D4Science: Advancing Ocean Science Through Collaborative Data Analysis

Massimiliano Assante (massimiliano.assante@isti.cnr.it), Leonardo Candela (leonardo.candela@isti.cnr.it), Luca Frosini (luca.frosini@isti.cnr.it), Francesco Mangiacrapa (francesco.mangiacraoa@isti.cnr.it), Elisa Molinaro (elisa.molinaro@isti.cnr.i), Pasquale Pagano (pasquale.pagano@isti.cnr.it)

Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche - CNR-ISTI (Italy)

Introduction

In the realm of ocean science, addressing intricate challenges necessitates collaborative analysis of extensive datasets. This underscores the significance of infrastructures that facilitate multidisciplinary collaboration, effective communication, and timely data sharing. D4Science [Assante et al., 2019], an operational infrastructure initiated 18 years ago with European Commission funding, has evolved into an efficient solution. Utilizing the "as a Service" paradigm, D4Science provides web-accessible Virtual Laboratories [Assante et al., 2023; Candela et al., 2023] (VLabs) that proved to be also suitable for ocean science collaboration [Schaap et al., 2022]. These VLabs simplify access to marine datasets, concealing underlying complexities. Key functionalities include a cloud-based Workspace for file organization, a platform for large-scale data analysis on a distributed computing infrastructure, a catalog for publishing research results, and a communication system based on social network practices.

D4Science has been actively supporting diverse marine and ocean science Virtual Laboratories (VLabs), adapting to evolving research needs. Notable initiatives include contributions to the European Open Science Cloud (EOSC), starting with the 'Blue-Cloud' project in 2020 and its subsequent extension, 'Blue-Cloud2026.' In 2015, D4Science played a pivotal role in the BlueBRIDGE Horizon 2020 Project, which aimed to provide user-friendly data services and tools for the aquaculture, fisheries, and environmental sectors. Additionally, in 2013, D4Science contributed to the iMarine FP7 Project, which has since evolved into the current iMarine initiative. This ongoing effort is dedicated to establishing and operating an e-infrastructure that aligns with the principles of the ecosystem approach to fisheries management and the conservation of marine living resources, further supporting the Food and Agriculture Organization's (FAO) Blue Growth Initiative.

D4Science is currently supporting over 20 scientific communities and over 150 VLabs, and pioneers Open Science in ocean research. It fosters collaboration, offers user-friendly environments, and provides service for accessing, sharing, analyzing, and publishing oceanographic data. A detailed description of these services is given in the following.

D4Science services overview for Ocean Science Virtual Laboratories

The D4Science services are instrumental in advancing Open Science practices within VLabs, empowering researchers to harness the advantages of state-of-the-art e-infrastructures. By leveraging these services, ocean science researchers can capitalise on the power of the Cloud and of e-infrastructures, driving scientific progress and enabling collaborative research efforts within the realm of Open Science.

The D4Science services offer a comprehensive array of features, fostering collaboration, facilitating data analytics, enabling result dissemination, and ensuring seamless integration with external systems. In fact, they cater to the entirety of the research lifecycle, providing diverse services. Specifically, *(i)* the *Collaborative Storage Framework* fosters collaboration among VLab users. The Workspace provides a platform for VLab members to collaborate, share resources,

and work together on projects. This collaborative environment enhances the efficiency and effectiveness of research activities within VLabs. In terms of data analytics, *(ii)* the *Analytics Engine Framework* empowers VLabs with powerful tools and resources, *(iii)* the *Publishing Framework* within VLabs facilitates the dissemination of research outcomes by means of the Metadata Catalogue and the Spatial Data Catalogue, that provide a means to organise and publish research results, making them accessible to the wider scientific community. This framework ensures transparency, reproducibility, and the sharing of valuable knowledge generated within VLabs.

These services are made available either by default or through specific requests. Every VLab can be equipped with:

- **Communication area** for collaborative and open discussions on any topic and disseminating information of interest for the community, for example, the availability of a research outcome;
- **Administration area.** User and groups Management dashboard for managing membership and roles;
- **Analytics Computing Framework area:**
  - The D4Science Analytics Engine, specifically developed to leverage advancements in IT and software engineering over the past decade where analytical methods can be integrated by means of the Container technology. The engine provides multiple execution infrastructures selectable by users, uses OGC API Processes (JSON) as a standard protocol, integrates with various CVS systems, and supports parallel executions. It offers many out-of-the-box methods, with a focus on flexibility and ease, provides JupyterLab and RStudio integrations, a web component-based front-end technology, and streamlined code generation tools.
  - **RStudio** allows users to perform online statistical analyses. Rstudio is no longer shared and it is now persistent. It offers a predefined list of R packages, and each VLab can define its RStudio servers configuration, Standard (4 cores/8GB RAM) and Large (8 cores/32GB RAM).
  - **JupyterLab**, a web-based interactive development environment for Jupyter notebooks, code, and data. It allows users to configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning.
- **Workspace cloud storage** for storing, organising and sharing items, such as datasets, scripts or outputs. This cloud-based solution provides two storage options: the Workspace and the Dataspace. Both can be accessed through RStudio and JupyterHub. The computations run in the analytical services can take inputs from the Workspace.
- **Spatial Data Infrastructure (SDI) services** provides users with the capability to store, discover, access, and manage vectoral and raster georeferenced datasets. The SDI exploits the following technologies: GeoServer equipped with PostgreSQL and PostGIS, GeoNetwork, Thredds.
- **Catalogue Framework** to document and publish any generated research product. Its primary component is the VRE Data Catalogue. The VRE Data Catalogue service is a catalogue service built on open-source technology for data catalogues (CKAN ckan.org). Via this Catalogue users can also search and browse data, products, and resources (posters, deliverables, etc) of interest from the Ocean Science community.

References

Assante M. et al., (2019). *Enacting open science by D4Science*. Future Gener. Comput. Syst. 101: 555-563, https://doi.org/10.1016/j.future.2019.05.063

Assante M. et al., (2023). *Virtual research environments co-creation: The D4Science experience*. Concurrency Computat Pract Exper. 2023; 35(18):e6925, https://doi.org/10.1002/cpe.6925

Candela L., Castelli D., and Pagano P., (2023). *The D4Science Experience on Virtual Research Environments Development*. In IEEE Computing in Science & Engineering, https://doi.org/10.1109/MCSE.2023.3290433

Schaap D., Assante M., Pagano P., and Candela L., (2022). *Blue-Cloud: Exploring and demonstrating the potential of Open Science for ocean sustainability*. IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea), Milazzo, Italy, pp. 198-202, https://doi.org/10.1109/MetroSea55331.2022.9950819