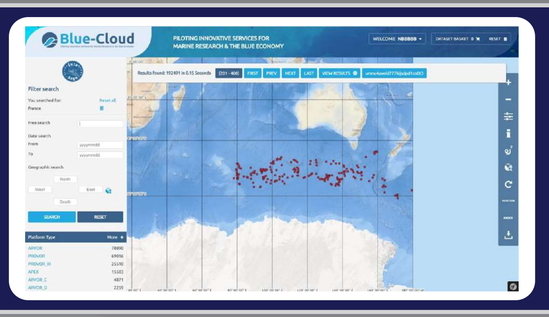
eosc | Blue-Cloud2026

BLUE-CLOUD DD&AS: Federated discovery and access to a wealth of data from leading Blue Data infrastructures

Scan the QR Code to visit our website **blue-cloud.org**

The Blue-Cloud Data Discovery and Access service is one of the two main components of the Blue-Cloud technical framework, next to the Blue Cloud Virtual Research Environment (VRE). It facilitates federated discovery and retrieval of data sets and data products, managed in leading Blue Data Infrastructures (BDIs).

A common interface is provided, and including facilities for mapping and viewing the locations of data sets, aspart of the query dialogue. Moreover, the interface has a shopping mechanism, facilitating users to compose and submit mixed shopping baskets with requests for data sets from multiple BDIs.



Scan to visit **data.blue-cloud.org**

Currently, the following BDIs are federated in the Blue-Cloud Data Discovery and Access service

Blue Data Infrastructure	Types of data sets	Logo and link
SeaDataNet CDI service	Marine physics, bathymetry, chemistry, geology, geophysics, and biology observation data sets	
EMODnet Chemistry data products	Marine chemistry data collections and interpolated map products	
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)	Nucleoticle sequencing data and information on marine species	
EcoTaxa	Taxonomic annotation data of images on planktonic biodiversity	
SeaDataNet data products	Aggregated marine data collections and climatologies, such as for Temperature & Salinity	
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	
EMODnet Bathymetry	EMODnet Bathymetry World Base Layer is used as base map in the interface	
EMSO		EMSO
SIOS		SIOS

Use is made of web services and APIs, following protocols such as CSW, OAI-PMH, ERDDAP, or otherwise, as provided and maintained by the BDIs. These are used to deploy machine-to-machine interactions for harvesting metadata, submitting queries, and retrieving resulting metadata, data sets and data products.

The query mechanism has a two-step approach:

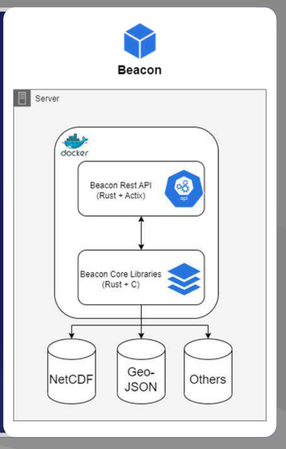
- The first step enables users to identify interesting data collections, with free search, geographic and temporal criteria as main query operators;
- The second step enables users to drill down per interesting BDI to get more specific data sets, but this time at granule level, and including additional search criteria, specific for a selected BDI;
- Finally, users are able to compose and submit shopping requests at the granule level and to retrieve the data sets by downloading from their MyBlueCloud dashboard.

BEACON: High performance data lake for sub-setting of big data sets

Blue-Cloud is also developing and deploying data sub-setting and extracting services, in addition to discovery and access, and building Blue-Cloud Data Lakes for use by Blue-Cloud WorkBenches, Virtual Labs, and beyond. In order to achieve this, the new Beacon data lake technology (© MARIS) is implemented.

Beacon

The Beacon software system comes with a unique indexing system that can, on the fly, extract specific data based on the user's request from millions of observational data files. . Beacon exposes a REST API so that clients can query data via a simple JSON request, integrated in Jupyter notebooks. The system returns one single harmonised file as output, regardless of whether the input contains many different data types or dimensions. In practice, Beacon functions as a data lake bringing together millions of e.g. NetCDF files from multiple repositories, and after initial indexing (taking several hours), allowing extraction of subsets and exporting these in one coherent NetCDF file in seconds.

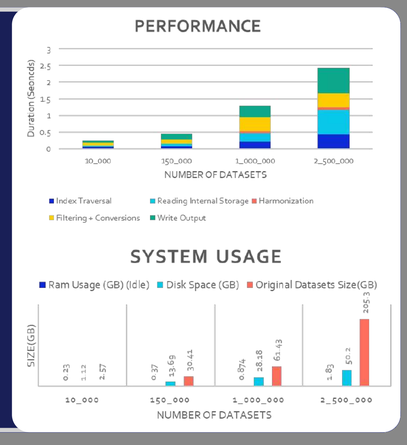


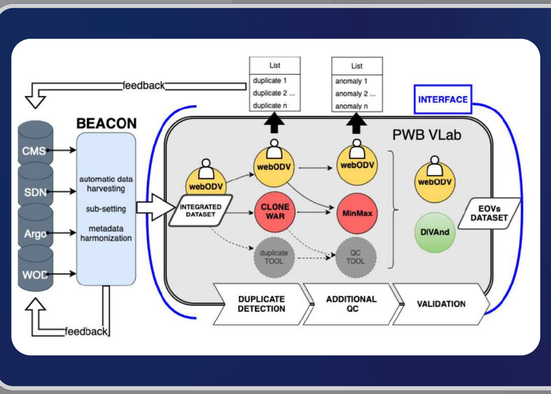
Performance

Performance of Beacon queries for the case with an increasing number of SeaDataNet NetCDF files loaded into Beacon instance. We loaded into Beacon all SeaDataNet CDI records (2.5 million datasets), constituting to a 200GB NetCDF Data Query with filters; Longitude from -8 to 12, Latitude from 50 to 61, Depth from 0 to 50, Time period from 2010 to 2012. All the temperature parameters were aggregated and harmonized in degrees Celsius --> Result: 12M points!

System Usage

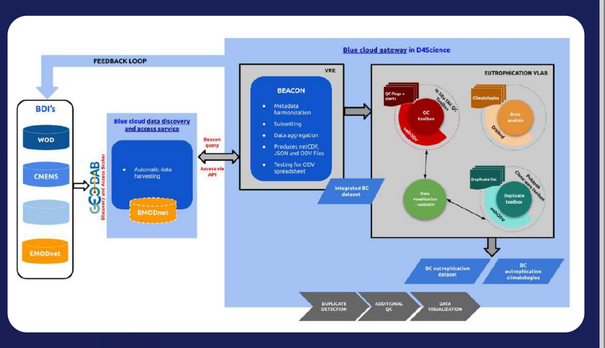
Beacon can run on laptops, home PC's and servers and only uses what's necessary to process the query.





WorkBenches

Blue-Cloud is establishing three big data processing WorkBenches to facilitate the generation of validated and harmonised data collections for selected Essential Ocean Variables (EOVs), namely for Physics, Chemistry and Ecosystems. Several datasets from different EU and non-EU BDIs will be integrated, harmonised and validated. The resulting high-quality EOVs datasets and analytical workflows will be instrumental to EU operational services and the Digital Twins of the Oceans (DTO).
WorkBenches are looking for an easy and efficient way to get their input organised of large amounts of data sets for selected parameters from multiple data repositories and preferably already in a homogeneous way considering formats, parameters, units and other core metadata. The Figures illustrates a schematic overview of the Physical andEutrophication WorkBenches with Beacon integrated in the workflow.



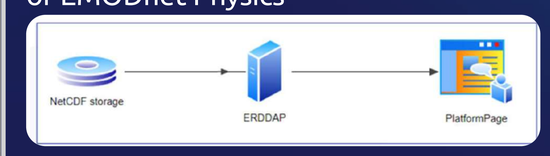
Benchmark

A total of 2.174.403 NetCDF files, amounting to 2.5TB of data, were ingested. The NetCDF files concerned monthly data for each platform, totaling approximately 75,000 platforms (multiple files per platform).The Beacon data lake currently utilises 127GB of disk space. The importing phase was successfully completed in just a couple of days, whereby the ingesting of data is streamlined with the user-friendly Beacon API.

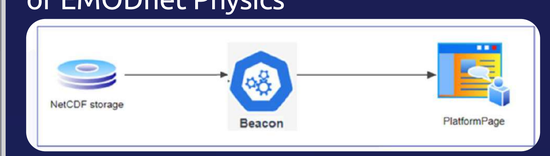
Various types of queries were executed to test performance:

As a result, Beacon was found to be 10 times faster than ERDDAP (sometimes even more).

Ingestion through ERDDAP as intermediary between the NetCDF storage and the Graphics engine and presentationwindow of EMODnet Physics



Ingestion through Beacon as intermediary between the NetCDF storage and the Graphics engine and presentationwindow of EMODnet Physics



Overview of tests performed for situation with ERDDAP versus situation with Beacon

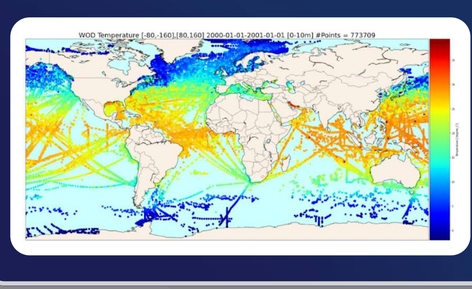
- Filtered by **dataset ID** (to select a single platform)
- Filtered by **metadata** (to select a single platform)

To retrieve: Time, Latitude, Longitude, Parameter, Parameter _qc values

Scan to visit **beacon.maris.nl**

Beacon Instances

Two main use cases have been formulated:
A number of monolithic Beacon instances, each for selected data from a specific BDI, and made available for all Blue-Cloud VRE users.
Two integrated Beacon instances for WorkBench 1 respectively WorkBench 2, merging and harmonising selected data from multiple BDIs, and with regular and controlled updating mechanism, in connection with the DD&AS, and made available to the 2 WorkBenches.



Notebooks

There are currently eight demonstrator notebooks available on GitHub and the Blue-Cloud VRE connected to the Beacon instances. The Figure illustrates observations from the World Ocean Database (WOD) Beacon instance.

Funded by the European Union