Blue-Cloud2026 project - Deploying BEACON data lakes for harmonizing ocean data access for Virtual Research Environments

The HE EOSC Blue-Cloud 2026 project aims at a further evolution of the pilot Blue-Cloud open science infrastructure 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 also strives to become a major data and analytical component for the Digital Twins of the Oceans (DTO's) as well as a blueprint for a thematic EOSC instance.

Within the project, a one-year long consultation and prototyping phase has resulted in concrete plans for establishing data subsetting capabilities in the Blue-Cloud ecosystem and implementing that vision by deploying a series of data lakes at the Blue-Cloud VRE for data repositories and data collections in support of VRE users and developers of the WorkBenches that aim to generate harmonised and validated data collections of Essential Ocean Variables (EOVs), **demonstrating true federated data access from 8 different sources**.

After consideration and positive tests on selected Blue Data Infrastructure (BDI) data collections, it was decided to embrace the open-source BEACON technology, as developed by MARIS, for deploying data lakes at the Blue-Cloud VRE for arranging data provision for WorkBenches 1 and 2 and also for VLabs.

BEACON is able to provide users with fast and easy access to multidisciplinary data originating from large collections, on the fly with high performance, and extract specific data based on the user's request. This software has been customised and deployed in the Blue-Cloud2026 project and several other European projects and is designed to return one single harmonised file as output, regardless of whether the input contains different data types. It allows everyone to set-up their own BEACON 'instance' to enhance the access to their data **or** use existing BEACON instances from well-known data infrastructures such as Euro-Argo or the World Ocean Database for fast and easy access to harmonized data subsets. More technical details, example applications and general information on BEACON can be found on the website https://beacon.maris.nl/.

In order to use BEACON for providing access to harmonised subsets for the WorkBenches, a set of in total 8 monolithic BEACON instances were set-up for relevant data collections such as the WOD, Copernicus Marine Cora, Euro-Argo, SeaDataNet, and more. The term 'monolithic' is used to indicate that the BEACON instance concerns one data collection or BDI. After initial configuration at a MARIS server, all 8 instances have been deployed operationally at the Blue-Cloud Virtual Research Environment (VRE). All have been integrated with the D4Science federated AAI service, by which access is arranged. All BEACON instances also have been provided with its own dedicated Jupyter-notebook which sits on the BEACON API and which makes it easier for the users to interact. The notebooks already contain several queries and users can adapt their own notebooks.

After considerable testing and feedback rounds on monolithic BEACON instances by members of the WorkBench teams, as a next step, two integrated BEACON instances have been initiated. These are merging data set collections from several monolithic BEACON instances. The challenge is to deliver harmonised collections, federated from the monolithic instances with different data models. For that purpose, a core metadata – data profile has been formulated by the WorkBench 1 and 2 teams which is populated by extracting and mapping from the monolithic instances, by retrieving and loading contents 'as-is' from monolithic BEACON instances into the merged BEACON instances, giving a common structure for variables, units, values, quality flags, and common metadata profile fields. The structured metadata and data will be supplemented by additional metadata data as available for each of the monolithic BEACON instances. For the merging use is made of the federation capabilities of the BEACON technology, while for the semantic mapping, use is made of the NERC Vocabulary Service, as well as the Semantic Analyser system of NOC-BODC. Any necessary unit conversion rules can be retrieved and applied using factors and offsets as listed.

This presentation will cover an introduction of the Blue-Cloud 2026 project and developments of the BEACON instances, explaining how it can practically serve as data lakes for many VRE applications and how it is extendable to other domains. By using examples from the WorkBenches demonstrating the federated data access - a huge challenge in many domains -, the reduction in time and effort spent for the researchers to collect the data are highlighted.