eosc | Blue-Cloud2026

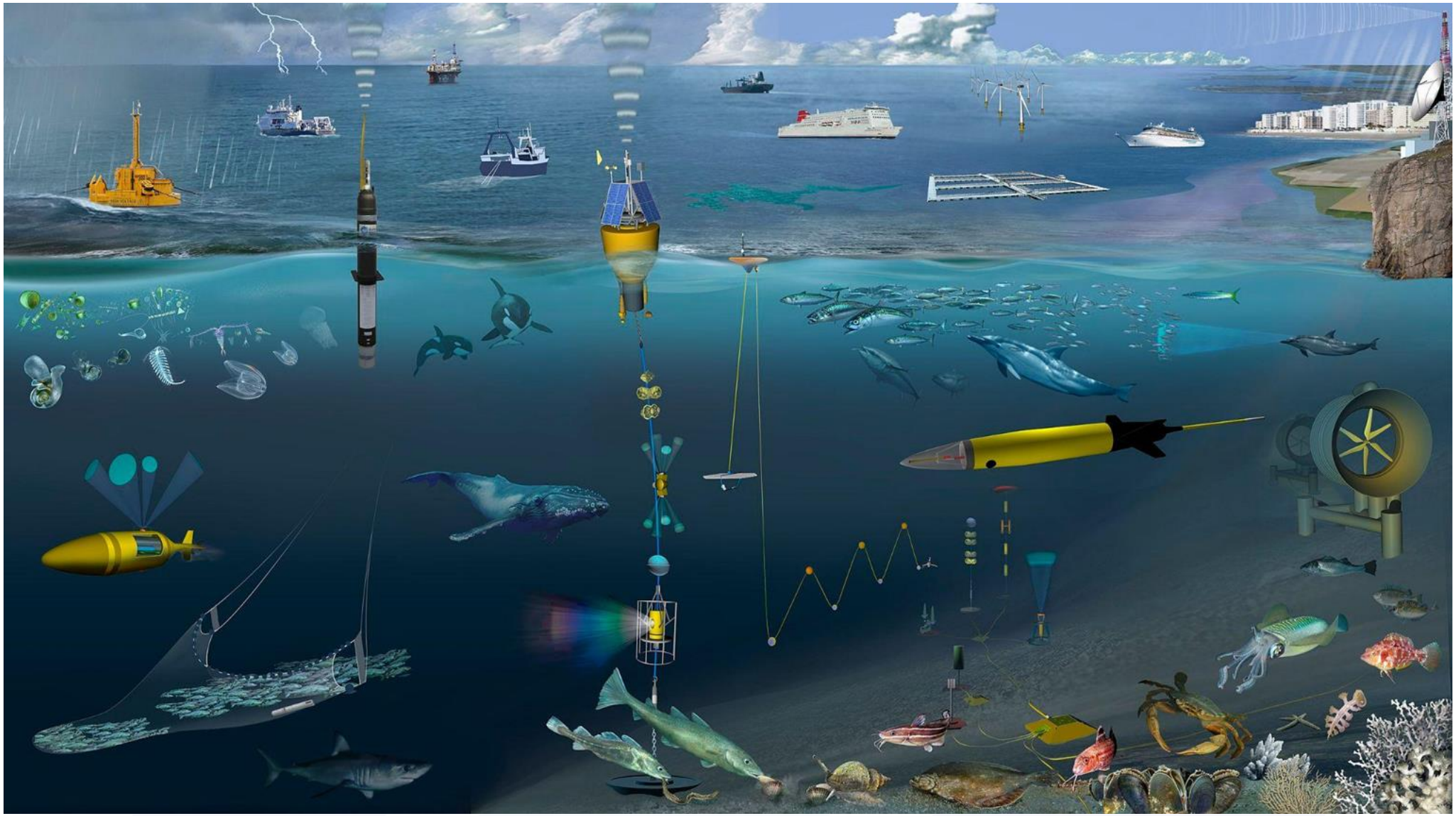
FAIR compliant Discovery and Access services for marine domains & beyond

Optimising FAIRness of federated Blue Data Infrastructures webinar
6 Dec 2023

Dick Schaap, MARIS, Blue-Cloud 2026 Technical Coordinator

Funded by the European Union

Marine and Ocean Observations



Blue Data Infrastructures

Blue Cloud Discovery and Access service

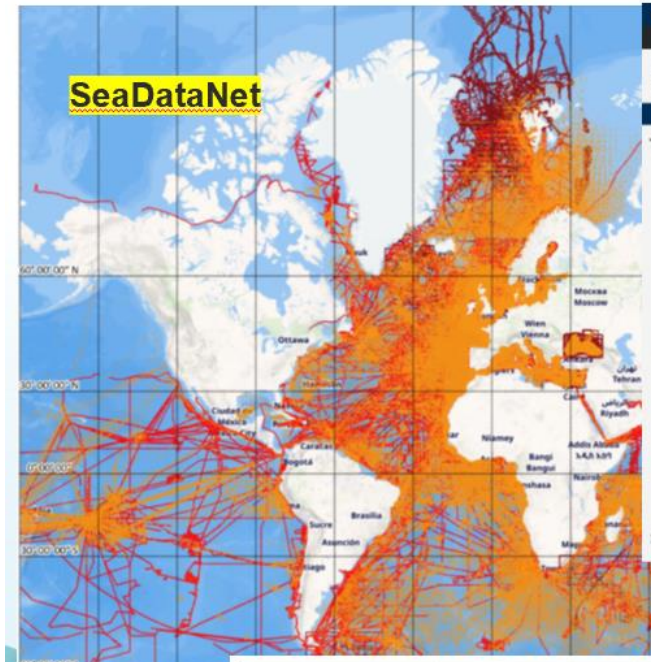
Facilitates users:

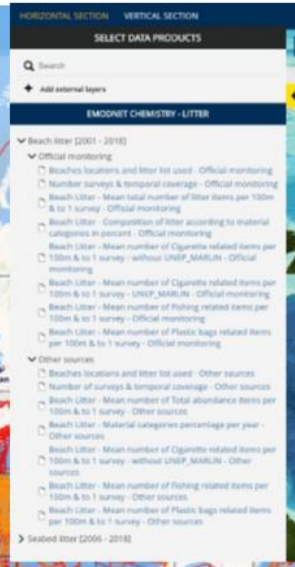
- Federated search for discovering interesting data sets (currently more than 10 million) in a common way
- Federated retrieval of identified data sets using a shopping basket mechanism
- Download of data sets or push to Blue-Cloud VRE

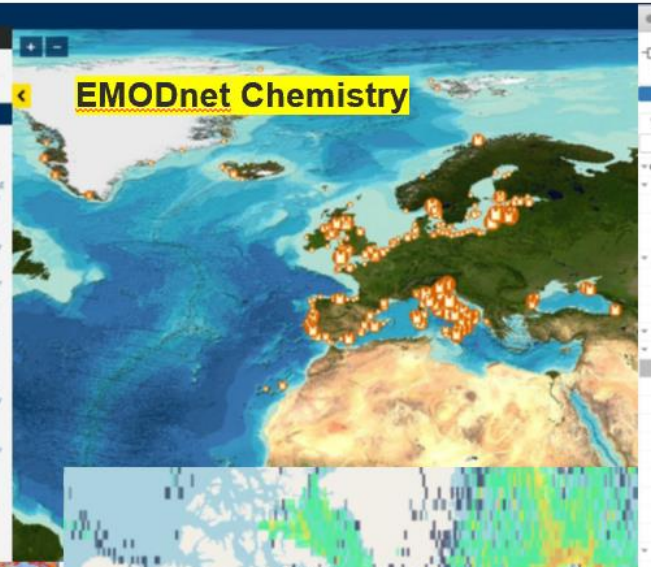
Facilitates managers of Blue Data Infrastructures:

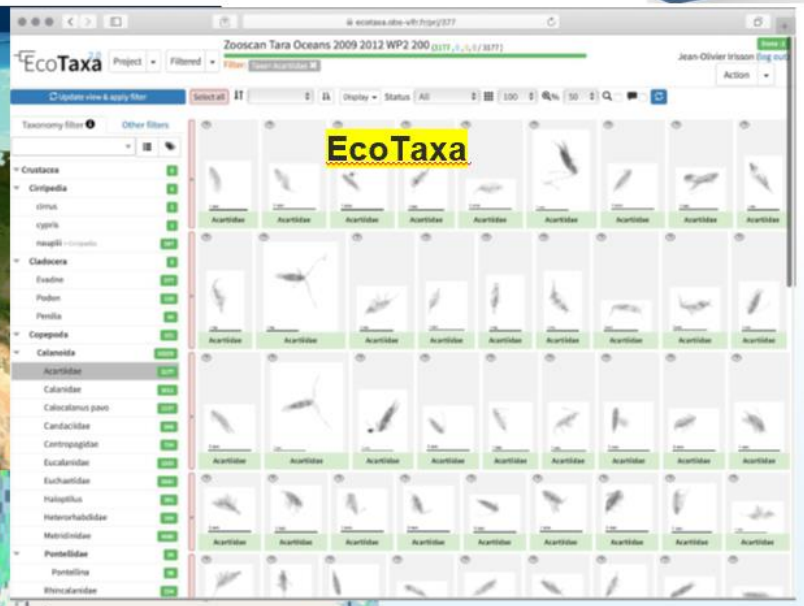
- Wider outreach to potential users
- Stay informed about data requests and users for their repository
- Periodic reporting of downloads from their repository

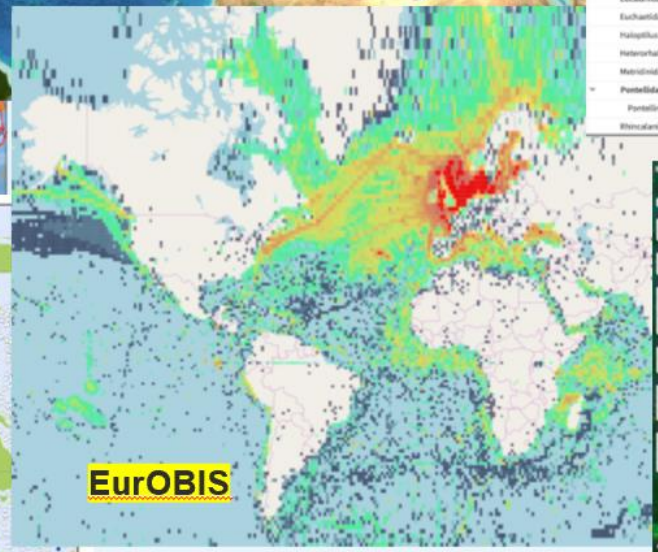
Illustrations of data coverage

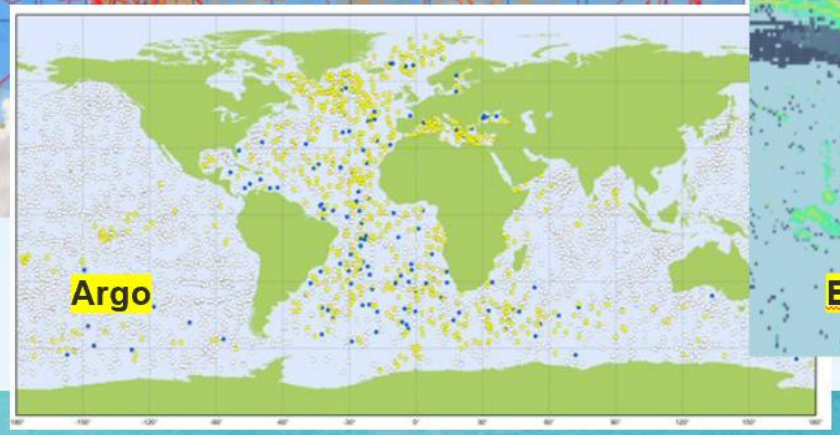


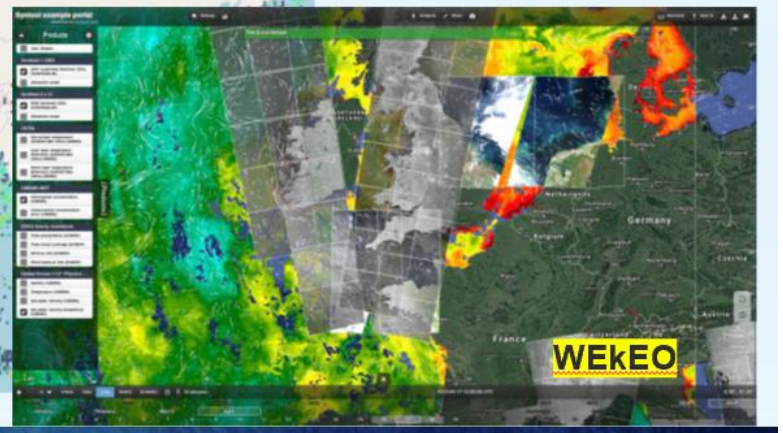












DD&AS conceptual approach

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

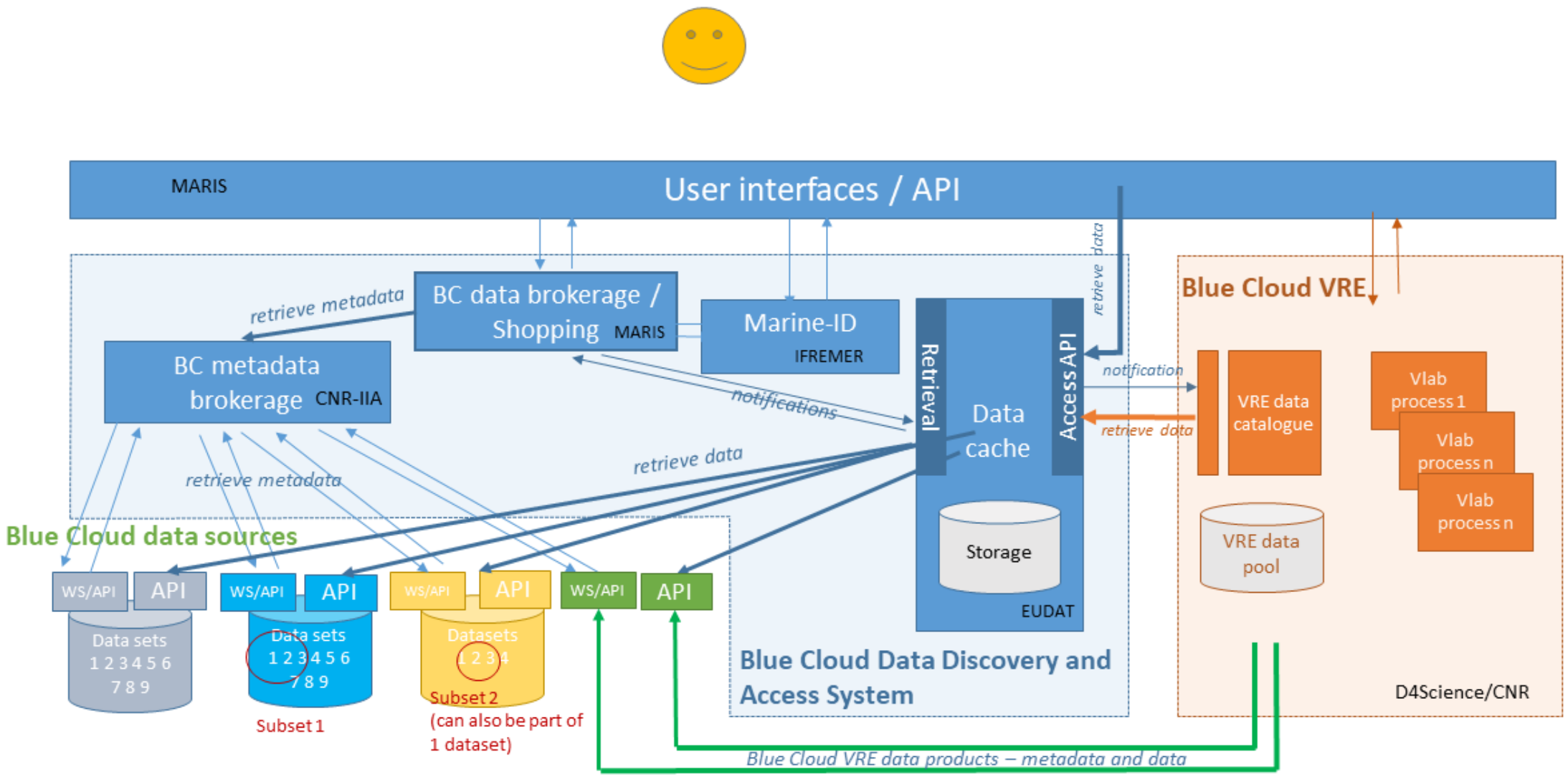
Concept of two-step search approach:

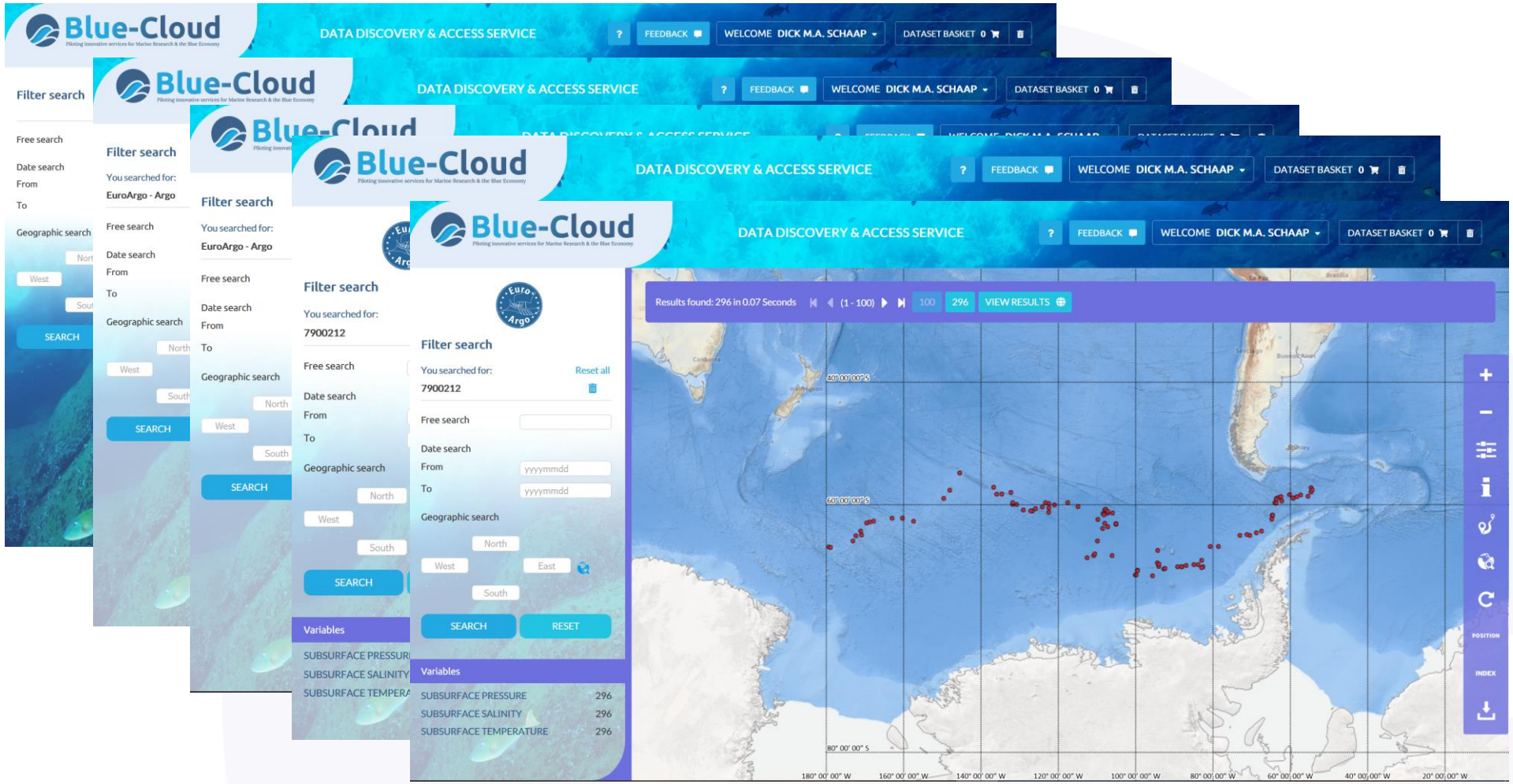
First step: identifying interesting data collections and products with few criteria

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

DD&AS architecture



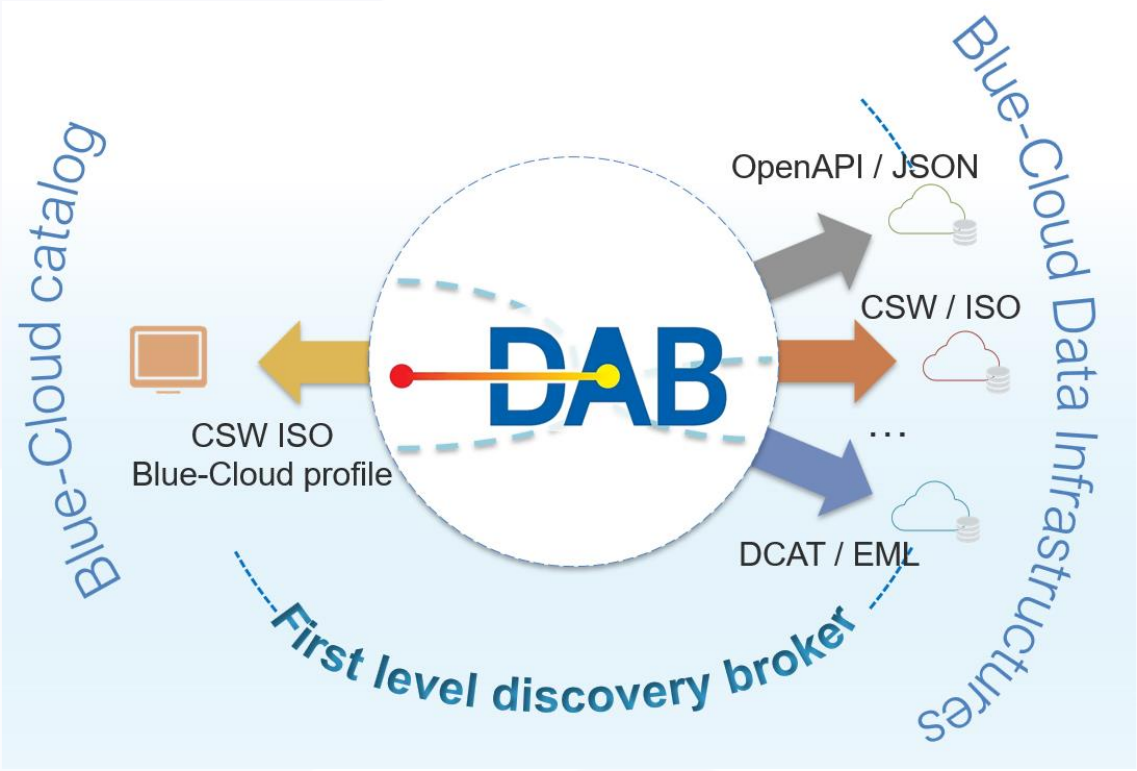


https://data.blue-cloud.org

Level 1 – Common Blue-Cloud metadata catalogue at collection level

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



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 **Blue-Cloud Data Discovery & Access service** (DD&AS) and its FAIRness will be expanded and optimized by:

- harmonising and expanding functionality of web services as operated by each BDI for discovery and access of managed data resources, and as used in DD&AS, following FAIRness review
- developing and deploying semantic brokering as part of DD&AS interface
- federating additional BDIs into the DD&AS (**EMSO, SIOS, EMODnet Physics, ELIXIR-Mgnify)**
- developing and deploying data sub-setting and extracting services, in addition to discovery and access, for feeding Blue-Cloud 'raw data' Data Lakes,
- tuning Data Lakes developments with Digital Twin of the Ocean (DTO) developments, in particular EDITO-Infra.

Actions for DD&AS semantic harmonisation

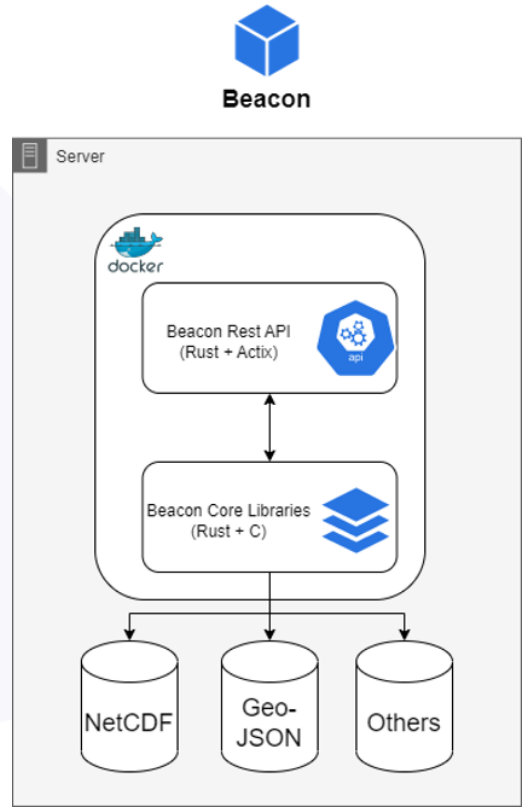
- Each BDI is requested to express their use of vocabularies in their metadata output by including ‘triples’ for:
  - literal description of term => example: sea level
  - Term code => example: https://vocab.nerc.ac.uk/collection/P02/current/ASLV
  - Used vocabulary => example: https://vocab.nerc.ac.uk/collection/P02/current/
- Each BDI is also requested, where missing, to adopt using vocabularies, where possible, focusing on uptake of SeaDataNet standards:
  - BODC – SeaDataNet controlled vocabularies
  - EDMO (organisations)
  - EDMERP (projects – programmes)
  - CSR (Cruise Summary Reports)
  - WoRMS for taxonomy
- In a number of cases, BDIs are requested to improve their web services
- CNR-IIA will adapt the DAB broker common metadata profile (XML) to include the semantic ‘triples’ in OGC – CSW output
- MARIS will adapt the level 2 data broker to include the semantic triples in search and output
- NOC-BODC, MARIS and CNR-IIA will work on analysing used vocabularies and ranges, currently in use, and deploying a semantic brokerage for mapping towards harmonised semantics

Other actions for optimising central components of DD&AS

- Optimising current federations with BDIs at level 1 and level 2
- Replace Marine-ID AAI for being fit for federated EOSC AAI:
  - Make use of the Keycloak service instance at CNR as used for VRE
  - CNR Keycloak is part of federation with EGI Check-IN
- Arrange federated monitoring for DD&AS components incl BDIs, where needed, possibly using EOSC core service for monitoring
- Developing and deploying API for DD&AS (once all works fine)
- Developing and deploying subsetting functionalities for feeding Data Lakes for WorkBenches and external use applications:
  - further developing MARIS BEACON tool
  - analysing and optimising existing subsetting services at BDIs
  - analysing subsetting services for international data sources, such as WOD
- Expanding federation with new BDIs: EMSO, SIOS, EMODnet Physics, and ELIXIR-Mgnify

BEACON data lake tool

- BEACON is a high-performance climate & marine data lake solution used to store and subset millions of NetCDF datasets and terabytes of data with powerful query possibilities and lightning fast retrieval
- BEACON makes use of the power of Rust with advanced optimization techniques and parallel processing capabilities to deliver unmatched performance. It enables you to explore and query millions datasets on the fly with ease.
- BEACON: Use the API to query on metadata, subset with precision using ranges on every single parameter and unit, and visualize spatial data with advanced geo polygon filters.
- BEACON is easy to set up with an intuitive Rest API and seamless integration with the existing NetCDF ecosystem.
- BEACON simplifies your data analysis with its harmonized single file output! BEACON seamlessly combines and harmonizes data, providing a unified view that eliminates the need for complex transformations, conversions, and mappings.



- https://beacon.maris.nl/

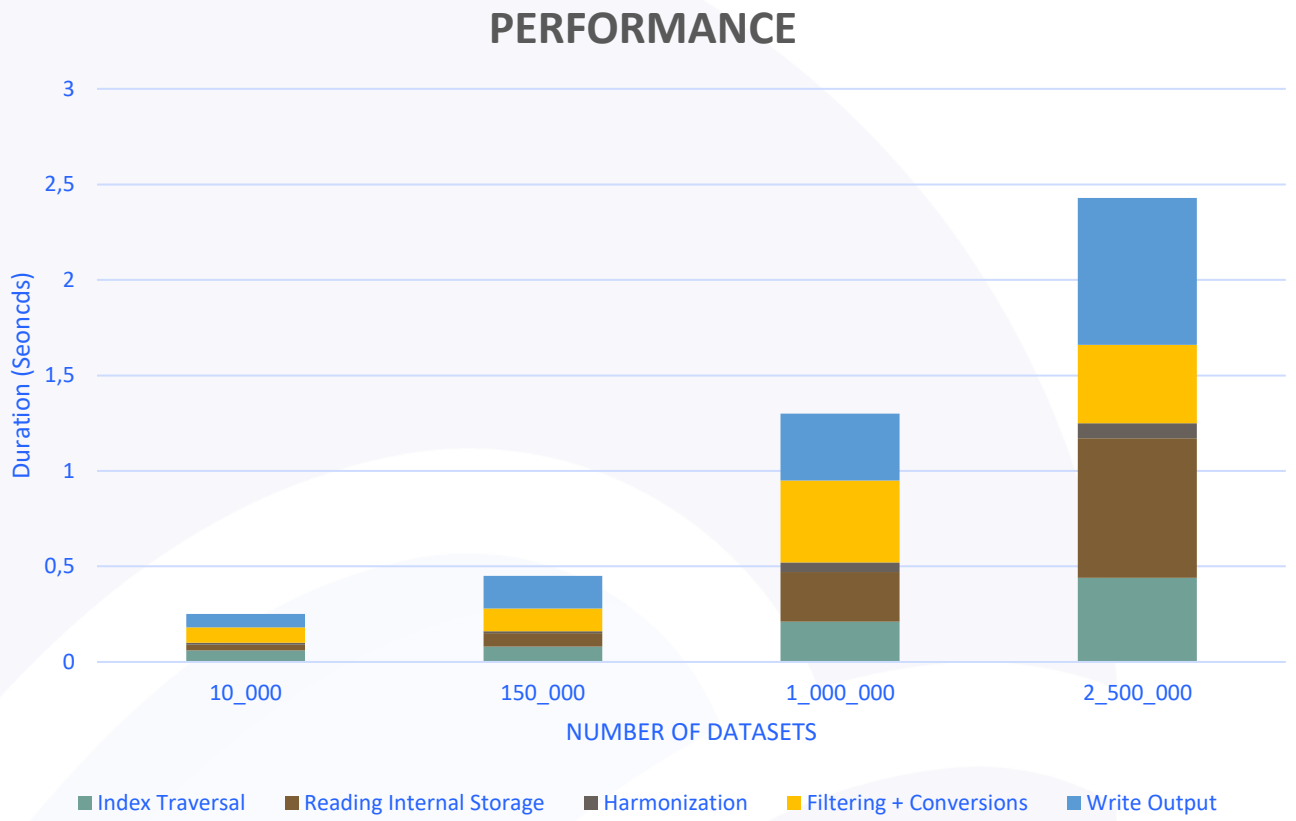
BEACON data lake tool

Loaded into beacon all SDN CDI records:

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

Query:

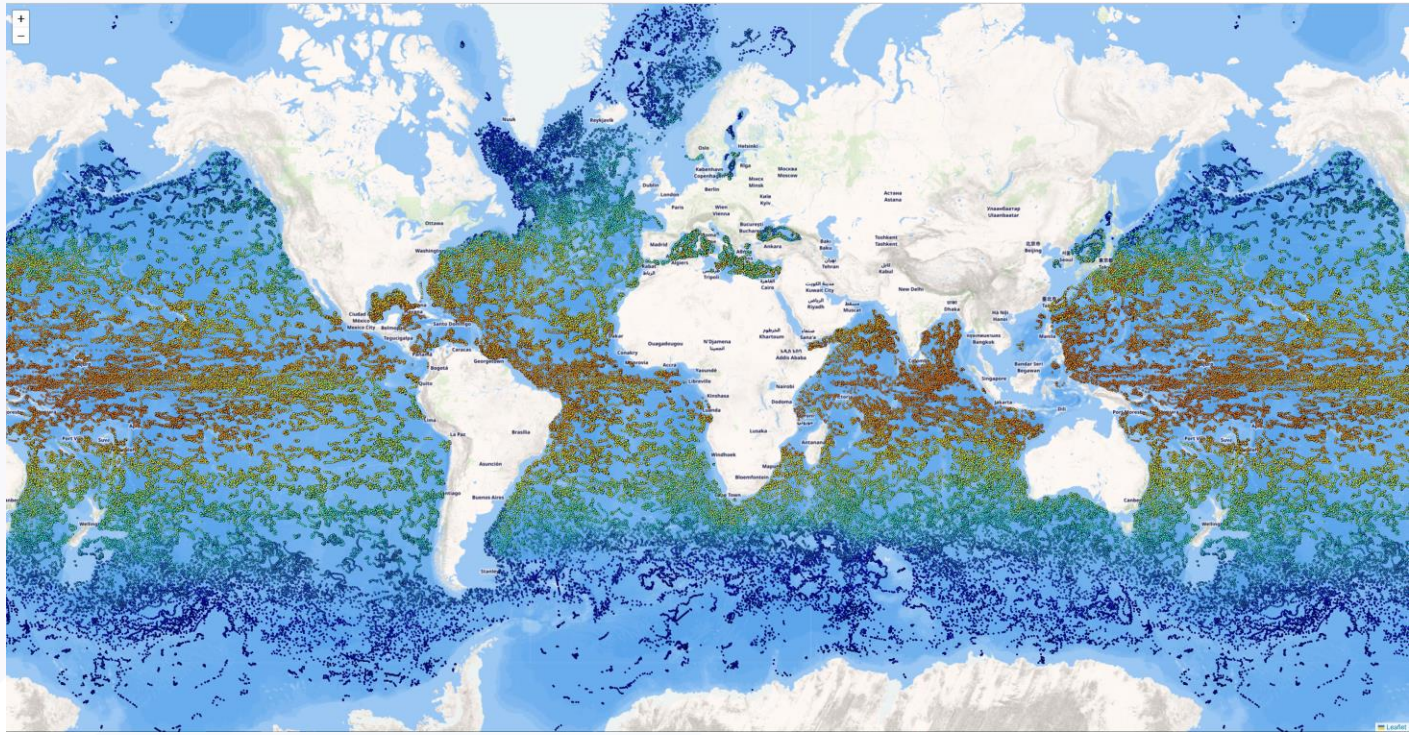
- Longitude from -8 to 12
- Latitude from 50 to 61
- Depth from 0 to 50
- Time from 2010 to 2012
- All the temperature parameters aggregated and harmonized in degrees Celsius
- Result: 12M points!



BEACON combined with fast data viewer as part of EOSC-FUTURE use case

- Action takes place on the web server instead of the client's computer
- Using Beacon Binary Format to cache API results
- Tiling WMS on top of beacon API instead of geoJSON
- Better browser performance (no memory issues, responsive UI)

https://eosc-future.maris.nl

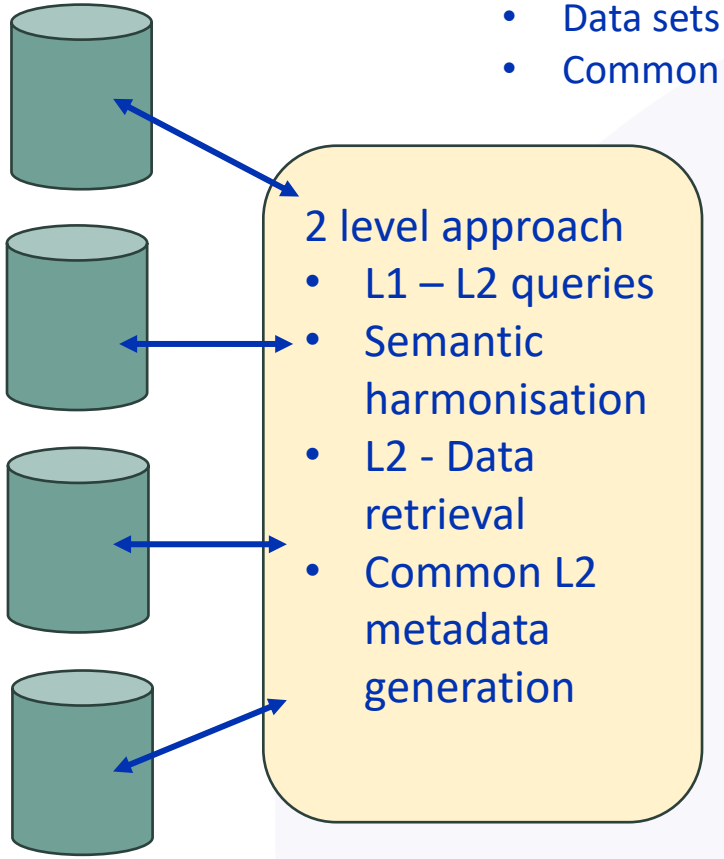


Argo measurements at surface layer for 2020

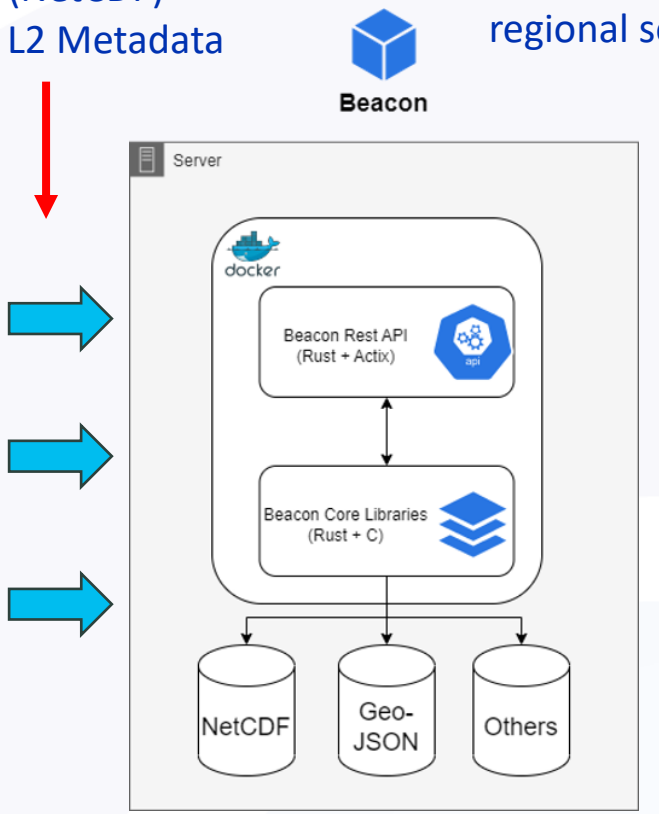
Proposed solution: Coupling DD&AS with BEACON

Multiple queries, retrieving:
- Data sets (NetCDF)
- Common L2 Metadata

Queries for specific variables, regional seas, time periods



2 level approach
- L1 – L2 queries
- Semantic harmonisation
- L2 - Data retrieval
- Common L2 metadata generation



Set of common L2 metadata

Homogeneous NetCDF with all selected data and harmonised data metadata

Building big data buffers
Multiple NetCDF formats
Inserting Common Metadata
Harmonisation P01 and P06 – select menu

One NetCDF per Query

Planning of deliverables

D2.1: Existing DD&AS and Blue Data Infrastructures – Review and Specifications for Optimisation Report: Report with FAIRness review of existing BDI web services for discovery and access and existing DD&AS central services **(MARIS)** - **Submitted to EU – end October 2023 (M10)**

D2.2: New Blue Data Infrastructures – Service Analysis Report: Detailed descriptions of the new to be connected BDIs (EMSO, EMBRC, SIOS, EMODnet Physics) with details on their local discovery and access mechanisms, types of data, metadata format, data formats, use of vocabularies, possible restrictions, and existing web services **(MARIS**) – **Submitted to EU – end November 2023 (M11)**

D2.3: Optimised and expanded Blue Cloud Data Discovery and Access Service – Documentation Report: Report documenting the new release of the Blue Cloud Data Discovery and Access Service with optimised services, semantic interoperability, new BDIs connected, and new or adapted discovery and access BDI web services **(MARIS) – planned for end December 2024 (M24)**

D2.4: BDI sub-setting APIs and Data Lakes – Concept and Specifications Report: Report with descriptions and analyses of existing sub-setting services at each of the BDIs, formulation of common requirements for BDI sub-setting APIs, and how BDIs might adapt existing or develop new sub-setting APIs **(MARIS) – planned for end April 2024 (M16)**

NOTE: All deliverables will be made available at the Blue-Cloud website and ZENODO

eosc | Blue-Cloud2026


blue-cloud.org


@bluecloudeu


blue-cloud org

Funded by
the European Union