Organised by:

EuroGOOS
European Global Ocean Observing System

Foras na Mara
Marine Institute

10th EuroGOOS International Conference

3-5 Oct 23
Galway, Ireland

European Operational Oceanography
for the Ocean we want – addressing
the UN Ocean Decade Challenges

Conference Proceedings

December 2023

2021 2030 United Nations Decade
of Ocean Science
for Sustainable Development

10th EuroGOOS International Conference

3-5 Oct 23
Galway, Ireland

European Operational Oceanography
for the Ocean we want – addressing
the UN Ocean Decade Challenges

Conference Proceedings

Edited by: Dina Eparkhina and Joseph E. Nolan

Published by

EuroGOOS AISBL
29 Rue Vautier
1000 Brussels
Belgium
www.eurogoos.eu

To be quoted as follows

Proceedings of the 10th EuroGOOS International Conference.
European Operational Oceanography for the ocean we want - Addressing the UN Ocean Decade challenges.
3-5 October 2023, Galway, Ireland.
Eparkhina, D., Nolan, J.E. (Eds).
EuroGOOS. Brussels, Belgium. 2023.
DOI: http://hdl.handle.net/10793/1883

Blue-Cloud-2026, a Federated European Ecosystem to deliver FAIR & Open data and analytical services, instrumental for the Digital Twins of the Oceans

Authors

Dick Schaap[1], Sara Piitonet[2] and Pasquale Pagano[3]

[1] MARIS BV, Gildeweg 7A, 2632BD Nootdorp, The Netherlands, dick@maris.nl

[2] Trust-IT, Via Francesco Redi, 10, Apt. #11-12, 4th floor, 56124 Pisa, Italy, s.pittonet@trust-itservices.com

[3] CNR-ISTI, Via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy, pasquale.pagano@isti.cnr.it

Keywords

Open science, EOSC, data discovery, data access, EOVs, data lakes, subsetting API, VRE, VLabs, WorkBenches

Abstract

The pilot Blue-Cloud H2020 project combined interests of developing a thematic marine EOSC cloud and serving the Blue Economy, Marine Environment and Marine Knowledge agendas. It deployed a versatile cyber platform with smart federation of multidisciplinary data repositories, analytical tools, and computing facilities in support of exploring and demonstrating the potential of cloud based open science for ocean sustainability, UN Decade of the Oceans, and G7 Future of the Oceans. The pilot Blue-Cloud delivered: 1) Blue-Cloud Data Discovery & Access service (DD&AS), 2) Blue-Cloud Virtual Research Environment infrastructure (VRE) and 3) Five multi-disciplinary Blue-Cloud Virtual Labs (VLabs). Since early 2023, Blue-Cloud 2026 aims at a further evolution into a Federated European Ecosystem to deliver FAIR & Open data and analytical services, instrumental for deepening research of oceans, EU seas, coastal & inland waters, and building a major data ground segment for the Digital Twins of the Oceans (DTO's).

The EMODnet Data Ingestion portal plays a role in the pathways towards the EMODnet data portal. Specifically, the services it provides to data holders include: (a) data submission, with integrated services such as the online submission form, user management service, tracking service, (b) discovery and access, operating on the ingested and completed data submissions, and (c) operational data integration.

1. INTRODUCTION

1.1. Blue-Cloud 2026 Project

Over the past decades, Europe developed an impressive capability for aquatic environmental observation, data-handling and sharing, modelling and forecasting, second to none in the world. This builds upon national environmental observation and monitoring networks and programs, complemented with EU infrastructures such as the Copernicus satellite observation programme and related thematic services, the European Marine Observation and Data Network (EMODnet), as well as a range of environmental European Research Infrastructures and major R&D projects.

Within this framework, since October 2019, the pilot Blue-Cloud project (https://blue-cloud.org/about-h2020-blue-cloud) combined both the interests of the European Open Science Cloud (EOSC), aiming to provide a virtual environment with open and seamless access to services for storage, management, analysis and re-use of research data, across borders and disciplines, and the blue research communities by developing a collaborative web-based environment providing simplified access to an unprecedented wealth of o multi-disciplinary datasets from observations, analytical services, and computing facilities essential for blue science.

The successor Blue-Cloud 2026 project (https://blue-cloud.org) aims at a further evolution of this pilot ecosystem into a Federated European Ecosystem to deliver FAIR & Open data and analytical services, instrumental for deepening research of oceans, EU seas, coastal & inland waters. It develops a thematic marine extension to EOSC for open web-based science, serving the needs of the EU Blue Economy, Marine Environment and Marine Knowledge agendas.

Blue-Cloud 2026's overall Objective is to expand the federated approach of Blue-Cloud, involving more aquatic data stakeholders, and interacting with EOSC developments, in support of the EU Green Deal, UN SDG, EU Destination Earth, and the EU Mission Starfish on healthy oceans, seas, coastal and inland waters, ultimately to provide a core data service for the Digital Twins of the Ocean.

1.2. Blue-Cloud Tools and Services

The pilot Blue-Cloud project delivered:

- **Blue-Cloud Data Discovery & Access service (DD&AS)**, federating key European data management infrastructures, to facilitate users in finding and retrieving multi-disciplinary datasets from multiple repositories);

- **Blue-Cloud Virtual Research Environment infrastructure (VRE)** providing a range of services and facilitating orchestration of computing and analytical services for constructing, hosting and operating Virtual Labs for specific applications;

- **Five multi-disciplinary Blue-Cloud Virtual Labs (VLabs)**, configured with specific analytical workflows, targeting major scientific challenges, and serving as real-life **Demonstrators,** which can be adopted and adapted for other inputs and analyses.

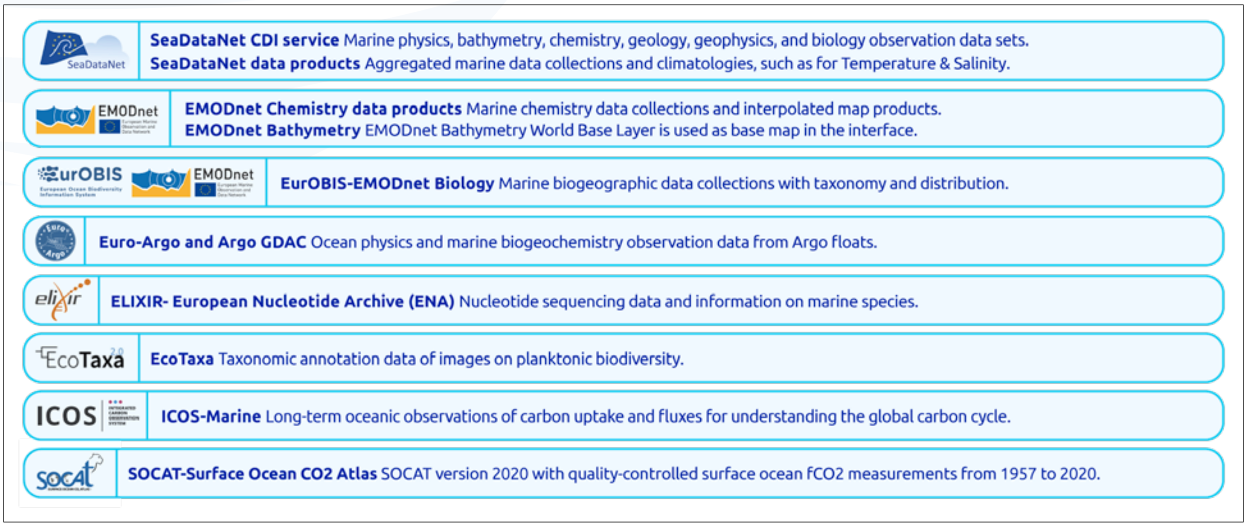
Over the course of 42 months starting in January 2023, Blue-Cloud 2026 will evolve these core services, integrating more blue analytical services, configuring more Virtual Labs, improving services for uptake of new data sets from a multitude of data originators (such as SeaDataNet, EurOBIS, Euro-Argo, ELIXIR-ENA, SOCAT, EcoTaxa, and ICOS-Ocean), and major e-infrastructures, namely EUDAT, D4Science, EGI, and WEkEO (CMEMS DIAS) and for discovery and access to their structured data collections.

Moreover, it will develop:

- **Three Workbenches for Essential Ocean Variables (EOVs)**, implementing efficient operational workflows that allow ocean and data scientists to harmonise, validate and qualify large and various *in situ* data sources into high quality EOV data collections, which are key input for many applications, including the Digital Twins of the Ocean.

2. BLUE-CLOUD DATA DISCOVERY & ACCESS SERVICE

The federated Data Discovery & Access Service (DD&AS) provides users with an easy and FAIR service for discovery and access to multi-disciplinary data sets and data products managed and provided by leading Blue Data Infrastructures (BDIs). The federation facilitates sharing of datasets as input for analytical and visualisation services and applications, that are hosted and further developed as VLabs and WorkBenches in the Blue-Cloud VRE. The DD&AS has been developed, is operated, and is being upgraded and expanded by MARIS together with CNR (IIA) and CINECA (EUDAT), interacting with each of the BDIs. The figure below gives an overview of the currently federated BDIs and their data resources.



SeaDataNet CDI service Marine physics, bathymetry, chemistry, geology, geophysics, and biology observation data sets.
SeaDataNet data products Aggregated marine data collections and climatologies, such as for Temperature & Salinity.

EMODnet Chemistry data products Marine chemistry data collections and interpolated map products.
EMODnet Bathymetry EMODnet Bathymetry World Base Layer is used as base map in the interface.

EurOBIS-EMODnet Biology Marine biogeographic data collections with taxonomy and distribution.

Euro-Argo and Argo GDAC Ocean physics and marine biogeochemistry observation data from Argo floats.

ELIXIR- European Nucleotide Archive (ENA) Nucleotide sequencing data and information on marine species.

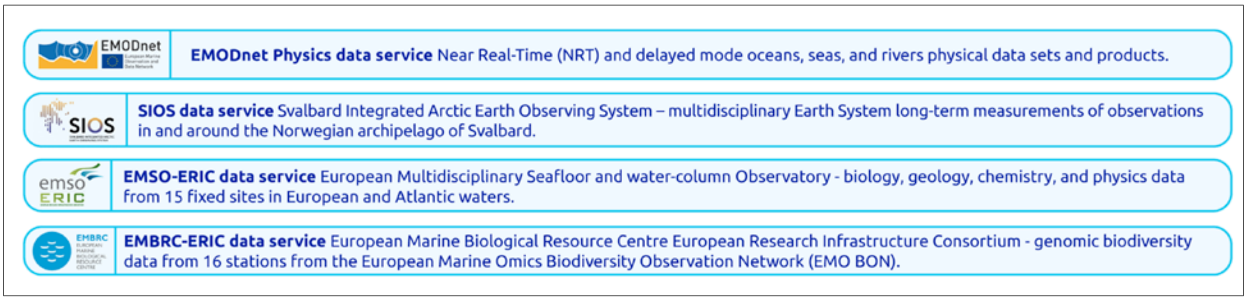
EcoTaxa Taxonomic annotation data of images on planktonic biodiversity.

ICOS-Marine Long-term oceanic observations of carbon uptake and fluxes for understanding the global carbon cycle.

SOCAT-Surface Ocean CO2 Atlas SOCAT version 2020 with quality-controlled surface ocean fCO2 measurements from 1957 to 2020.

Figure 1. Current Blue Data Infrastructures (BDIs) federated in the DD&AS.

The DD&AS this way facilitates common discovery and access to more than 10 million marine datasets for physics, chemistry, geology, bathymetry, biology, biodiversity, and genomics. It is fully based on machine-to-machine brokering interactions with web services as provided and operated by the BDIs. As part of Blue-Cloud 2026 it will expand by federating more leading European Aquatic Data Infrastructures, as illustrated in the next figure:



EMODnet Physics data service Near Real-Time (NRT) and delayed mode oceans, seas, and rivers physical data sets and products.

SIOS data service Svalbard Integrated Arctic Earth Observing System – multidisciplinary Earth System long-term measurements of observations in and around the Norwegian archipelago of Svalbard.

EMSO-ERIC data service European Multidisciplinary Seafloor and water-column Observatory - biology, geology, chemistry, and physics data from 15 fixed sites in European and Atlantic waters.

EMBRC-ERIC data service European Marine Biological Resource Centre European Research Infrastructure Consortium - genomic biodiversity data from 16 stations from the European Marine Omics Biodiversity Observation Network (EMO BON).

Figure 2. Blue Data Infrastructures (BDIs) to be added to the DD&AS federation.

The DD&AS works with brokerage services both at metadata and data level. Discovery and selection are done in a two-step approach - from data collections using a common metadata profile, to detailed records using extended metadata profiles- and fully based on web services (such as OGC-CSW, OAI-PMH, ERDDAP, DCAT, and dedicated APIs), as published and maintained by connected BDIs. For the first step - collection level - the DAB Metadata Brokerage mechanism (developed by CNR (IIA)) is used, harmonising individual outputs of BDI discovery web services to a common syntactic metadata model (ISO19115 – 19139). The second step drills down within identified collections to get more specific data, using free search, geographic and temporal criteria, but this time at granular level, and including additional BDI-specific search criteria. Users are able to download and store the retrieved data collections on their own machines or in a Data Pool in the Blue-Cloud VRE. The two-step approach is effective to go from coarse to fine and to determine at an early stage which of the BDIs might have interesting datasets.

The shopping mechanism is based on the shared experience and services of MARIS, IFREMER, and EUDAT from developing and managing the SeaDataNet CDI service, from which selected services were adapted. The DD&AS is available by means of a GUI, while level 1 – the collections catalogue – is also published as OGC CSW service, which is e.g. used for sharing the Blue-Cloud collection records with the EUDAT B2FIND data catalogue service at the EOSC Marketplace. This way, EOSC users can find relevant Blue-Cloud data collections and are then directed to the Blue-Cloud DD&AS for further down drilling and actual download of datasets at granule level.

As part of Blue-Cloud 2026, the DD&AS is being elaborated in a number of ways:

- Optimising and refining current web services at BDIs and DD&AS federation principles for increasing the FAIRness of the integrated services and the provided output;

- Achieving semantic interoperability for common metadata tags, such as parameters, instruments, platforms, sea regions, etc: Each BDI is using own vocabularies for several metadata tags. A semantic brokerage service component will be developed and integrated into the DD&AS;

- Adding sub-setting and extracting services to the DD&AS by means of APIs at BDIs: Currently, the DD&AS supports discovery and download of predefined data objects, while several applications might require specific extracts and slices of data. Also, repositories like WEkEO (CMEMS DIAS) are managing big datasets for which sub-setting and slicing is more appropriate as otherwise very large files are downloaded, which might be updated and increased regularly as results of model runs. New APIs will be developed or existing APIs will be adapted, where required to facilitate the compilation of Blue-Cloud 'Data Lakes' for specific data types;

- Defining, developing, and operating Blue-Cloud Data Lakes: Data Lakes will function as 'harmonised buffers' of observation data as combined from multiple BDIs and also from major international repositories like the NOAA World Ocean Database (WOD) and others. Data Lakes will improve data access both in terms of data harmonisation and of technical efficiency of data access. The data lakes will be very relevant for organising input for the Vlabs and Work Benches, while they also will be relevant for providing high quality output to other initiatives such as the Digital Twins of the Ocean (DTO).

3. BLUE-CLOUD DATA VIRTUAL RESEARCH ENVIRONMENT

The Blue Cloud Virtual Research Environment (VRE) provides an Open Science platform for collaborative marine research, using a wide variety of datasets and analytical tools, complemented by generic services such as sub-setting, pre-processing, harmonising, publishing and visualisation. For each Virtual Lab and each Workbench, accounts of researchers will be configured at the VRE. Each will enact a family of analytical workflows which consist of a series of applications and make use of selected datasets as input. The multi-disciplinary datasets can be retrieved from the BDIs using the Blue Cloud DD&AS and its Data Lakes, and external resources. Outputs, such as data products, data collections, maps, notebooks, software applications, and services can be documented with DOIs for citation, provenance for reproducibility, and published in the Blue-Cloud Catalogue. All methods and services in this Catalogue are exchanged with the EOSC Portal Catalogue & Marketplace.

The Blue-Cloud VRE is organised as a multi-site digital infrastructure with a central hub and peripheral sites. The central hub is located at the D4Science data centre, operated by CNR. It is responsible for four common services: 1) Identity and Access Management (IAM) Service, 2) the Information System (IS), 3) the Resource Manager (RM), and 4) the persistent, fault-tolerant and replicated Storage Manager (SM). The peripheral sites host most of the computing resources and the tailored storage devices that offer low-latency and efficient storage solutions for supporting large and complex data analytics processes. Overall, it offers an aggregated shared capacity of 3,650 CPU cores with 13.7 TB RAM and 0.6 PB persistent storage and will power the Blue-Cloud 2026 VRE at the beginning of the project. This initial capacity will be then expanded with additional sites, each with a minimum capacity of 256 CPU cores with 512 GB RAM and 10 TB persistent storage to enlarge the overall computing capacities and enable distribution of the load, fault tolerance, advanced resilience and exclusive assignment of resources to the data and computing intensive WorkBenches for EOVs. By exploiting those digital resources, services will be able to join and leave the VRE according to the provisioning policies specified at the service integration time. This "system of systems" will give a good grade of autonomy at the site level (independence and evolution), openness (join and leave; dynamic reconfiguration); distribution (interdependence and interoperability) which makes it easier to define policies for the addition of new site providers to the VRE. All physical resources of the infrastructure will be manageable through a single platform for both hardware and software layers, which will simplify the RI management, enhance scaling of the VRE deployment, and reduce the total cost of ownership.

Most of this physical architecture is invisible to the final users, which will see and access the resources from a single and unified access point (i.e., the Blue-Cloud Gateway). The Analytics Computing Framework - i.e., the Kubernetes clusters, used to deliver Jupyter Notebooks via the JupyterLab web-based interactive development environment, RShiny and RStudio applications, the Docker Swarm clusters used to operate containerised applications, the computing clusters used to support high-throughput computing (HTC) tasks, and the worker clusters used to support map-reduce jobs - will be located in several sites to ensure scalability, reliability and fault-tolerance.

As part of Blue-Cloud 2026 project, the VRE will be elaborated in a number of ways:

- Further evolution of the 4 VRE common services and operation: IAM, IS, RM, and SM components will be evolved to better serve the needs of the enlarged community and to deliver them a high-quality operation. IAM will join Identity Federation of the EGI Check-in service in support of Single Sign On (SSO) for the users at the integrated VRE platform. IS and RM will manage new resources, computing and services, joining the Blue-Cloud VRE for enlarging its overall capacities and capabilities. SM will offer more tailored configurations for analytical services;

- Enhancing the computing facilities: The Analytics Computing Framework will support HTC on (Docker and Linux) containerised applications. This will improve the portability of the application and its reproducibility in other infrastructures, isolation of application packages, increased security, and reduced operational costs. The JupyterHub component will be enhanced for deployment over multiple Kubernetes clusters to ensure high scalability;

- Expanding the VRE by federating multiple e-infrastructures: the VRE powered by D4Science, will federate resources and services, namely provided by EGI, WEkEO, EUDAT, and JERICO-CORE. It will provide a return on investment for each provider joining the Blue-Cloud VRE since compliance with the EU Regulation (as for GDPR), security, monitoring, accounting, user management, fault management, and alerting management will be granted by the VRE with no cost for the provider of the digital resources. In turn, the provider will be acknowledged by the users in all products and services generated by exploiting the provider's resources. This seamless expansion will support orchestrating workflows, with algorithms and computing resources, divided over and running at the different e-infrastructures. This way, it will also open the connectivity to applications in new EU projects such as iMAGINE (AI applications for marine domain), and EGI-ACE (applications for ocean use cases);

- Expanding the monitoring of availability and usage of the integrated VRE platform: the initial Blue-Cloud VRE monitoring system, empowering central dashboards of uptime and users and uses of all services will be expanded towards the newly developed services and the additional e-infrastructures.

Finally, **an EOSC Blue Task Force for Blue-Cloud integration with EOSC core services** will be established to ensure the compliance of the Blue-Cloud technical framework with the EOSC principles for service management, including Service Level Agreement (SLA), Operation Level Agreement (OLA), incident and service request management, service availability and continuity management.

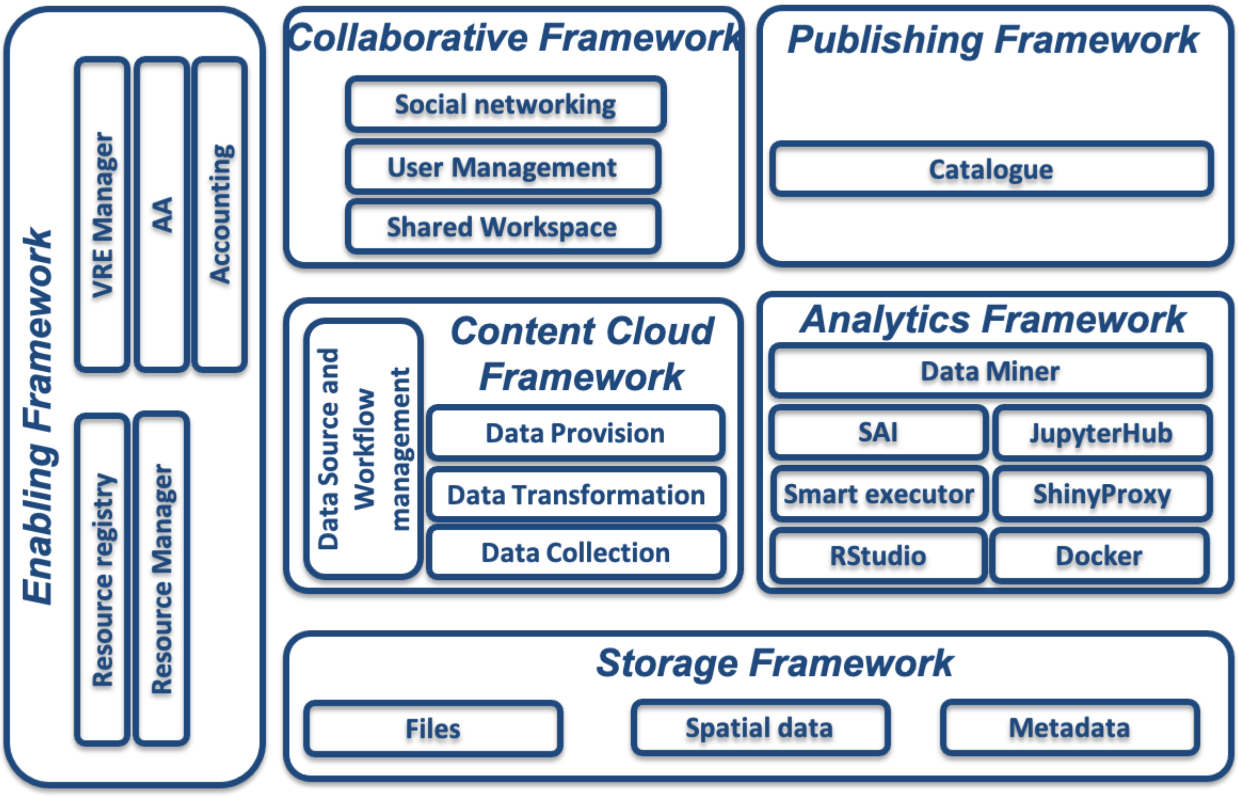


Figure 3. High-level architecture of the D4Science infrastructure.

ACKNOWLEDGEMENTS

he Blue-Cloud 2026 project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement N° 101094227.

REFERENCES

Schaap D. (2023) Exploring and demonstrating the potential of Open Science for ocean sustainability, https://zenodo.org/record/8063865

Pagano P., Pittonet S., Drago F., Giuffrida M. (2023), Providing computing platforms and analytical services to facilitate the collaboration between researchers, https://zenodo.org/record/7827629

Assante M., (2023) Blue-Cloud VRE Open Science services for building, hosting and operating Virtual Labs, https://zenodo.org/record/8083123

Palermo, F. (2023) Marine Environmental Indicators VLab in the pilot Blue-Cloud, https://zenodo.org/record/8083142

Schaap D., Pittonet S., Drago F., Giuffrida M. (2023), Supporting marine data discovery and accessibility to enable cross-domain research, https://zenodo.org/record/7646200

Pesant S. (2023) Data life cycle for life science, https://zenodo.org/record/7630274