general public, scientific and commercial users. All data products are distributed in the Sentinel-Standard Archive Format for Europe (SAFE) format as in section 2.1.2.1.2.

The PEPS platform, the French "mirror site", redistributes the products of Sentinel satellites, S1A, S1B, S2A and S2B from COPERNICUS, the European system for Earth monitoring. This program is coordinated and managed by the European Commission. Access to the data is open and free. It concerns all products except level 0 (level-0 product is compressed raw image data in Instrument Source Packet (ISP) format). These products were generated by ESA.

Funded by
the European Union

2.1.3.2.2 Terms of use

2.1.3.2.3 Link to dataset and other relevant information

- **Dataset name:** Sentinel-1 SLC data
- **URL provided:** https://peps.cnes.fr/rocket/#/home
- **Dataset type:** SAR image. One image per burst and per sub-swath. 8 bursts per sub-swath, 3 sub-swaths. At least 2 products are needed to compute an interferogram.
- **Data format:** TIFF (radar geometry, not georeferenced), SAFE (Standard Archive Format for Europe)
- **Access:** Public
- **M2M access:** Yes, with authentication.
- **Alternative access to data:**
  - PEPS API: https://peps.cnes.fr/resto/api/collections/S3/describe.xml
  - OpenSearch tool (only in French) https://peps.cnes.fr/rocket/plus/img/PEPS-IF-0-0170-ATOS_01_00_[2].pdf
  - Sentinel Open Access Hub: https://scihub.copernicus.eu/dhus/#/home
  - COPERNICUS API: https://scihub.copernicus.eu/twiki/do/view/SciHubWebPortal/APIHubDescription

2.2 UC2 - Environmental Bio-geochemical Assets

The definition of standard methodologies, reference collections, technologies, and quality control for BioGeoChemical (BGC) data typical of specific environmental assets (soil, water, air) is a major challenge for addressing fundamental scientific questions and improving our understanding of the earth and environmental conditions and dynamics [10].

2.2.1 Ocean Bio-Geo-Chemical Observatory (Pilot 5.2.1):

The observation of marine biogeochemical processes (BGC) is useful to monitor and understand the dramatic changes influencing ecosystem functioning and ocean health accompanying ecosystem dynamics and tracing the negative effect of human practice on environmental global changes [10].

2.2.1.1 Argo datasets

2.2.1.1.1 Dataset description

Argo is a global array of more than 3,800 active and free-drifting profiling floats that measures the temperature and salinity of the upper 2000 m of the ocean, even 6000 m for few Argo floats. This allows, for the first time, continuous monitoring of the temperature, salinity, and velocity of the upper ocean, with all data being relayed and made publicly available within hours after collection. The array provides 100,000 temperature/salinity profiles and velocity measurements per year distributed over the global oceans at an average of 3-degree spacing. 12% of floats of the global network (active and inactive floats) provide additional bio-geo parameters that are oxygen, nitrate, chlorophyll-a, pH, suspended particulates and downwelling irradiance (Figure 4). All data collected by Argo floats are publicly available in near real-time via the Global Data Assembly Centers (GDACs) in Brest (France) and Monterey (California) after automated quality control (QC), and in a scientifically quality controlled form, delayed mode data, via the GDACs within six months of collection.

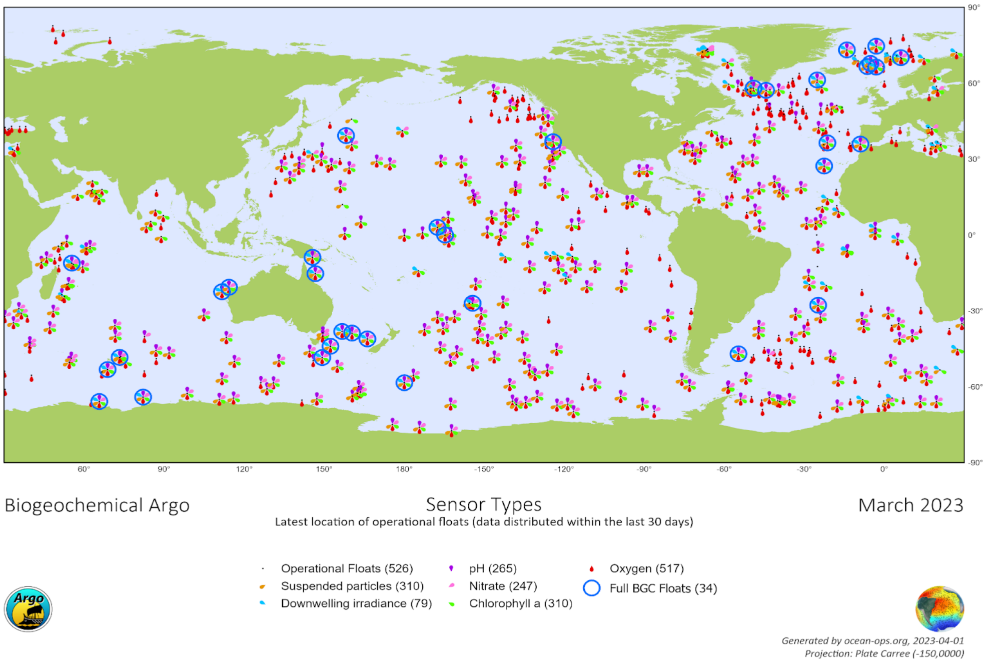

90°
60°
30°
0°
-30°
-60°
-90°
60°
90°
120°
150°
180°
-150°
-120°
-90°
-60°
-30°
0°
Biogeochemical Argo
Sensor Types
Latest location of operational floats (data distributed within the last 30 days)
March 2023
Operational Floats (526)
Suspended particles (310)
Downwelling irradiance (79)
pH (265)
Nitrate (247)
Chlorophyll a (310)
Oxygen (517)
Full BGC Floats (34)
Argo
Generated by ocean-ops.org, 2023-04-01
Projection: Plate Carree (-150,0000)

Figure 4 - Last location of BGC-Argo instruments.

2.2.1.1.2 Terms of Use

A user of Argo data is expected to read and understand the manual and the documentation about the data contained in the "attributes" of the NetCDF data files, as they contain essential information about data quality and accuracy. A user should acknowledge the use of Argo data in all publications and products where such data are used, preferably with the DOI and the following standard sentence: "These data were collected and made freely available by the international Argo project and the national programs that contribute to it."

Funded by
the European Union

2.2.1.1.3 Link to dataset and other relevant information

- **Dataset name:** Argo float data and metadata from Global Data Assembly Centre (Argo GDAC).
- **URL provided:** https://www.seanoe.org/data/00311/42182/
- **Dataset type:** Profiles
- **Data format:** NetCDF
- **Access:** Public
- **M2M access:** Yes
- **Alternative access to data:**
  The first two are accessible only with an FTP client (e.g Filezilla)
    - ftp.ifremer.fr/ifremer/argo
    - usgodae.org/pub/outgoing/argo
    - https://dataselection.euro-argo.eu/
    - https://erddap.ifremer.fr/erddap/tabledap/ArgoFloats.html
    - https://tds0.ifremer.fr/thredds/catalog/CORIOLIS-ARGO-GDAC-OBS/catalog.html
    - https://fleetmonitoring.euro-argo.eu/dashboard

2.2.1.2 ERA5 reanalysis

2.2.1.2.1 Dataset description

ERA5 is the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis for the global climate and weather for the past 8 decades. Data is available from 1940 onwards. Reanalysis combines model data with observations from across the world into a globally complete and consistent dataset. ERA5 provides hourly estimates for a large number of atmospheric, ocean-wave and land-surface quantities.

The data set presented is a regridded subset of the full ERA5 data set on native resolution. It should satisfy the requirements for most common applications. Data has been regridded to a regular lat-lon grid of 0.25 degrees for the reanalysis and 0.5 degrees for the uncertainty estimate (0.5 and 1 degree respectively for ocean waves). There are four main subsets: hourly and monthly products, both on pressure levels (upper air fields) and single levels (atmospheric, ocean-wave and land surface quantities).

2.2.1.2.2 Terms of use

Copernicus is funded under the Copernicus Regulation and operated by ECMWF under the ECMWF agreement. Access to all Copernicus (previously known as GMES or Global Monitoring for Environment and Security) Information and Data is regulated under Regulation (EU) No 1159/2013 of the European Parliament and of the Council of 12 July 2013 on the European Earth monitoring programme, under the ECMWF Agreement and under the European Commission's Terms and Conditions. Access to all Copernicus information is regulated under Regulation (EU) No 1159/2013 and under the ECMWF agreement. More detailed information at: https://cds.climate.copernicus.eu/api/v2/terms/static/licence-to-use-copernicus-products.pdf

2.2.1.2.3 Link to dataset and other relevant information

- **Dataset name:** ERA5 hourly data on single levels from 1940 to present.
- **URL provided:** https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form
- **Dataset type:** Gridded
- **Data format:** GRIB and NetCDF
- **Access:** Public.
- **M2M access:** yes, with authentication.
- **Alternative access to data:**
    - cdsAPI: https://cds.climate.copernicus.eu/api-how-to
    - CDS Toolbox: https://cds.climate.copernicus.eu/toolbox/doc/index.html

2.3 UC3 - Marine Omics Observatory

In the framework of the Biodiversity Observation Network (BON), this Use Case focuses on one of the fundamental real Life-Science challenges requiring tools useful to characterise the biodiversity asset, monitoring and predicting changes that can be crucial for several different aspects impacting human socio-economic interest [11].

2.3.1 Marine Omics Observatory (Pilot 5.3.1):

Starting from an ongoing effort undertaken by the European Marine Biological Resource Centre (EMBRC) infrastructure, with the establishment of the European Marine Omics Biodiversity Observation Network (EMO-BON), this pilot focuses on the challenge to set up a web-based VRE to provide products and services orientated to non-specialist researchers interested in omics approaches to study marine biodiversity (Figure 5). Today, EMO-BON includes several marine stations that will sample for genomic microbial marine biodiversity, essential ocean variables (EOVs), and essential biological variables (EBVs).

At the time of writing the practical data management for EMBRC 's EMO-BON project is actively being set up with large portions of infrastructure, workflow and automation in full development. As a consequence, our current assessment is not based on a practical exploration of available services, but purely on a description of the design and the intended services.

Funded by
the European Union

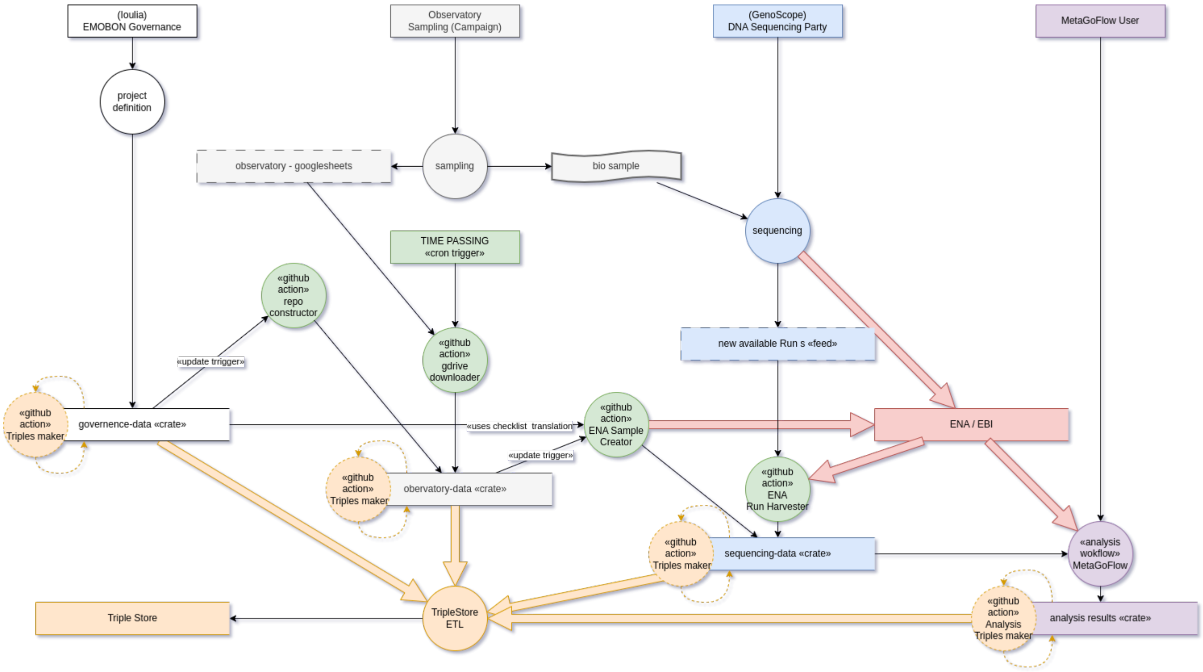


Figure 5 - Extended Data Flow Diagram for the Marine Omics Use Case

2.3.1.1 RO-Crate datasets

2.3.1.1.1 Dataset description

The raw data from EMO-BON is maintained in data files (mostly csv) under git version control and shared via various repositories on the github platform (see https://github.com/emo-bon).
By applying so-called 'github-actions', the changes to these data repositories trigger an orchestrated workflow that augments, converts and publishes these as interlinked mini-websites that use Linked Open Data principles and Semantic web technology to effectively build a so-called knowledge graph of interconnected data entities.

These entities are grouped into actual datasets that mirror the various git-repositories. The metadata of these sets in turn is captured using a novel packaging standard called RO-Crate (Research Object Crate, see: https://www.researchobject.org/ro-crate/1.2-DRAFT/).
Second to this containment of actual data, the application of RO-Crate allows to reference external data (such as sequencing data which is stored on the ENA, European Nucleotide Archive, or externally maintained image files). Next, RO-Crate provides a central semantic description of the dataset and its content as well as its declared external references by using schema.org (an application to create, organise and maintain structured data on the internet) extensible with any specific semantic vocabulary useful to the domain. Lastly, a set of recommended practices make this information naturally discoverable to both end users and automated consumption strategies.

Funded by
the European Union

2.3.1.1.2 Terms of use

EMBRC have chosen to have all EMO-BON data to be published via the RO-Crate datasets under an open CC-BY licence. An important nuance of this approach is that some of the data (images and sequence data) are not kept inside these RO-Crates, but only referred to. For these, the actual raw data access is kept under an embargo for six months after their acquisition. This period allows the partners in the EMBRC consortium to validate and complete the data space with the results of their analysis.

2.3.1.1.3 Link to dataset and other relevant information

Dataset name: EMO-BON Data in RO-Crates
URL provided: https://data.emobon.embrc.eu/{crate_name}-data/
Dataset type: tabular info of planning, registered samples, associated measurements of facts, references to DNA sequence data, and results from genomic analysis workflows
Data format: CSV and derived semantic triples in turtle format
Access: Public.
M2M access: Yes. Including automatic discovery of available sets through FAIR-signposting; a full semantic description of the available datasets (using DCAT), each RO-Crate dataset (using schema.org inside the ro-crate-metadata.json); data-level semantics in the generated turtle files; a feed publishing the updates to data-entities (using LDES).
Alternative access to data: The SPARQL endpoint mentioned below

2.3.1.2 Triple Store / SPARQL endpoint

2.3.1.2.1 Dataset description

The RO-Crate datasets described above will be automatically harvested and ingested into a triple store that will allow free exploration of the aggregated data through SPARQL queries. As such this service endpoint can be seen as a [[Derived Data Provider]] following the schema and nomenclature introduced by D4.1. Indeed, using the internal indexing of its triple storage this will effectively serve addressable subsets of the EMO-BON data-space.

For this to actually fit the concept of the [[Data Provider]] an additional access API will need to be designed, implemented and installed. This process will be driven by the outcome of the recently started VRE analysis.

2.3.1.2.2 Terms of Use

As with the RO-Crate datasets, access and usage of the SPARQL endpoint will be open under a CC-BY licence. Since sequences and images will not be stored here, but only referenced, the same embargo settings as above apply.

Funded by
the European Union

2.3.1.2.3 Link to dataset and other relevant information

Dataset name: EMO-BON Triple Store / SPARQL Endpoint
URL provided: https://sparql.emobon.embrc.eu/ (not yet active)
Dataset type: knowledge graph aggregated from the triples in the RO-Crates mentioned above
Data format: sparql response format in csv-tsv, xml or json as defined in the SPARQL protocol
Access: Public.
M2M access: Yes.
Alternative access to data: n/a

3 Assessment on UC dataset requirements.

As described in D4.1 the data lake is to be a distributed set of components that collaborate towards offering a variety of services and features expected from this conceptual "data lake". In this architectural view the realisation of "ingestion" is the shared responsibility between some central [[Discovery]] block that harvests its knowledge of the available datasets from various [[Data Provider]] blocks.

The focus on the selected datasets we are taking in this document urges us to question if and to what level these are fitting the expectations expressed for the [[Data Provider]] services that are to provide and describe them.

Quoting from D4.1:

Our proposed contract includes the following requirements (for the [Data Provider]):

1. All provided datasets must have a clear identifier, which should be a URI. This URI can be resolved to a URL for downloading the data. (**REQ #1**)
2. [[Data providers]] must produce a listing of the available datasets with their identifiers. This listing is the one that will be harvested and should use standard formats like DCAT and include an elaborate "Dataset Information Model" that allows for the needed cleverness in [[Discovery]] (**REQ#2**)
3. The Data Providers should allow for effective harvesting by chunking this information stream, ordered by the last modification date (descendingly), into separate change-blocks. This resulting "change-feed" or "stream of changes" should be encoded using standard techniques like LDES or OAI-PMH. (**REQ#3**)

Table 2 offers a handy overview of the requirements for each dataset:

Table 2 - Assessment of the requirements of each UC/pilot dataset. ☑ Yes ☒ No

Use Case / Pilot	Dataset	REQ#1	REQ#2	REQ#3
UC1: Coastal water Dynamics	Daily river Runoff	☑	☒	☒
	Eutrophication and Acidity data collection	☑	☒	☒
UC1: Earth Critical zones observatory	Satellite images from sentinel II	☑	☑	☒
UC1: Volcano Space Observatory	Sentinel-5P/TROPOMI SO2	☑	☑	☒
	Sentinel-1 SLC	☑	☑	☒
UC2: Ocean Bio-Geo-Chemical Observatory	Argo datasets	☑	☑	☒
	ERA5 Reanalysis	☑	☑	☑
UC3: Marine Omics observatory	RO-Crate	☑	☑	☒
	Triple-Store / Sparql endpoint	☒	☒	☒

4 References

[1] M.L. Chiusano, R. Schlitzer, S. Simoncelli, C. Troupin, G. Langella, F. Terribile, N. Pascal, M. Boichu, R. Grandin, V. Racape, C. Schmechtig, R. Sauzede, A. Sizun, A. Giorgetti, C. Reyes, C. Cox, K. Exter, M. Portier, S. Ninidakis, I. Santi, L. Bosso, L. Ambrosino and M. Miralto, "FAIR-EASE_D5.1_Report on key requirements from Use Cases/Pilots," Zenodo, 2023. DOI: 10.5281/zenodo.7588904.

[2] "FAIR-EASE - Use Cases: Earth & Environmental Dynamics," [Online]. Available: https://fairease.eu/use-cases/earth-environmental-dynamics. [Accessed 04.2023].

[3] R. Schlitzer and S. Mieruch-Schnuelle, "webODV Explore," 2021. [Online]. Available: https://explore.webodv.awi.de. [Accessed 03.2023].

[4] "European Marine Observation and Data Network (EMODnet) - Chemistry: Eutrophication," [Online]. Available: https://emodnet.ec.europa.eu/en/eutrophication. [Accessed 03.2023].

[5] A. Barth, L. Buga, L. Fyrberg, J. Gatti, A. Giorgetti, G. Giorgi, S. Iona, M. M. Larsen, M. Lipizer, D. Schaap, R. Schlitzer, M. Vinci, S. Watelet and M. Wenzer, "EMODnet Thematic Lot n° 4 –

Chemistry - Methodology for data QA/QC and DIVA products. V8," OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale), Oceanographic Section, Trieste, IT, 2015.

[6] L. Buga, G. Sarbu, L. Fryberg, K. Wesslander, J. Gatti, S. Iona, M. Tsompanou, M. M. Larsen, A. Østrem, M. Lipizer, M. M. Jack and A. Giorgetti, "Quality Control steps for EMODnet Chemistry Eutrophication aggregated datasets - v2021," OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale), Oceanographic Section, Trieste, 2021.

[7] "SENTINEL-2 MISSION GUIDE," [Online]. Available: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2. [Accessed 04.2023].

[8] "Legal notice on the use of Copernicus Sentinel Data and Service Information," [Online]. Available: https://sentinels.copernicus.eu/documents/247904/690755/Sentinel_Data_Legal_Notice. [Accessed 04.2023].

[9] "Sentinel-5 precursor/TROPOMI Level 2 Product User Manual Sulphur Dioxide SO_2" [Online]. Available: https://sentinels.copernicus.eu/documents/247904/2474726/Sentinel-5P-Level-2-Product-User-Manual-Sulphur-Dioxide.pdf/3ea6cac4-b40c-428c-b8cb-178f8ea35d91?t=1658408531468. [Accessed 04.2023].

[10] "FAIR EASE - Use Cases: Environmental Bio-geochemical Assets," [Online]. Available: https://fairease.eu/use-cases/environmental-bio-geochemical-assets. [Accessed 04 2023].

[11] "FAIR-EASE, Use Cases: Biodiversity Observation," [Online]. Available: https://fairease.eu/use-cases/biodiversity-observation. [Accessed 04 2023].

[12] "Sentinel-2 User Handbook" [Online]. Available: https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook. [Accessed 05 2023].

Funded by
the European Union