



Blue-Cloud2026

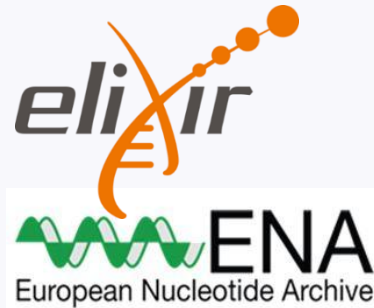
Federated Data Discovery & Access Service and high-performance Datalake for sub-setting of big data sets

Dick Schaap (MARIS)
Robin Kooyman (MARIS)

On behalf of BC2026 team

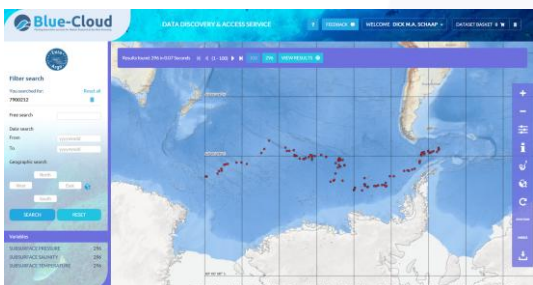
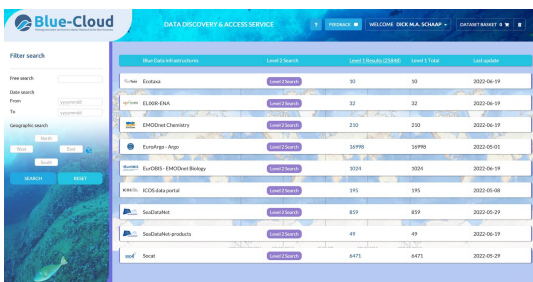
Blue-Cloud Federation Workshop 6 Nov 2024, Lisbon - Portugal





Blue Data infrastructures

E-infrastructures



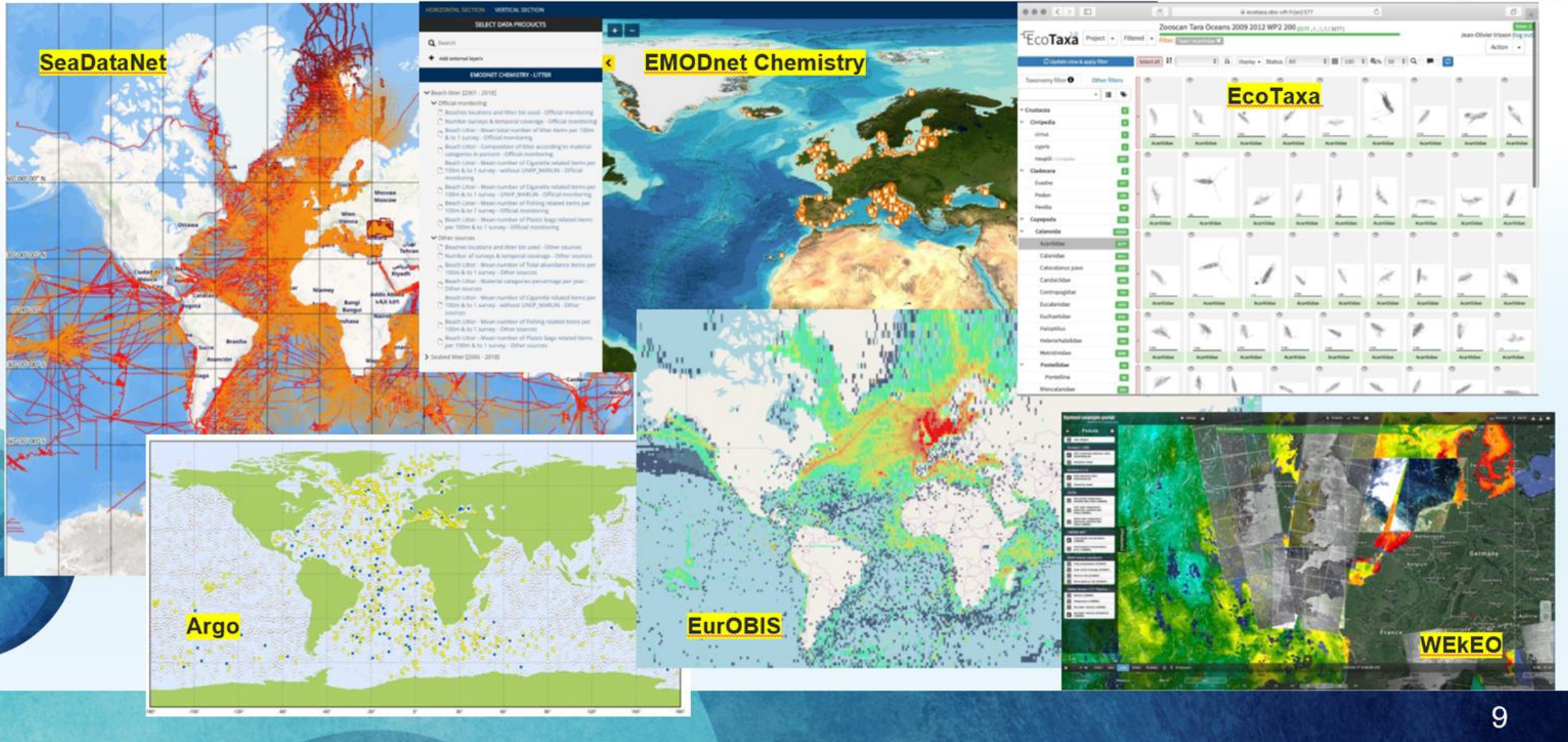
• **Facilitates Blue-Cloud users:**

- Federated search for discovering interesting data sets (currently more than 10 million) in a two step approach
- Federated retrieval of identified data sets using a shopping basket mechanism
- Download of data sets or push to Blue-Cloud VRE

• **Facilitates managers of Blue Data Infrastructures:**

- Stay informed about data requests and users for their repository
- Periodic reporting of downloads from their repository

• **Facilitates trying out the federation concept with the web services as offered by the Blue Data Infrastructures**



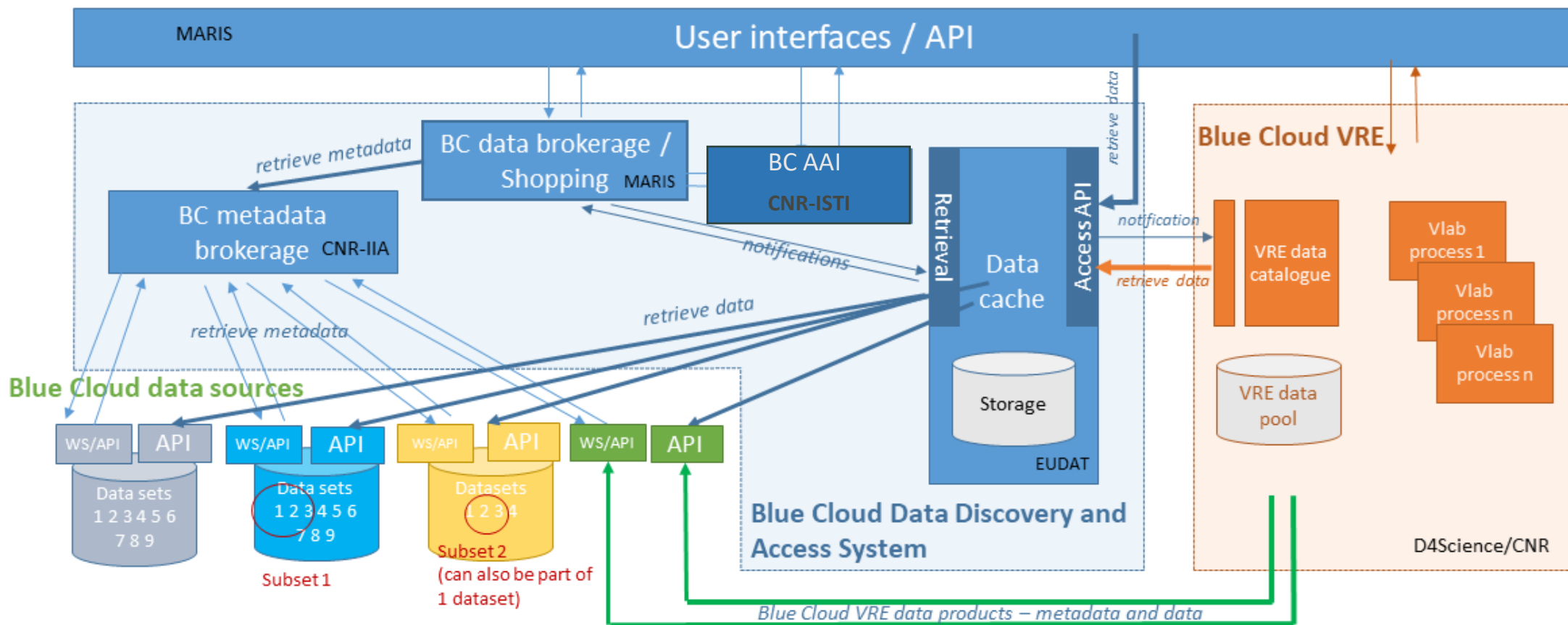
Federated discovery and retrieval of data sets and data products from the Blue Data Infrastructures

Concept of two-step search approach:

First step: identifying interesting data collections and products with few criteria

Second step: drilling down with more criteria to select specific data at granule level, where possible, otherwise at collection/products level

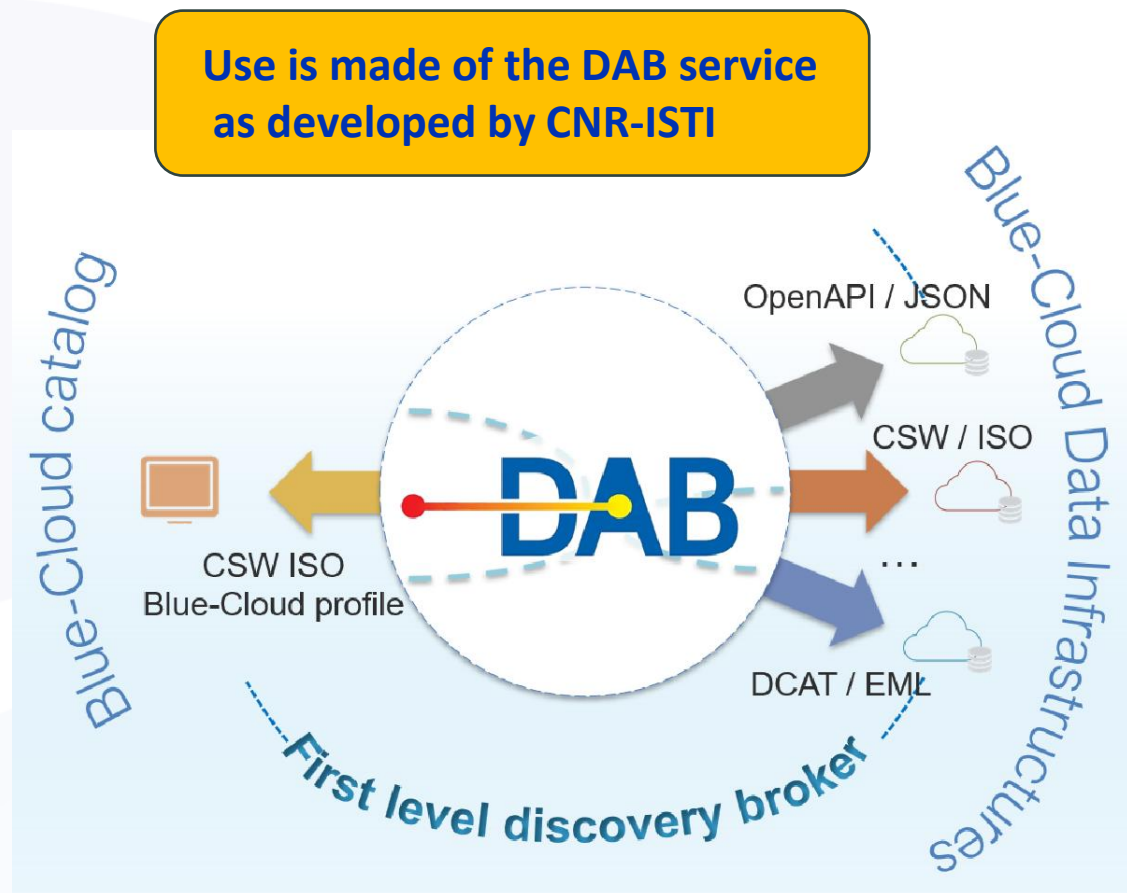
Metadata and Data Brokerage services interacting **Machine-to-Machine** with web services and APIs as provided and operated by the Blue Data Infrastructures



The common Blue-Cloud metadata elements are:

- IDENTIFIER: Blue-Cloud unique and persistent code for the metadata record
- TITLE: a characteristic, and often unique, name by which the collection is known
- ABSTRACT: a short description of the collection
- KEYWORD: a commonly used word, formalised word or phrase used to describe the subject
- BOUNDING_BOX: extent of the resource in the geographic space given as a bounding box
- TEMPORAL_EXTENT: time period covered by the content of the collection
- PARAMETER: name of the attribute described by the measurement value
- INSTRUMENT: measuring instrument used to acquire the data
- PLATFORM: platform from which the data were taken
- ORGANIZATION: organization associated with the collection
- DATESTAMP: the latest update date of the metadata description
- REVISION_DATE: the latest update date of the data
- RESOURCE_LINKS: download links where available and useful

DAB Service endpoint (global): <https://blue-cloud.geodab.eu/gsservice/services/essi/view/blue-cloud/csw>



SeaDataNet	Dedicated API
SeaDataNet Products	OGC CSW service
EMODnet Chemistry	OGC CSW service
EuroArgo - Argo	Dedicated API
EurOBIS – EMODnet Biology	DCAT service
Ecotaxa	Dedicated API
ELIXIR - ENA	Dedicated API
ICOS Marine	SPARQL service
SOCAT	ERDDAP service

The image displays a series of overlapping screenshots of the Blue-Cloud Data Discovery & Access Service web application. The top-most screenshot shows the main search interface with a navigation bar at the top containing the Blue-Cloud logo, the service name, a user welcome message ('WELCOME DICK M.A. SCHAAP'), and a dataset basket. The search filters on the left include 'Free search', 'Date search' (From/To), and 'Geographic search' (North/South, West/East). The search results section shows 'You searched for: EuroArgo - Argo' and a map of the Atlantic Ocean with 296 data points marked as red dots. A status bar above the map indicates 'Results found: 296 in 0.07 Seconds' and includes navigation controls. Below the map, a table lists the variables: SUBSURFACE PRESSURE (296), SUBSURFACE SALINITY (296), and SUBSURFACE TEMPERATURE (296). The interface is clean and modern, with a blue and white color scheme.

<https://data.blue-cloud.org>

In **Blue-Cloud 2026** activities are ongoing for expanding and optimising the **Blue-Cloud Data Discovery & Access service (DD&AS)** and its FAIRness by:

- harmonising and expanding functionality of web services as operated by each BDI for discovery and access of managed data resources
- expanding the DAB Common Metadata Profile with extra metadata fields
- developing and deploying **semantic brokering**
- federating additional BDIs into the DD&AS (**EMSO, SIOS, EMODnet Physics, ELIXIR-Mgnify**)

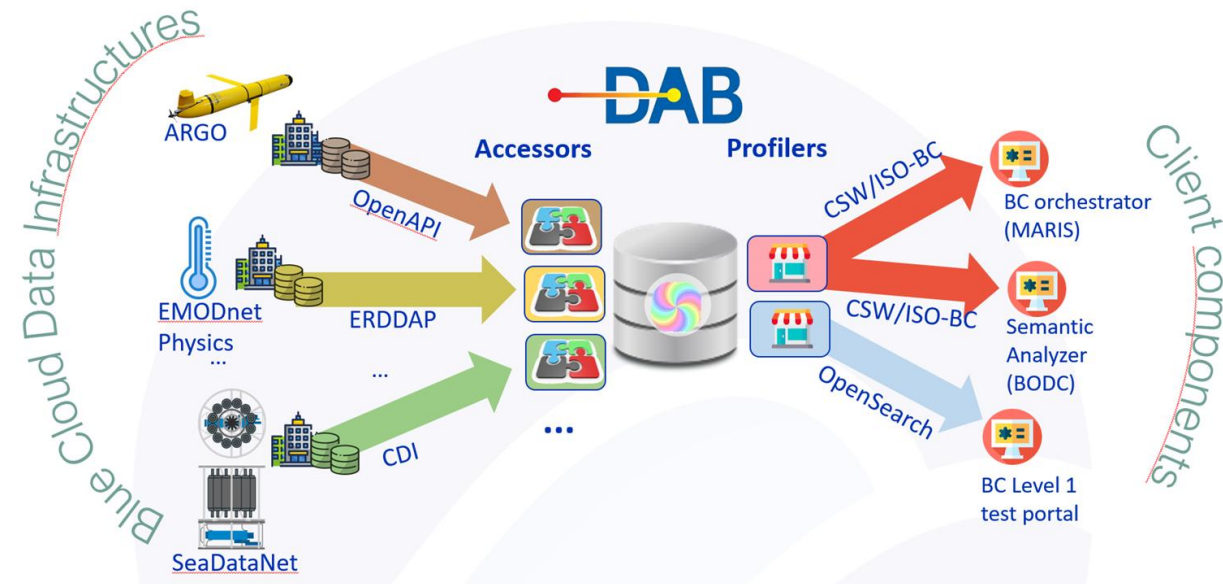


MGNify



The target common Blue-Cloud metadata elements are:





- IDENTIFIER: Blue-Cloud unique and persistent code for the metadata record
- TITLE: a characteristic, and often unique, name by which the collection is known
- ABSTRACT: a short description of the collection
- **KEYWORD**: a commonly used word, formalised word or phrase used to describe the subject
- BOUNDING_BOX: extent of the resource in the geographic space given as a bounding box
- TEMPORAL_EXTENT: time period covered by the content of the collection
- **PARAMETER**: name of the attribute described by the measurement value
- **INSTRUMENT**: measuring instrument used to acquire the data
- **PLATFORM**: platform from which the data were taken
- **ORGANIZATION**: organization associated with the collection
- **ORGANIZATION ROLE**: role of the cited organization
- **PROJECT**: project associated with the collection
- **CRUISE**: cruise associated with the collection
- DATESTAMP: the latest update date of the metadata description
- REVISION_DATE: the latest update date of the data
- RESOURCE_LINKS: download links where available and useful

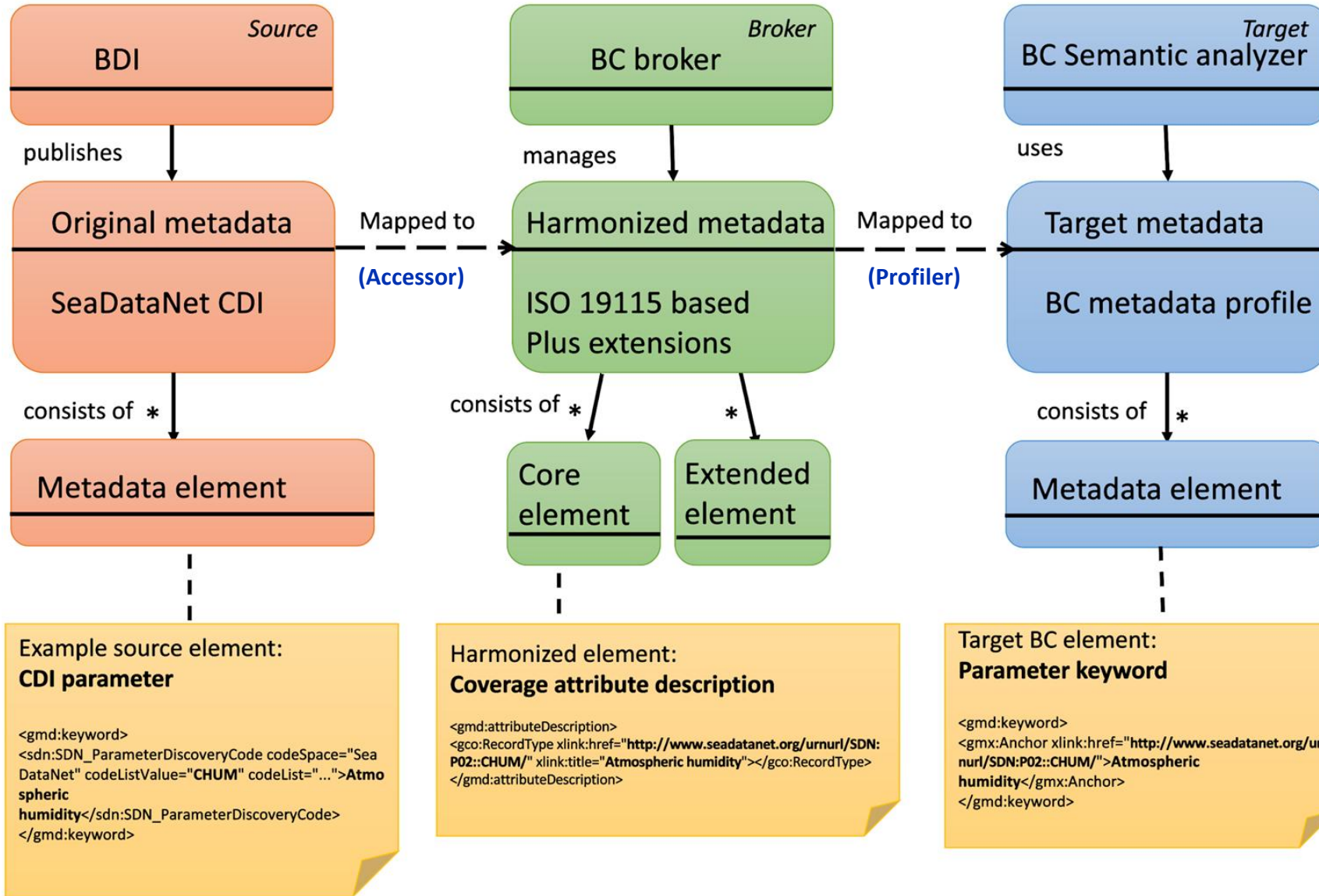


Adding URIs to Vocabularies

Additional metadata elements

DAB coupling to Semantic Analyser

Metadata element	Definition	Example
Instrument	Instrument or sensor used in the observation	Atomic absorption spectrometers
Instrument URI 	URI for unambiguously lookup and interpret an instrument Example vocab: L05, ...	http://vocab.nerc.ac.uk/collection/L05/current/LAB10/
Platform	Platform used in the observation	Polarstern
Platform URI 	URI for unambiguously lookup and interpret a platform Example vocab: C17, ...	http://vocab.nerc.ac.uk/collection/C17/current/06AQ/
Organization name	Name of a cited organization	Marine Information Service (MARIS)
Organization Role 	Role of the cited organization	Provider
Organization URI 	URI for unambiguously lookup and interpret an organization Example vocab: EDMO, ...	https://edmo.seadatanet.org/report/634
Datestamp	Metadata time stamp	2024-01-01
Revision date	Data time stamp	2024-01-01



semantics.bodc.ac.uk

British Oceanographic Data Centre

Metadata Sources **File Sources**

Source

Select a source from the list...

- ARGO
- Copernicus Marine Environment Monitoring Service (CMEMS)
- ELIXIR-ENA
- ELIXIR-MGnify
- EMODnet Chemistry
- EMODnet Physics
- EMSO ERIC
- EurOBIS
- European Environment Agency SDI Catalog
- European Marine Observation and Data Network (EMODnet)
- ICOS Data Portal
- ICOS SOCAT
- Joint Research Centre Data Catalog
- SeaDataNet - Open datasets
- SeaDataNet products
- Svalbard Integrated Arctic Earth Observing System (SIOS)
- US NODC Collections
- VITO /Copernicus Global Land Services
- WEkEO

British Oceanographic Data Centre

Metadata Sources

File Sources

File type

NetCDF XML TXT

Select a text file containing terms to parse. There should be one term per line. (Max number of terms: 300)

Choose File cora_probe_type.txt

Parse Text file

Showing results 1 to 13 of 15

Bottles
 CTD
 XBT/MBT
 Gliders
 Profilers
 Sail drones
 Thermistor chains
 Drifters
 Ferry boxes
 Thermo salinographs
 Moorings
 Animal mounted
 Other/Unknown

< 1 >

Terms analysis

Select a list to vocabularies

Select...

Leave empty to select from all

Match Properties

Identifier Preflabel Alllabel Definition

Exclude Deprecated Terms

Analyse terms (1 to 13)

(Please note that analysis is currently limited to 300 terms max)



- Through the federation, an analysis is made of the availability and presence of core metadata elements, as well as their use of semantics, in the output of the web services of each BDI; these results are reported by an interface to the BDI managers **for improvement**
- Used vocabularies and ranges, currently in use, are analysed and a **semantic brokerage** is deployed for mapping towards the target semantics
- Where missing, each BDI is requested to adopt using vocabularies, where possible, with a focus on uptake of SeaDataNet standards:
 - BODC – SeaDataNet controlled vocabularies
 - EDMO (organisations)
 - EDMERP (projects – programmes)
 - CSR (Cruise Summary Reports)
- As alternative, free text terms are analysed by the semantic analyser and mapped, where possible
- **As a result BDIs are becoming more FAIR in web services and contents, enabling their federation to become more streamlined for a rich and harmonised discovery, publishing, and access**



Physical Workbench

Implement a cloud-based workflow to generate harmonised, validated and customisable **EOV data collections for temperature and salinity** on Mediterranean sea.



Eutrophication Workbench

Define and implement an efficient production workflow to merge multi-source datasets and build highly qualified **EOV datasets for eutrophication variables: chlorophyll, nutrients, oxygen** on NorthEast Atlantic Sea.



Ecosystem Workbench

Improve the availability, quality and interoperability of large collections of plankton observations. Develop an analytical workflow using ML to produce global intercomparable **plankton biodiversity and biomasses** maps & products with clear QC protocols.



Heterogeneous, incomplete data

- Different data types & observation methods
- Data of various quality
- Duplicates of data in repositories

Data access challenges

- To access the latest version of (subset of) large datasets from diverse, distributed data repositories

Data harmonisation & semantic challenges

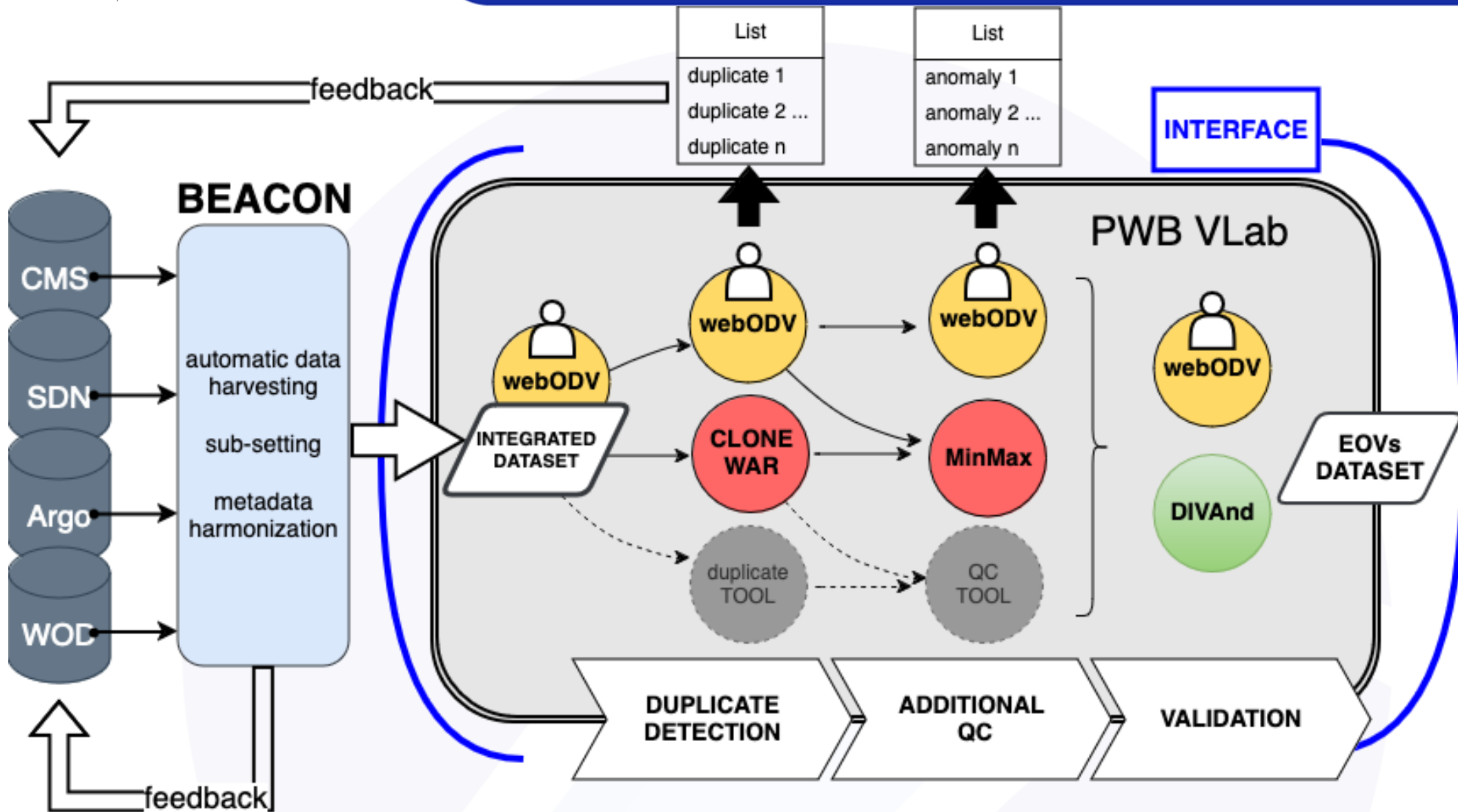
- To merge different datasets of different types, with different metadata and handle duplicates

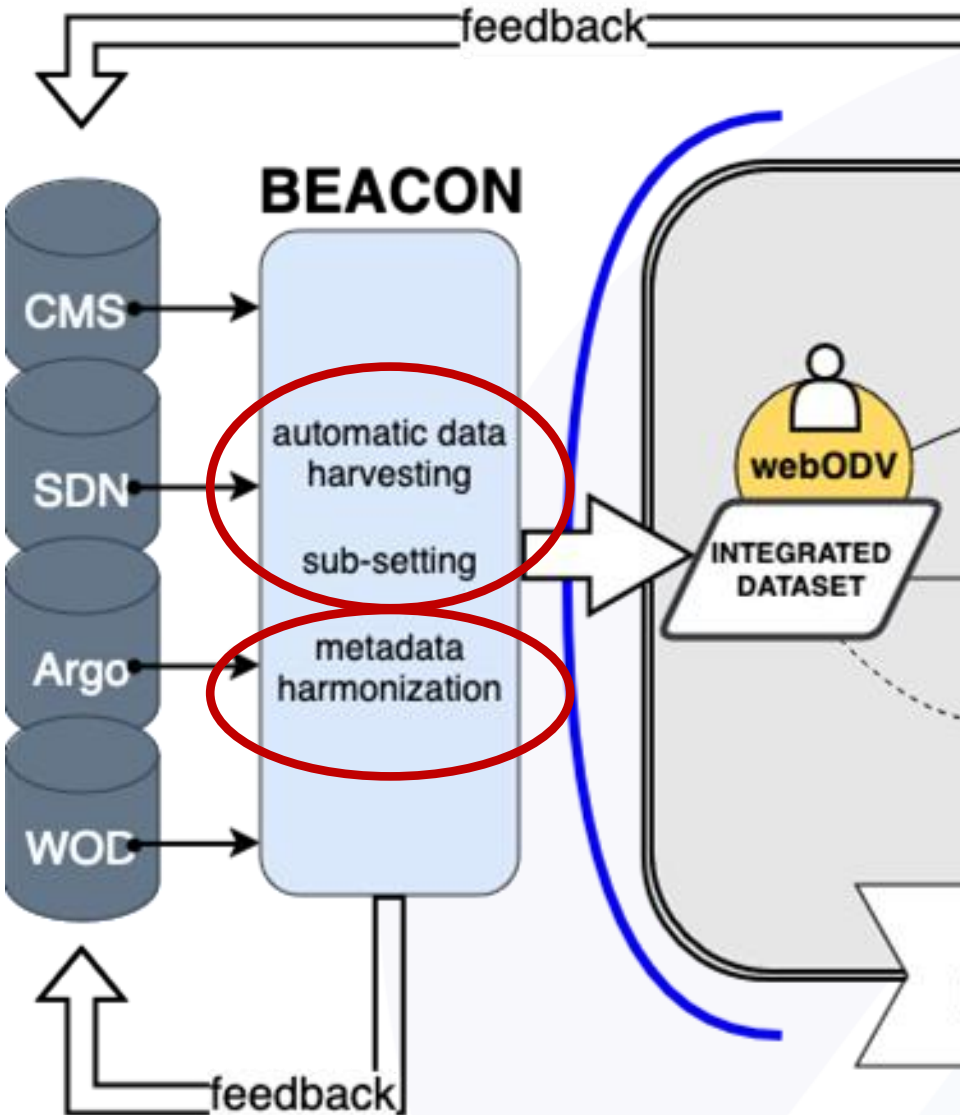
Data computing challenges

- To run Machine-Learning methods and analytical pipelines seamlessly regardless of the user IT resources

**Collaborative, integrated approach with Blue-Cloud 2026 services:**

- DD&AS + BEACON + Semantic Analyser:
 - To meet data access & harmonisation challenges (catalogue, Metadata harmonisation, Subsetting)
- D4Science:
 - To meet computing & processing challenges (HPC, cloud-optimized services)
 - To meet collaborative challenges (VRE)





Beacon

High-Performance data lake solution for fast big data harvesting and subsetting enabling efficient data consumption in the VRE

Semantic Analyzer

it allows semantic harmonization focusing on key metadata elements

- Platforms
- Instruments
- Parameters
- Units
- ...



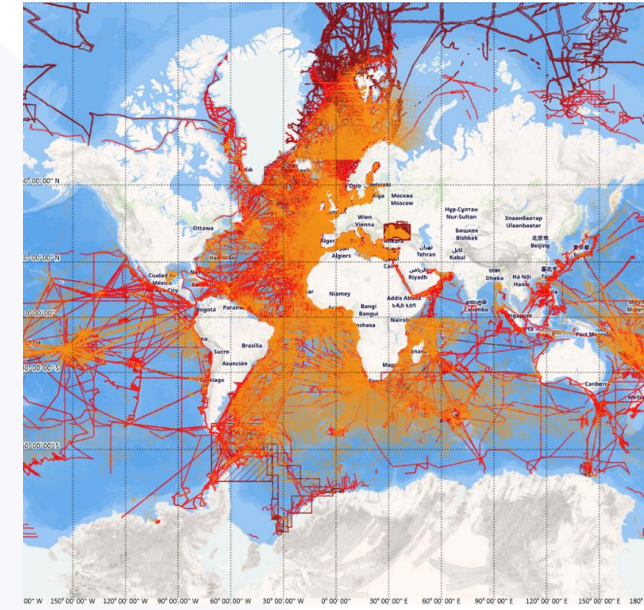
Challenge:

- Repositories can contain millions of data files
- How to optimize **Machine2Machine access to subsets**, enabling easy access to Jupyter Notebooks and other applications.
- **How to go from files to serving applications as an actual “Data lake”?**

Example:

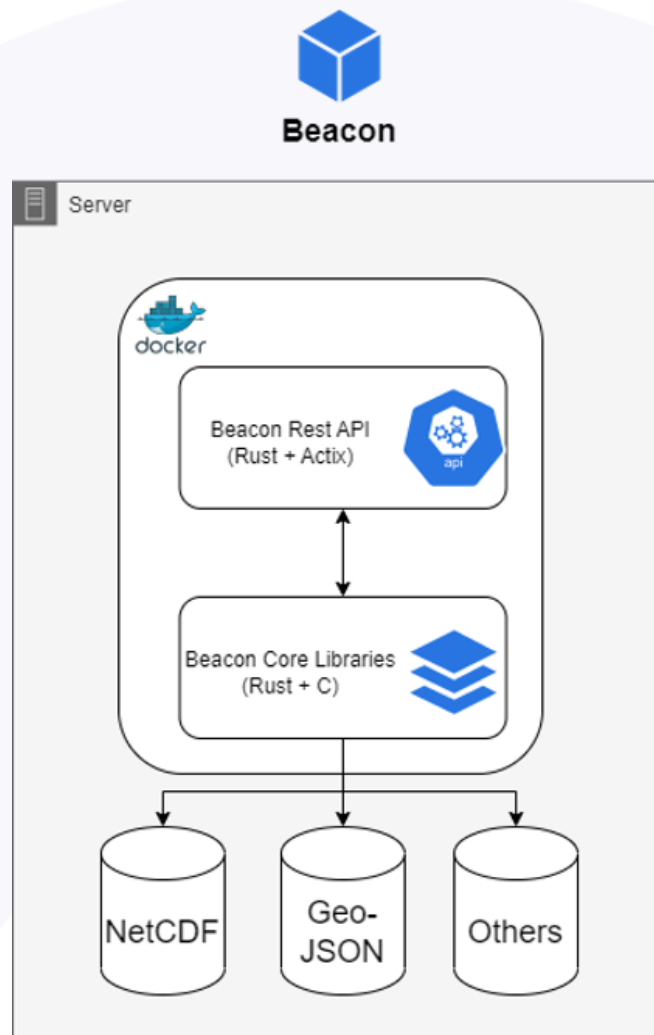
- Request: ‘Give me all the temperature data in the North Sea, from 2010-2020, in degrees Celsius, at a depth of 0-50 m’.
- Response: One NetCDF file containing exactly this data, which is then on the fly, directly usable in a Jupyter notebook and for HPC.

SeaDataNet CDI



BEACON is developed to provide an easy-to-use, fast, reliable, and scalable solution for storing, processing, and retrieving data from large amounts of data files

- **Written in Rust + C**
- **High Performance Data Lake**
- **Runs on:**
 - Linux
 - Windows
 - MacOS
 - Docker Containers
- **Consists of:**
 - Rest API
 - Core Libraries
 - Real-Time sub-setting
 - Data harmonization (single output file)
 - Dynamic Chunking
- **Produces different output formats:**
 - NetCDF
 - CSV
 - JSON
 - GeoJSON
 - IPC (Apache Arrow)
 - Parquet
 - WebODV ASCII



- **Handle any NetCDF Structure (E.g. Timeseries, Cruises, Gridded)**
- **Powerful query capabilities**
- **Filter on:**
 - Ranges
 - Polygons
 - Metadata
 - Union/Aggregation Queries
 - Federated Queries

Performance

Loaded into BEACON all SDN CDI records:

- 2.5 millions datasets
- > 4 billion data points
- 200GB of NetCDF Data

Query:

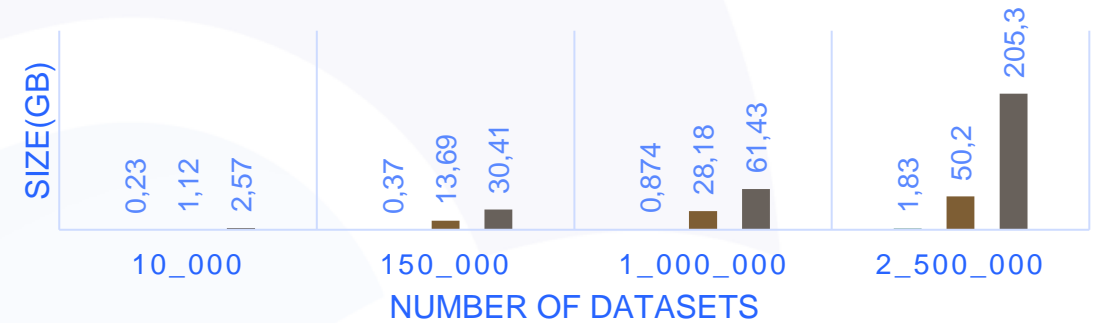
- Longitude from -8 to 12
- Latitude from 50 to 61
- Depth from 0 to 50
- Time from 2010 to 2012
- All the temperature parameters aggregated and harmonized in degrees Celsius
- Result: 12M points!

Beacon System Usage

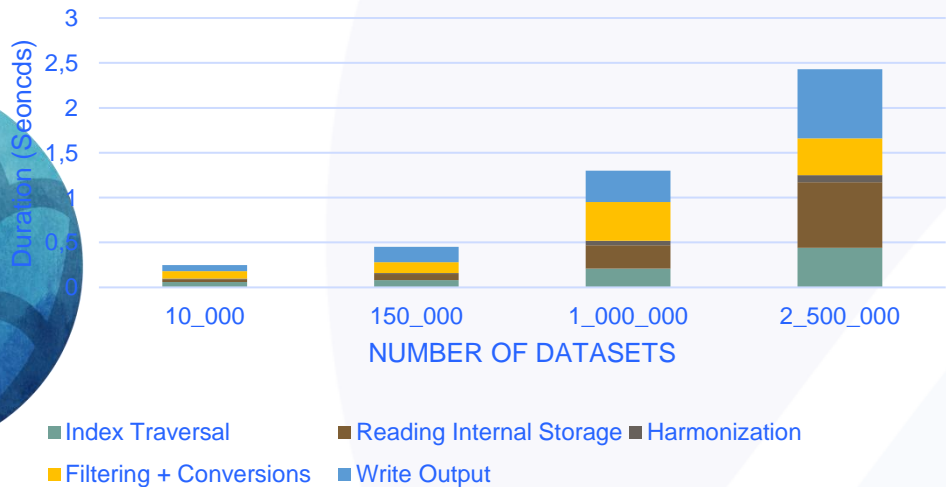
- Can run on Laptops, Home PC's and Servers
- Only uses what's necessary to process the query

SYSTEM USAGE

- Ram Usage (GB) (Idle)
- Disk Space (GB)
- Original Datasets Size(GB)



PERFORMANCE



Beacon ERA5

11TB of Data
 11K daily 0.05 gridded SST datasets.
 260 Billion measurements
 Resource Usage:
 1.2GB of RAM,
 250GB of Disk Space

Beacon Argo Data

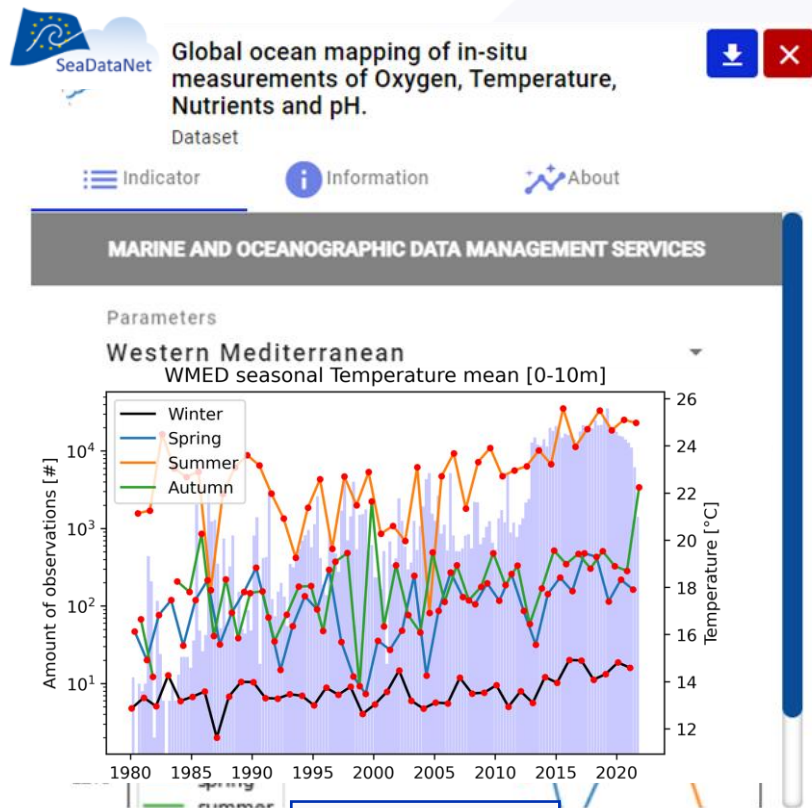
600GB of NetCDF Data.
 3.3M datasets.
 Resource Usage:
 600MB of RAM
 50 GB of Disk Space

Beacon SeaDataNet

250 GB of NetCDF Data.
 2.8M datasets.
 Resource Usage:
 1.8GB of RAM
 60 GB of Disk Space

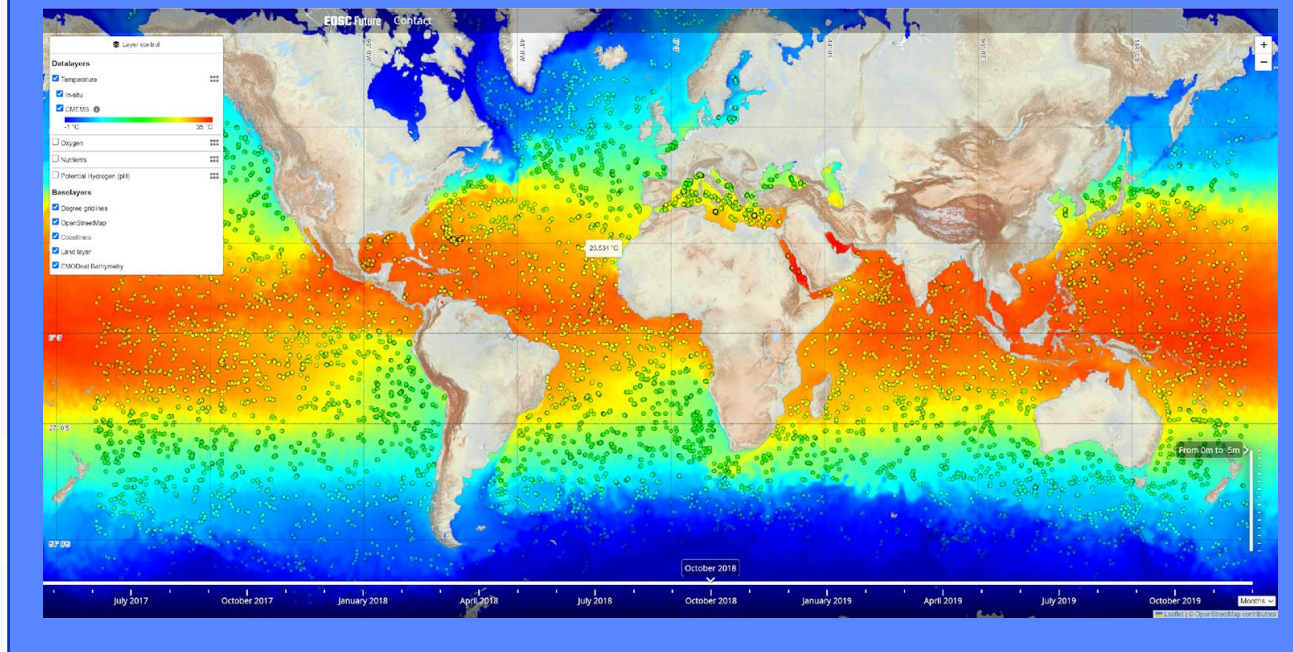
Beacon as data lake and engine, giving instantaneous access to data values from **EuroArgo-Argo** and **SeaDataNet**, co-located with **Copernicus Marine** data products

Level 1: Ocean indicators



Map Viewer

Level 2: Co-location data values as-is Map Viewer



Sliding by depth, time period, parameters, and geo location

Notebook implementation example

```

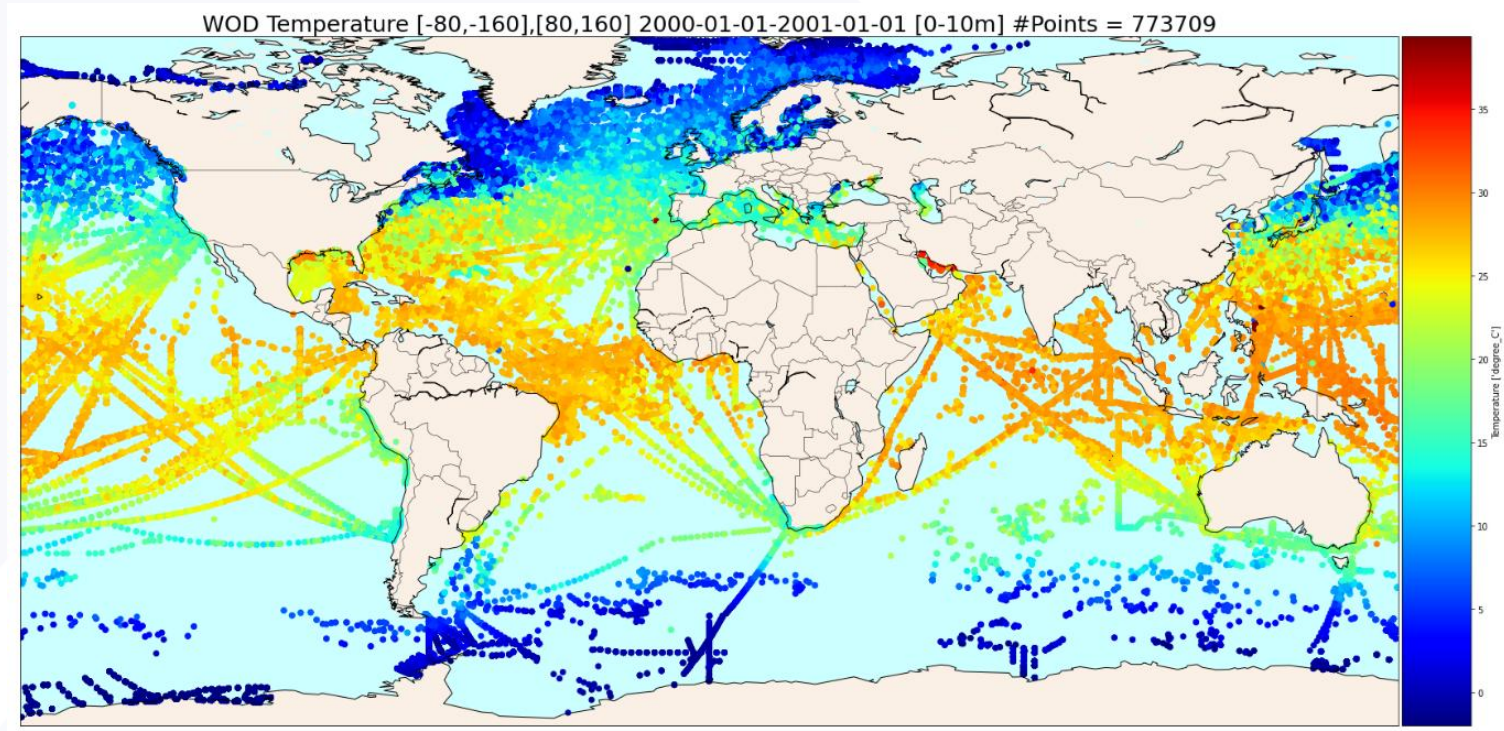
body = {
  "query_parameters": [
    {
      "column_name": "parameter",
      "alias": "parameter"
    },
    {
      "column_name": "time",
      "alias": "TIME"
    },
    {
      "column_name": "z",
      "alias": "DEPTH"
    },
    {
      "column_name": "lon",
      "alias": "LONGITUDE"
    },
    {
      "column_name": "lat",
      "alias": "LATITUDE"
    },
    {
      "column_name": "dataset",
      "alias": "DATASET",
      "optional": True
    },
    {
      "column_name": "WOD_cruise_identifier",
      "alias": "cruise-identifier",
      "optional": True
    },
    {
      "column_name": "wod_unique_cast",
      "alias": "cast",
      "optional": True
    },
    {
      "column_name": "WMO_ID",
      "alias": "WMO_ID",
      "optional": True
    },
    {
      "column_name": "country",
      "alias": "country",
      "optional": True
    }
  ],

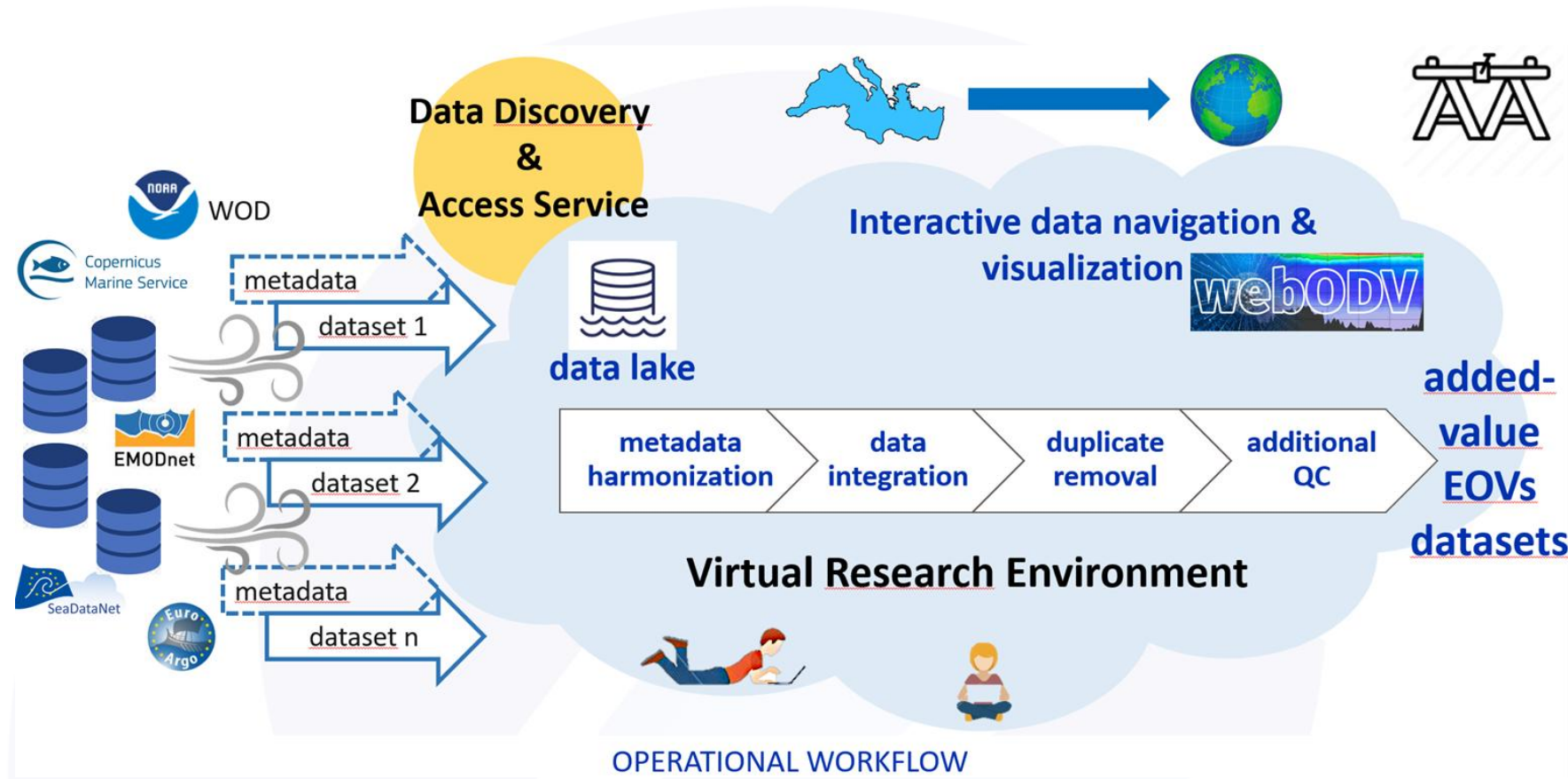
```

```

"filters": [
  {
    "for_query_parameter": "TIME",
    "min": mintemporal,
    "max": maxtemporal
  },
  {
    "for_query_parameter": "DEPTH",
    "min": mindepth,
    "max": maxdepth
  },
  {
    "for_query_parameter": "LONGITUDE",
    "min": minlon,
    "max": maxlon
  },
  {
    "for_query_parameter": "LATITUDE",
    "min": minlat,
    "max": maxlat
  },
  {
    "for_query_parameter": "parameter",
    "min": -2,
    "max": 40
  }
],
"output": {
  "format": "netcdf"
}
}
return body

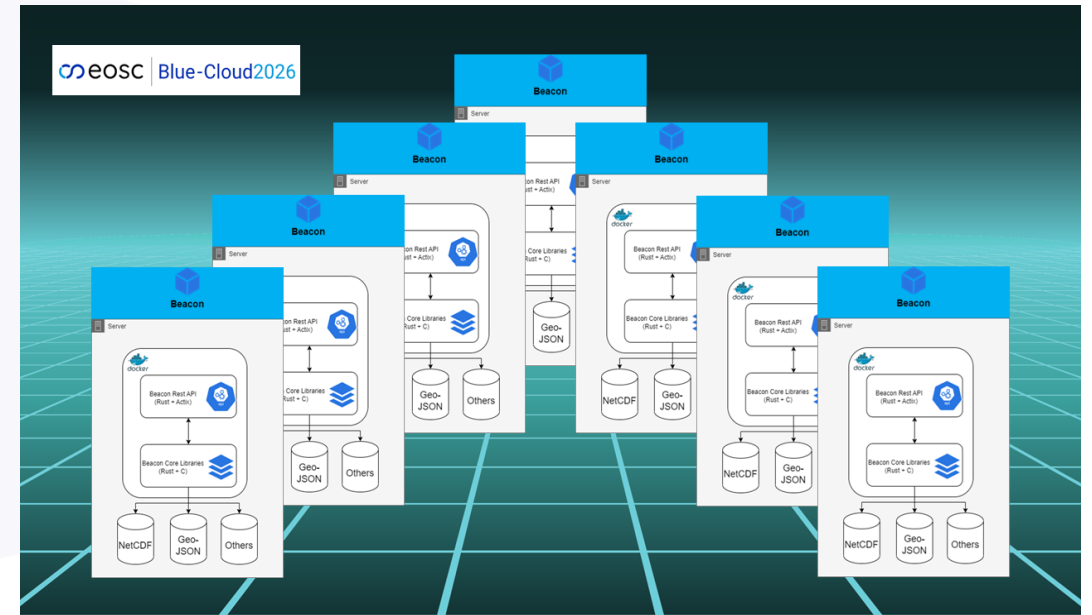
```





- WorkBench 1 (T&S) => BEACON to combine data from: SeaDataNet, WOD, EuroArgo, CMEMS IN-SITU TAC
- WorkBench 2 (Eutrophication) => BEACON to combine data from: SeaDataNet, CMEMS IN-SITU TAC, WOD, EMODnet Chemistry

- Euro-Argo data
 - *retrieved from S3 bucket*
- CORA Profile data
 - *retrieved from CMEMS*
- CORA Timeseries data
 - *retrieved from CMEMS*
- EMODnet Chemistry data
 - *retrieved from EMODnet Chemistry WebODV*
- WOD data
 - *retrieved from ncei.noaa.gov*
- SeaDataNet CDI TS data
 - *retrieved from EGI-ACE webODV*
- SeaDataNet CDI Incremental
 - *retrieved from SeaDataNet CDI service*
- CMEMS BGC data
 - *retrieved from CMEMS*

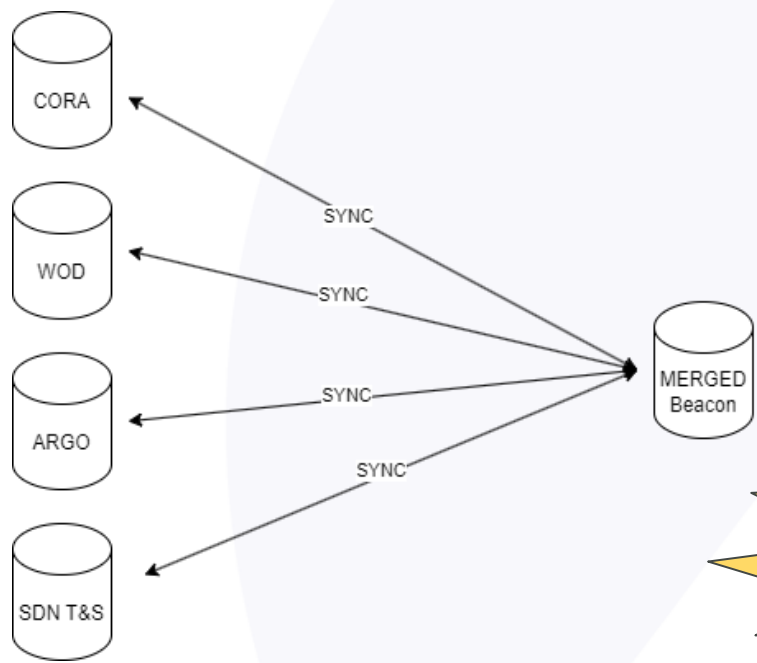


First deployed on MARIS servers

Now deployed on Blue-Cloud VRE, accessible for Blue-Cloud users

From monolithic nodes to merged Beacons

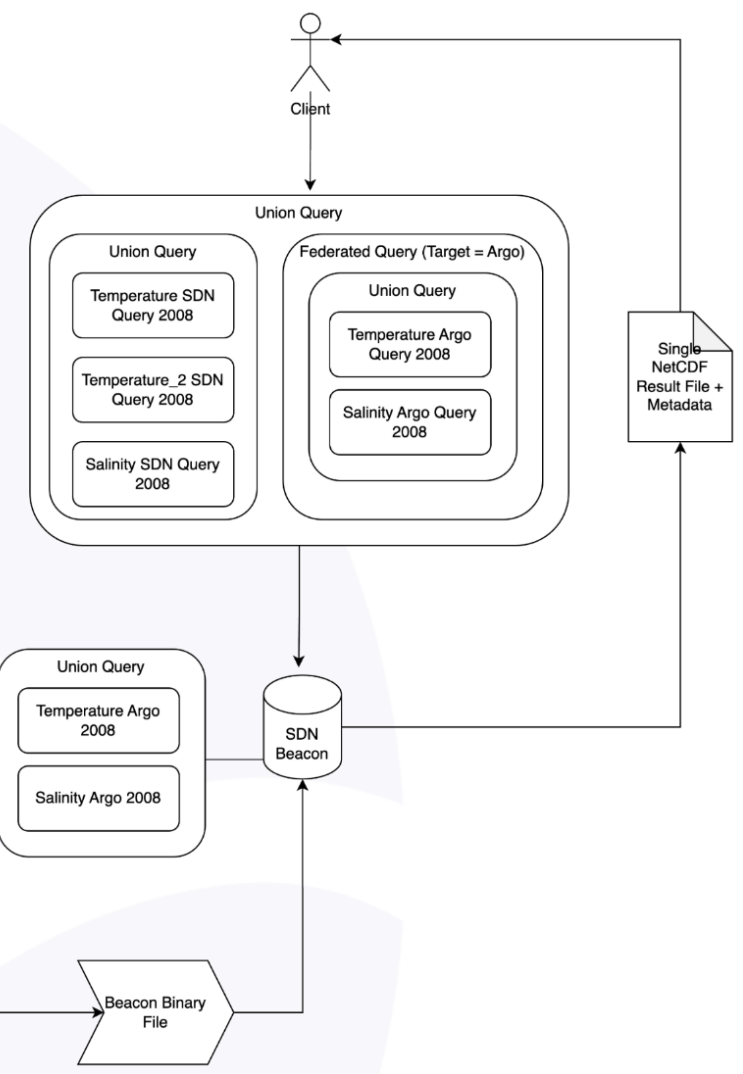
- Building on top of the monolithic BEACON instances
- Adopting common metadata profile as core + additional tags
- Semantic harmonisation using the semantic broker
- Output via M-to-M to WebODV for duplicate checks and QA-QC



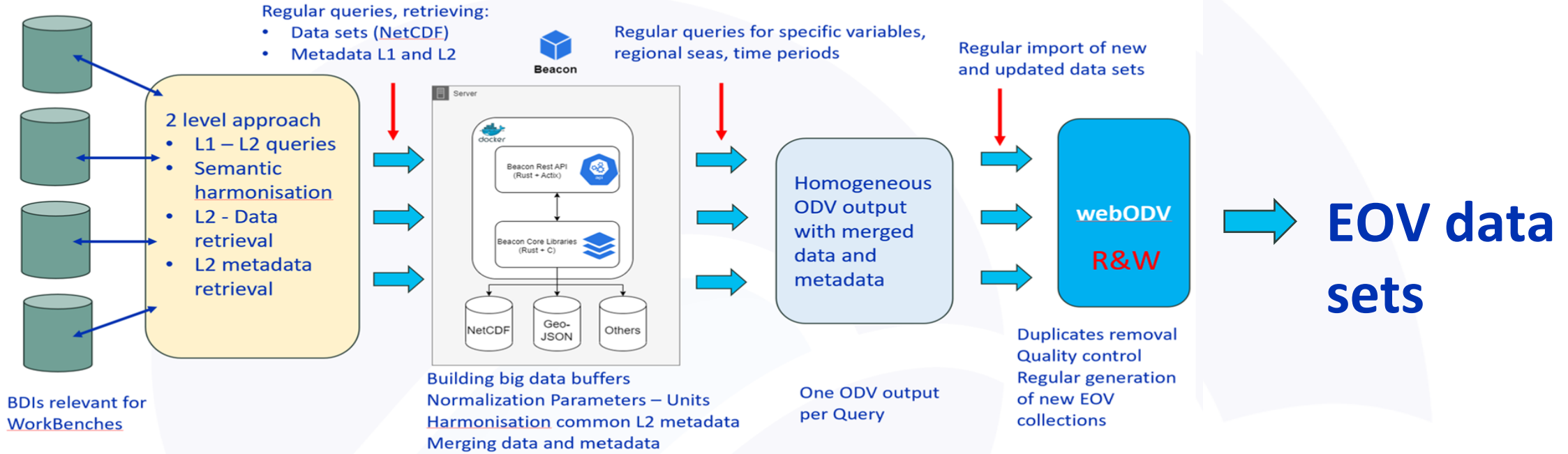
MERGED Beacon

The merged Beacon Instance will have the triples from the sync feed and mappings applied during the syncing process.

This way we can update mappings from the triples, and apply the updated mappings when re-syncing from the monolithic instances.



Dashboard for controlling and administering new harvests and ingestions into webODV



eosc | Blue-Cloud2026



blue-cloud.org



[@bluecloudeu](https://twitter.com/bluecloudeu)



[blue-cloud org](https://www.linkedin.com/company/blue-cloud-org)



Funded by
the European Union