Blue-Cloud: Exploring and demonstrating the potential of Open Science for ocean sustainability

Dick Schaap
MARIS
Nootdorp, The Nederlands
dick@maris.nl

Pasquale Pagano
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Pisa, Italy
pasquale.pagano@isti.cnr.it

Leonardo Candela
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Pisa, Italy
leonardo.candela@isti.cnr.it

Massimiliano Assante
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Pisa, Italy
massimiliano.assante@isti.cnr.it

***Abstract*— The Blue-Cloud project is part of 'The Future of Seas and Oceans Flagship Initiative' of the European Commission and runs since October 2019. It has established a *pilot cyber platform*, providing researchers access to multi-disciplinary datasets and derived data products from observations, in-situ and satellite-based, analytical services, and computing facilities essential for blue science to better understand & manage the many aspects of ocean sustainability. A number of core services have been delivered and are now in a phase of wider dissemination and uptake by marine researchers. Core services are the Federated Data Discovery & Access Service (DD&AS), the Blue-Cloud Virtual Research Environment (VRE), and five Blue-Cloud Virtual Labs.**

Keywords*— *Web-based Open science, Marine and ocean data, Virtual Research Environment, Virtual Labs, Ocean sustainability, Federated discovery and access, Federated computing and analytics, EMODnet, Copernicus, Blue-Cloud

I. Introduction

Over the past decades, Europe already has developed an impressive capability for marine 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. This way, an expanding European capacity is provided for collecting, managing, and processing of in-situ and remote sensing data, while federating and interacting with national, regional, and global activities and initiatives. Research infrastructures, including the European Open Science Cloud (EOSC) (https://digital-strategy.ec.europa.eu/en/policies/open-science-cloud), and e-infrastructures are crucial enablers of research and technological innovation and drivers of multidisciplinary and data-intensive science. By means of EOSC, Europe should benefit from an integrated, inter-operable and effective ecosystem of these infrastructures to facilitate and accelerate fundamental knowledge creation and technology deployment in support of Open Science and European technology leadership. The aim is to develop the EOSC in a more cohesive and structured manner so that it becomes a fully operational enabling ecosystem for the whole research data lifecycle.

Oceans, seas, coastal and inland waters are vital for our societies and the future of our planet, and there are challenges in the aquatic sciences domain that may be addressed with better and broader use of existing data resources and wider application of web-based analytical services, including services for Big Data analysis in support of multidisciplinary and collaborative research. Aquatic bodies are home to diverse ecosystems and habitats, provide a wealth of resources, strongly regulate climate, and offer many resources for economic opportunities. The combination of long-term global change and multiple local stressors affects ecosystems in unpredictable ways to a point of no return with significant socio-economic impact. Therefore, a better understanding of the dynamics and complex geochemical interactions is key to allow a sustainable use, conservation and implementation of mitigation, and/or restoration plans for these essential ecosystems.

II. Blue-Cloud project

Since October 2019, the pilot Blue-Cloud project (www.blue-cloud.org) is well underway as part of 'The Future of Seas and Oceans Flagship Initiative' of the EU H2020 programme. It combines both the interests of EOSC and the blue research communities. It is undertaken by leading European ocean and marine data and knowledge initiatives such as EMODnet and Copernicus Marine Service (CMS), together with leading marine environmental research infrastructures such as SeaDataNet, EurOBIS, Euro-Argo, ELIXIR-ENA, SOCAT, EcoTaxa, and ICOS-Ocean (qualified as Blue Data Infrastructures (BDIs)), and major e-infrastructures, namely EUDAT, D4Science, and WEkEO (CMS DIAS).

The pilot Blue-Cloud project has delivered:

- **A Blue-Cloud Data Discovery & Access service (DD&AS):** a "beta release" federation of key European Blue Data Infrastructures to facilitate users in finding and retrieving multi-disciplinary datasets in a common and efficient way.
- **Blue-Cloud Virtual Research Environment (VRE):** facilitates collaborative research offering computing, storage, analytical, and generic services to be orchestrated with a large variety of data resources by

researchers for constructing, hosting and operating analytical workflows for specific applications;

- **Blue-Cloud Virtual Labs:** five VLabs have been developed by Blue-Cloud scientific experts as pilot open science demonstrators of complex ocean related challenges. Each VLab combines application services and makes use of selected datasets as input for delivering data products and/or dedicated services.

The Blue Cloud initiative is relevant in the context of major EU data and knowledge initiatives with a marine focus such as EMODnet (European Marine Observation and Data Network), Copernicus (The European Earth Observation Programme), EuroGOOS (European Global Ocean Observing System), ESFRI (European Strategy Forum on Research Infrastructures), and as an implementation as part of the European Open Science Cloud (EOSC), sharing and using key features and principles of EOSC to develop and shape a more integrated European marine and ocean data management and research landscape.

III. Federated Data Discovery and Access Service

The 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). FAIR [10] stands for Findable, Accessible, Interoperable, and Reusable and covers guiding principles for scientific data management and stewardship. These principles refer to three types of entities, namely data (or any digital object), metadata (information about that digital object), and infrastructure. In the DD&AS the following BDIs have been federated: EMODnet, SeaDataNet, EurOBIS, ICOS, SOCAT, EcoTaxa, Argo, and ELIXIR-ENA. Together these BDIs currently manage more than 10 million data sets for physics, chemistry, geology, bathymetry, biology, and genomics. The DD&AS works with brokerage services both at metadata and data level. Discovery and selection are done in a two-step approach. The first step has a focus on identifying interesting data at an aggregated collection level, with free search, geographic and temporal criteria as main query operators. In this first step use is made of a common metadata profile. The BDIs are operating web services such as OGC-CSW, OAI-PMH, ERDDAP, DCAT, dedicated APIs, and FTP with quite a diversity of functionalities and formats. The GEODAB Brokerage mechanism is used for harmonising individual outputs of BDI discovery web services to a common syntactic metadata model (ISO19115 – 19139) at data collection level. The second step drills down within identified collections to get more specific data, using free search, geographic and temporal criteria, but this time making use of the BDI own discovery web services at granule level, and including additional BDI-specific search criteria. Finally, users are able to download and store the retrieved data sets on their own machines or push these to the Blue-Cloud VRE. The two-step approach for data discovery and access is effective to go from coarse to fine and to determine in an early stage which of the BDIs might have interesting data sets. It is also effective to keep the number of entries relatively limited in the exploratory first step of discovery. The granule level as a second level is applicable to several of the blue data infrastructures, in particular in cases with observation (raw) data which often can be very large collections with numerous data sets. For instance, the SeaDataNet CDI service currently gives discovery and access to more than 2.5 million individual observation data sets for physics, chemistry, geology, biology, geotechnics, and bathymetry. At first level, there are circa 800 CDI aggregated records at collection level, which then give access to the more than 2.5 million granule records, which in the end can be downloaded. For the data access part of the Blue-Cloud DD&AS, a data brokerage service has been developed, integrating the internal Blue-Cloud level 1 metadata catalogue (see above), a series of machine-to-machine interfaces to the blue data infrastructures for level 2 queries, and a shopping mechanism to support the actual discovery and retrieval functions. This part makes use of the experience and software services that partners have developed and are managing for the SeaDataNet CDI service. For the Blue-Cloud selected services have been adopted and/or adapted. Implementing this approach largely depends on the interfaces of the BDIs, that should be supportive. Therefore, as a preparatory step, the existing web services and APIs of each BDI were analysed, tested and documented. This included inter alia finding the best ways for the deployment of level 2 queries and how to construct the download URLs. As an overall result, users of the DD&AS now have a common user interface to discover and download data sets from the federated BDIs which themselves have different user interfaces. Moreover, the applied approach is scalable to expand the federation to more BDIs and more data content per BDI. Further innovations are planned in the near future, such as adding semantic brokerage to harmonise terminologies as used for parameters, platforms, devices and others by mapping between vocabularies used in BDIs. Another expansion will include developing and adding subsetting functionality to facilitate querying and extracting data sets for specific data criteria, like all temperature values at specific depth range and geographic area and with specific quality flags. This will require the availability or development of subsetting APIs at the various BDIs, thereby also considering sufficient performances. These developments will be undertaken in the successor Blue-Cloud 2026 project which will start from January 2023.

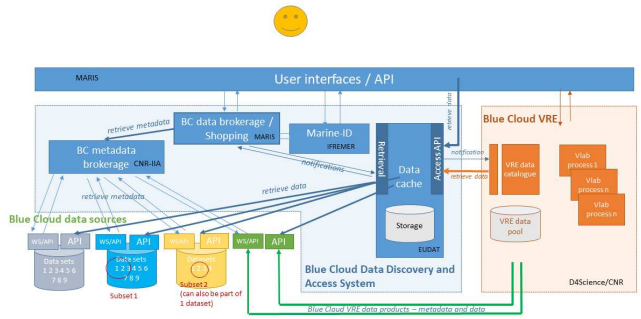

Fig. 1. Architecture of the Blue-Cloud discovery and access service

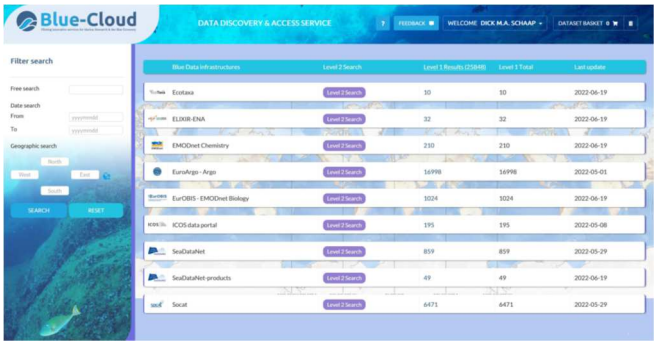

Fig. 2. Homepage of the DD&AS comon user interface

IV. 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.

This platform follows the *system of systems* [1] approach, where the constituent systems offer "resources" (namely services) for the implementation of the resulting system facilities. In particular, such a platform aggregates "resources" from "domain agnostic" service providers (e.g. D4Science, EGI) as well as from community-specific ones (e.g. WEkEO) to build a unifying space where the aggregated resources can be exploited via *Virtual Laboratories* [2]. This system of systems approach is enabled by D4Science [3,4].

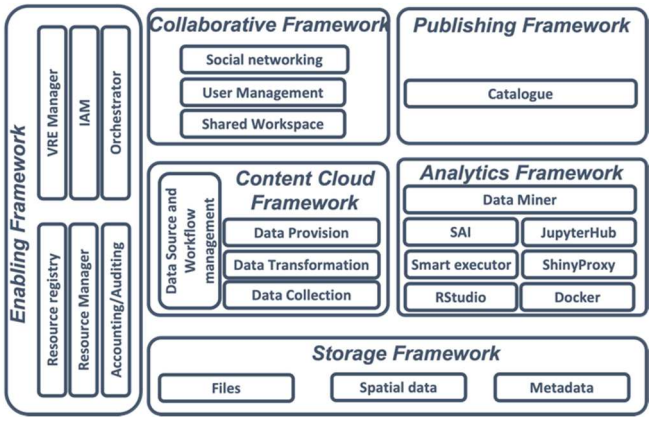

Fig. 3. Blue Cloud Virtual Research Environment Architecture

D4Science [3,4] is at the heart of the Blue Cloud platform. In fact, this service provider offers the core services to implement the resulting platform, namely: (*a*) the Blue Cloud VRE *Gateway*, realising the single access point to the rest of the platform; (*b*) the *authentication and authorisation infrastructure*, enabling users to seamlessly access the aggregated services once managed to log in the gateway; (*c*) the *Workspace*, for storing, organising and sharing any version of a research artefact, including dataset and model implementation; (*d*) the *Social Networking* area enabling collaborative and open discussions on any topic and disseminating information of interest for the com- munity, e.g. the availability of a research outcome; (*e*) the *Catalogue* recording the assets worth being published thus to make it possible for others to be informed and make use of these assets. These basic facilities are complemented by diverse facilities for *data analytics* including the DataMiner proprietary platform with its integration facility [3], JupyterHub, and a cluster of RStudio instances.

The overall architecture of the Blue Cloud VRE is depicted in Fig. 3. The *Enabling Framework* includes services required to support the operation of all services, VREs and VLabs. In particular, it includes (a) a resource registry service, to which all e-infrastructure resources (data sources, services, computational nodes, etc.) can be dynamically (de)registered and discovered by users and other services; (b) Identity and Access Management (IAM) services, as well as accounting/auditing services, capable of granting and tracking access and usage actions from users; (c) a VRE management framework for deploying specialized VREs/VLabs based on a selected subset of "applications". The *Storage Framework* includes services for efficient, advanced, and on-demand management of digital data represented by files in a distributed file system, collections of metadata records, and time series in spatially-enabled databases. The *Content Cloud Framework* includes services required to collect, transform, harmonize, and provide, via a variety of APIs, all metadata records published by the D4Science community and those provided by the organizations integrated by the D4Science consortium. The *Collaborative framework* supports all VREs and VLabs deployed and provides social networking services, user management services, shared workspace services, and Web-based User Interface access to the information cloud and to the analytics framework, via analytics laboratory services. The *Analytics Framework* includes the services required for executing analytics methods and processes provided by scientists. The *Publishing framework* includes services enabling users to document and make "public" any artifact, i.e., made available online.

This platform enacts an effective mechanism for Virtual Labs *co-creation* [4]. Every VLab is conceived to provide its designated community with the web-based working environment needed to accomplish their tasks by making seamlessly available the data, services and capacity by the "*as-a-Service*" delivery mode. Each enacts a family of scientific 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 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.

Every VLab consists of two complementary parts: the *community-agnostic part*, i.e. services offering basic or advanced functionality exposing an expected behaviour when instantiated in diverse contexts; the *community-specific part*, i.e. services offering a peculiar functionality or data, sometimes implemented by combining into specific workflows the community-agnostic part with context-specific services or data. The collaborative development of each VLab is supported by (*a*) an almost immediate delivery of every envisaged environment with the basic functionalities and features that are ready to use; (*b*) a dynamic and shared VLab development plan implemented by feature-oriented teams including both community members and D4Science engineers. Community members often implement the new features on their own by using the integration patterns supported [4]; (*c*) a DevOps strategy where new features are released, revised, and improved with rapid iterations.

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 the enabling framework and (part of) the other services. 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. This can be expanded with additional sites. Most of this physical architecture is invisible to the final users, which can 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 containerized applications, the computing clusters used to support high-throughput computing (HTC) tasks, and the worker clusters used to support map-reduce jobs - is located in several sites to ensure scalability, reliability and fault-tolerance.

V. Blue Cloud Virtual Labs

The Blue Cloud platform supports a series of designated communities manifested by selected demonstrators, each developing one or more VLab. All of them are made available by the Blue Cloud VRE https://blue-cloud.d4science.org/.

A. *Demonstrator #1 – Zoo- and Phytoplankton Essential Ocean Variable products*

This demonstrator implements a VLab to describe the current state of the plankton communities and forecasts their evolution, representing valuable information for the modelling, assessment and management of the marine ecosystems [5]. It is of interest for fundamental research (researchers and consultants from environmental agencies) contributing to the understanding of the environmental conditions and top-down factors at different scales of observations (regional/global, seasonal and time series). This knowledge will help the marine policy officers to address threats such as food insecurity, as foreseen under the EU Biodiversity Strategy for 2030. Moreover, fisheries advisory organisations can use these plankton products to study the availability of food resources for fish stocks.

B. *Demonstrator # 2 – Plankton Genomics*

This demonstrator implements a VLab supporting the discovery of unknown genes using the large dataset collected during the Tara Oceans Expedition [6]. Marine plankton is far more diverse than previously thought, with hundreds of thousands of genetically distinct taxa and more than 150 million genes documented. However more than half of the planktonic 'omic' sequences have still unknown taxonomy and/or function, especially in terms of sequences with eukaryotic origin. These unprecedented amounts of data on planktonic communities call for innovative data-driven methodologies to quantify and observe their biogeographic importance. The service allows retrieving unknown genes from annotation files for 4 different plankton size classes and then building gene clusters by similarities of sequences and larger metabolic pathways. The output file is used for the second service of this demonstrator: mapping the geographic distribution of plankton functional gene clusters using habitat prediction models. This service is designed for expert users with a strong genomic and bioinformatics background.

C. *Demonstrator # 3 – Marine Environmental Indicators*

This demonstrator implements a VLab offering a number of services for producing Marine Environmental Indicators [7] including: (i) *Marine environmental indicator generator* generating statistical analyses of the quality and characteristics of the marine environment for the Mediterranean Sea region, with possibility to scale up to the Global; (ii) *Ocean pattern and ocean regime indicators* to discover pattern/regime indicators based on machine learning and a simplified way to analyse oceanographic data; (iii) the *Storm severity index*, a quantitative impact modelling of severe wind/storms for different areas in the Mediterranean Sea region and for different time periods, up to 40 years (1979 - 2020).

D. *Demonstrator # 4 – Fish, a matter of scales*

This demonstrator manifests in two VLabs:

- Fisheries atlas of EU waters and beyond, a harmonised time-series of catch, commodities and trade data. This atlas is an expansion of the FAO Tuna Atlas, is scalable and offers to users features for data analysis using indicators, statistics, and interactive maps, accessible through a map Viewer, ISO/OGC metadata and data services, analytical and reporting tools and R Shiny, Jupyter and Markdown reporting services.
- Global record of stocks and fisheries (GRSF) [8]: a scalable and robust open data portal for fisheries data in EU waters and beyond, with a focus on assessment status and management of natural living resources. The GRSF contains Unique Identifiers of Stocks and Fisheries and the descriptions of the area and management structure. It is thus a natural companion of the Fisheries Atlas that has more catch location specific information.

E. *Demonstrator # 5 – Aquaculture monitor*

This demonstrator implements a VLab supporting the development of a tool to produce national aquaculture sector overviews whereby a country can make use of OGC compliant data services to monitor its aquaculture sector, not in an isolated way, but built on interoperable services where teams can compute and publish reproducible experiments [9]. An aquaculture monitor can become an important information source for local and national governments that lack the capacity to implement a national monitoring tool. It can thus provide a monitor for several Sustainable Development Goals such as 14 "Life below water", and 2 "Food Security".

VI. Conclusions

The Blue-Cloud pilot platform implements collaborative Open Science in marine research and provides solid cross-disciplinary assets in support of the EU Mission Ocean of a sustainable Blue Economy. It places itself as the EOSC Thematic Cloud for ocean science for the benefit of marine researchers and institutions. It develops a thematic marine EOSC for open web-based science, and serves needs of the EU Blue Economy, Marine Environment and Marine Knowledge agendas. The modular architecture of the VRE is scalable and sustainable, fit for connecting additional e-infrastructures, integrating more blue analytical services, configuring more Virtual Labs, and targeting broader (groups of) users. As part of the Blue-Cloud project also a Roadmap to 2030 is being prepared, including dialogues and surveys to major stakeholders from policy to research level. These confirm that the EU marine research community is becoming more aware of the Blue-Cloud potential and the role it can play in accelerating knowledge and science-based solutions to marine and ocean challenges. The initial Blue-Cloud project will be succeeded from January 2023 by the Blue-Cloud 2026 project which will focus on a further evolution of its pilot ecosystem into a Federated European Ecosystem to deliver FAIR [10] and open data, analytical services, instrumental for deepening research of oceans, EU seas, coastal and inland waters. Such an advanced ecosystem will also provide a core data service for the Digital Twin of the Ocean (DTO) which is under development.

ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Blue Cloud project (grant agreement No. 862409),

REFERENCES

[1] MW Maier. "Architecting Principles for Systems-of-Systems". *INCOSE International Symposium* 1996; 6(1): 565-573. doi: 10.1002/j.2334-5837.1996.tb02054.x

[2] L. Candela, D. Castelli, P. Pagano "Virtual Research Environments: an Overview and a Research Agenda" *Data Science Journal* 2013; 12: GRDI75-GRDI81. doi: 10.2481/dsj.GRDI-013

[3] M. Assante, L. Candela, D. Castelli, et al. "Enacting open science by D4Science". *Future Generation Computer Systems* 2019; 101: 555 - 563. doi: https://doi.org/10.1016/j.future.2019.05.063

[4] M. Assante, L. Candela, D. Castelli, et al. "Virtual Research Environments Co-creation: the D4Science Experience". Concurrency Computat Pract Exper. 2022;e6925. doi: 10.1002/cpe.6925

[5] P. Cabrera, S. Pint, L. Schepers, et al. "Developing zoo & phytoplankton EOV products in Blue-Cloud". Zenodo. 2021. https://doi.org/10.5281/zenodo.5896680

[6] P. Debeljak, A. Schickele, S.-D. Ayata, L. Bittner, J.-O. Irisson, F. Drago. "Exploring and mapping plankton genomics data with Blue-Cloud". Zenodo. https://doi.org/10.5281/zenodo.6224852

[7] L. Bachelot, K. Balem, F. Drago, M. Drudi, A. Garcia Juan "Applying machine learning methods to ocean patterns and ocean regimes indicators". Zenodo. https://doi.org/10.5281/zenodo.5896651

[8] Marketakis, Yannis, Ellenbroek, Anton, Gentile, Aureliano, & Drago, Federico. (2021). Federating knowledge on stocks and fisheries in Blue-Cloud with the GRSF. Zenodo. https://doi.org/10.5281/zenodo.5902487

[9] J. Augot, C. Fatras, E. Lavergne, et al. "Monitoring aquaculture activities through high-resolution satellite images". Zenodo. https://doi.org/10.5281/zenodo.5902386

[10] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18