International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-024-00664-0

REGULAR PAPER

Blue-cloud DAB: developing a platform to harmonize, assess and disseminate marine metadata collections

Enrico Boldrini[1] · Roberto Roncella[1] · Fabrizio Papeschi[1] · Paolo Mazzetti[1] · Dick Schaap[2] · Peter Thijsse[2] · Paul Weerheim[2] · Stefano Nativi[1]

Received: 30 June 2024 / Accepted: 29 September 2024

Abstract
The integration and harmonization of marine data from diverse sources are vital for advancing global oceanographic research and ensuring seamless discovery and access of critical datasets. This paper presents a comprehensive analysis of the metadata harmonization efforts within the Blue-cloud 2026 project, which brokers data from numerous Blue Data Infrastructures (BDIs), leveraging the Discovery and Access Broker technology. The platform enables discovery and analysis of marine data collections while facilitating interoperability with other components of the marine digital ecosystem, such as virtual laboratories and the Semantic Analyzer. It also supports the flow of Blue-cloud information to other initiatives like the Global Earth Observations System of Systems. For data managers, the findings emphasize the importance of enhancing metadata quality, revealing discrepancies in core metadata elements, and the need for more consistent use of controlled vocabularies. For cyberinfrastructure developers, the study details the challenges of accommodating a wide array of interfaces from different data systems, highlighting the adoption of an extensible brokering architecture that harmonizes metadata models and protocols. The study also emphasizes the importance of metadata analysis in ensuring effective searches for end users, highlighting challenges in aggregating diverse sources, where data providers may have structured the content with different objectives compared to those of the system of systems. End users will gain insights into the current metadata content of Blue-cloud, enabling them to search and access data from multiple BDIs with an understanding of the technical complexities behind the scenes.

Keywords Brokering approach · Marine data · Metadata analysis · System of systems · Digital ecosystems

1 Introduction

Earth's surface is predominantly covered by water, making the assessment and prediction of the state of marine environments crucial for various human interests, including food provision, transportation and environmental monitoring. The ability to accurately monitor and forecast marine conditions is vital for assessing and achieving policies such as the Sustainable Development Goals (SDGs) outlined in the 2030 Agenda for Sustainable Development [1].

Thousands of organizations engage in multidisciplinary projects and programs to continuously gather and disseminate near-real-time and real-time marine data from in situ and remote sensors, but also historical data and analyzed samples. This data collection aims to extract valuable information on essential marine variables, facilitating the calculation of indicators to assess targets and goal completions, and to produce curated data products and accurate forecasts.

In the European Union (EU) context, Horizon Europe Blue-cloud 2026 project[1] plays a pivotal role in the EU Strategy for Healthy Oceans, Seas, Coastal, and Inland Waters [2], as well as the European Open Science Cloud (EOSC) initiative [3]. The overarching objective of Blue-cloud 2026 is to further develop the European federation of marine and inland water data management infrastructures and services. This development aims to enhance the FAIRness (Findability, Accessibility, Interoperability, and Reusability) of data, provide advanced analytical capabilities, and ensure higher

✉ Enrico Boldrini
enrico.boldrini@cnr.it

[1] National Research Council of Italy, Institute of Atmospheric Pollution Research, Florence, Italy

[2] MARIS BV, Nootdorp, The Netherlands

[1] Blue-Cloud 2026, "Homepage", https://blue-cloud.org/

Published online: 16 October 2024

quality data provision, in line with the principles of the GO FAIR initiative.[2] The project, which started on January 1, 2023, lasts three and a half years and is about midway through its timeline.

Blue-cloud 2026 is undertaken by leading European ocean and marine data and knowledge initiatives, such as the European Marine Observation and Data Network (EMODnet) and the Copernicus Marine Environmental Monitoring Service (CMEMS), alongside prominent aquatic environmental research infrastructures. These entities, collectively referred to as Blue Data Infrastructures (BDIs), are at the forefront of this initiative, coordinating collection and publication of marine data worldwide, including both streaming data and curated data products.

A key feature of Blue-cloud is its data systems brokering approach, which builds a harmonized platform on top of autonomous heterogeneous systems. The broker relies on a harmonized common metadata model, known as the Blue-cloud metadata profile. This profile comprises elements deemed most important for discovering and accessing blue data, including URIs to unambiguously reference concepts following the linked data approach.

The metadata brokering platform detailed in this article and in charge of the Institute of Atmospheric Pollution Research (IIA) of the National Research Council of Italy (CNR) (CNR-IIA) is based on the Discovery and Access Broker (DAB) open technology and has been established through collaborative work among project technical coordinators, broker developers, and data providers, incorporating agile development cycles and feedback.

Thanks to the metadata brokering platform, users can search across harmonized dataset collections through the online available Blue-cloud portal.[3] Access and further processing of the matching data collections is the responsibility of other components of the Blue-cloud infrastructure, namely the data brokering component, the data cache, and the Virtual Research Environment (VRE).

Additionally, the metadata brokering platform allows for detailed analyses of the available metadata content, as discussed in this work. This additional capability can serve the marine community to assess and improve the information quality. Furthermore, the platform facilitates the information flow to other initiatives, such as the Global Earth Observations System of Systems (GEOSS) and enables seamless integration with other components that can generate added value from the metadata content, such as the Blue-cloud Semantic Analyzer.

[2] GO FAIR, "GO FAIR Initiative", https://www.go-fair.org/go-fair-initiative/

[3] Blue-Cloud, "Blue-Cloud Data Discovery & Access Service", https://data.blue-cloud.org/search

2 Blue data infrastructures

In the context of Blue-cloud 2026 project, BDIs are the key systems providing blue data. The European organizations responsible for BDIs are project partners selected for their regional and global relevance. Each of these organizations efficiently aggregates contributions from multiple local and regional data providers within their specific domains, establishing a domain-based normalization path. Nonetheless, currently, no single organization or infrastructure offers comprehensive data coverage across all marine domains. The Blue-Cloud 2026 project aims to bridge this gap by enabling multidisciplinary marine data search and access.

This section provides a short overview of the BDIs involved in the Blue-Cloud project. It introduces the data asset associated with each BDI, details the technical contact that was involved in Blue-Cloud activities and lists the available human-accessible portals. Finally, it covers the web service interfaces for data sharing, including the protocols, data models and vocabularies used for the machine-to-machine connection to Blue-Cloud, highlighting the existing heterogeneity and the challenges posed by big data variety [4]. Additional information is also available in the Blue-Cloud 2026 project deliverables [5, 6].

2.1 Argo

Data asset Salinity, temperature, biogeochemistry and ocean currents data from a robotic fleet that dives and glides through the oceans. A web portal is available.[4]

Technical contact The French national institute for ocean science and technology (Ifremer),[5] member of the European Research Infrastructure Consortium (ERIC) Euro-Argo,[6] the European partner of the Argo program.

Web service Swagger 2.0 Application Programming Interface (API),[7] with custom, JavaScript Object Notation (JSON) based data model. No controlled vocabularies seem to be used.

2.2 ELIXIR-ENA

Data asset The European Nucleotide Archive (ENA) is a comprehensive open repository of the world's nucleotide sequencing information. A web portal is available.[8]

[4] Euro-Argo ERIC, "Argo Fleet Monitoring", https://fleetmonitoring.euro-argo.eu/dashboard

[5] Ifremer, "Homepage", https://en.ifremer.fr/

[6] Euro-Argo ERIC, "Homepage", https://www.euro-argo.eu/

[7] Euro-Argo ERIC, "Argo Fleet Monitoring API", https://fleetmonitoring.euro-argo.eu/swagger-ui.html

[8] EMBL-EBI, "ENA browser", https://www.ebi.ac.uk/ena/browser/view/

Technical contact European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL) (EMBL-EBI).[9]

Web service Swagger 3.0 API,[10] with custom, JSON based data model. No controlled vocabularies seem to be used.

2.3 ELIXIR-MGnify

Data asset The MGnify platform [7] is an analytical component of the ELIXIR-ENA system. It is dedicated to the assembly, analysis and archiving of microbiome-derived nucleic acid sequences from different environments. A web portal is available.[11]

Technical contact EMBL-EBI.

Web service OpenAPI Specification 3.0 API,[12] with custom, JSON based data model. No controlled vocabularies seem to be used.

2.4 EMODnet chemistry

Data asset Chemistry observation data related to eutrophication, contaminants, and marine litter. The platform aims to produce validated, aggregated data collections and interpolated map products. A web portal is available.[13]

Technical contact National Institute of Oceanography and Applied Geophysics of Italy (OGS)[14] and Ifremer.

Web service Open Geospatial Consortium (OGC) Catalogue Service for the Web (CSW) 2.0.2 International Organization for Standardization (ISO) application profile, implemented using GeoNetwork technology.[15] Its data model is ISO 19115 based, encoded as ISO 19139 Extensible Markup Language (XML). The Infrastructure for Spatial Information in the European Community (INSPIRE) theme register[16] is used for keywords, Natural Environment Research Council (NERC) Vocabulary Server (NVS) vocabularies [8], 9 are used for parameters and the European Directory of Marine Organisations (EDMO)[17] for organizations.

2.5 EMODnet physics

Data asset In situ ocean physics time-series data, vertical profiles, and metadata, including parameters such as temperature, salinity, currents, sea level trends, wave height, wind speed, and more. A web portal is available.[18]

Technical contact ETT.[19]

Web service ERDDAP [10] service,[20] with custom, JSON exportable data model. NVS vocabularies are used for parameters.

2.6 EMSO ERIC

Data asset The European Multidisciplinary Seafloor and Water Column Observatory (EMSO) is a distributed infrastructure comprising ocean observation systems across 14 test sites, storing a wide range of data for marine monitoring purposes. A web portal is available.[21]

Technical contact EMSO-ERIC.[22]

Web service: ERDDAP service,[23] with custom, JSON exportable data model. NVS vocabularies are used for parameters and EDMO is used for organizations.

2.7 EurOBIS

Data asset Biogeographic data focused on taxonomy and distribution records in European marine waters and by European researchers globally, including species presence, abundance, biomass data, and length measurements. A web portal is available.[24]

Blue-cloud technical contact Flanders Marine Institute[25] (VLIZ), manager of the European node of the international Ocean Biodiversity Information System (EurOBIS).

Web service Linked data endpoint,[26] with Data Catalog Vocabulary (DCAT) based data model, Turtle Resource Description Framework Schema (RDFS) encoded. The World Register of Marine Species (WoRMS) [11] is used for keywords, NVS vocabularies are used for parameters and

[9] EMBL-EBI, "Homepage", https://www.ebi.ac.uk/

[10] EMBL-EBI, "ENA Portal API", https://www.ebi.ac.uk/ena/portal/api/

[11] EMBL-EBI, "MGnify", https://www.ebi.ac.uk/metagenomics

[12] EMBL-EBI, "MGnify API", https://www.ebi.ac.uk/metagenomics/api/docs/

[13] EMODnet, "EMODnet Chemistry", https://emodnet.ec.europa.eu/en/chemistry

[14] OGS, "Homepage", https://www.ogs.it/en

[15] EMODnet, "EMODnet Chemistry products catalogue", https://emodnet.ec.europa.eu/geonetwork/emodnet/eng/csw?

[16] INSPIRE, "INSPIRE theme register", https://inspire.ec.europa.eu/theme

[17] MARIS, "EDMO Homepage", https://www.seadatanet.org/Metadata/EDMO-Organisations

[18] EMODnet, "EMODnet Physics", https://emodnet.ec.europa.eu/en/physics

[19] ETT, "Homepage", https://ettsolutions.com/

[20] EMODnet, "EMODnet Physics ERDDAP", https://data-erddap.emodnet-physics.eu/erddap

[21] EMSO-ERIC, "EMSO data portal", https://data.emso.eu/home

[22] EMSO-ERIC, "Homepage", https://emso.eu/

[23] EMSO-ERIC, "EMSO ERDDAP service", https://erddap.emso.eu/erddap/index.html

[24] VLIZ, "EurOBIS data access & services", https://www.eurobis.org/data_access_services

[25] VLIZ, "Homepage", https://vliz.be/en

[26] VLIZ, "EurOBIS linked data endpoint", https://marineinfo.org/id/collection/619.ttl

instruments and VLIZ MarineInfo registry[27] for organizations.

2.8 EcoTaxa

Data asset Planktonic biodiversity data.[28]

Blue-cloud technical contact Laboratoire d'Océanographie de Villefranche (LOV)[29] of the Sorbonne University.

Web service EcoTaxa data is shared through the EurOBIS service, as such in this analysis isn't treated as a separate subset.

2.9 ICOS data portal

Data asset Long-term oceanic observations focused on the global carbon cycle and climate-relevant gas emissions, gathered by over 130 greenhouse gas measurement stations across Europe and neighboring regions. A web portal is available.[30]

Technical contact Integrated Carbon Observation System (ICOS).[31]

Web service SPARQL Protocol and RDF Query Language (SPARQL) endpoint service,[32] with custom, JSON based data model. ICOS community registry is used for parameters, organizations and projects.

2.10 ICOS SOCAT

Data asset Surface ocean CO_2 Atlas (SOCAT) measurements, featuring 35.6 million quality-controlled fCO_2 observations from 1957 to 2023, sourced from over 10 countries. A homepage is available.[33]

Technical contact University of Bergen[34] (UiB), Pacific Marine Environmental Laboratory[35] (PMEL) of the National Oceanic and Atmospheric Administration (NOAA).

Web service ERDDAP service,[36] with custom, JSON exportable data model. No controlled vocabularies seem to be used.

2.11 SeaDataNet—open datasets

Data asset Marine open datasets and data products from European research cruises and observational activities, covering European coastal marine waters, regional seas, and the global ocean. A web portal is available.[37]

Technical contact Mariene Informatie Services[38] (MARIS).

Web service XML based inventory service,[39] encoded as SeaDataNet Common Data Index (CDI) ISO profiles [12]. The INSPIRE theme register is used for keywords; NVS vocabularies are used for parameters, instruments and platforms; the EDMO is used for organizations; the Cruise Summary Report[40] (CSR) Inventory is used for cruises and the European Directory of Marine Environmental Research Project[41] (EDMERP) is used for projects.

2.12 SeaDataNet products

Data asset Derived data products including aggregated data collections and climatologies, such as temperature and salinity datasets. A web portal is available.[42]

Technical contact MARIS.

Web portal SeaDataNet Data Products.

Web service OGC CSW 2.0.2 ISO application profile, implemented using GeoNetwork technology.[43] Its data model is ISO 19115 based, encoded as ISO 19139 XML. The INSPIRE theme register is used for keywords; NVS vocabularies are used for parameters and the EDMO is used for organizations.

2.13 SIOS

Data asset Long-term measurements in and around Svalbard, focusing on Earth System Science questions related to Global Change. A web portal is available.[44]

Technical contact The Svalbard Integrated Arctic Earth Observing System (SIOS).[45]

[27] VLIZ, "MarineInfo", https://marineinfo.org/

[28] EcoTaxa, "EcoTaxa exploration", https://ecotaxa.obs-vlfr.fr/explore/

[29] LOV, "Homepage", https://lov.imev-mer.fr/web/

[30] ICOS, "ICOS data portal", https://data.icos-cp.eu/portal

[31] ICOS, "Homepage", https://www.icos-cp.eu/

[32] ICOS, "ICOS SPARQL endpoint", https://meta.icos-cp.eu/sparql

[33] ICOS-SOCAT, "Homepage", https://socat.info/

[34] UiB, "Homepage", https://www.uib.no/en

[35] NOAA PMEL, "Homepage", https://www.pmel.noaa.gov/

[36] ICOS-SOCAT, "ICOS-SOCAT ERDDAP service", https://data.pmel.noaa.gov/socat/erddap/tabledap/socat_v2023_fulldata

[37] SeaDataNet, "SeaDataNet CDI", https://cdi.seadatanet.org/search

[38] MARIS, "Homepage", https://www.maris.nl/

[39] SeaDataNet, "SeaDataNet Open Datasets inventory", https://cdi.seadatanet.org/report/aggregation/open

[40] SeaDataNet, "Cruise Summary Report Inventory", https://csr.seadatanet.org/

[41] SeaDataNet, "EDMERP", https://edmerp.seadatanet.org/

[42] SeaDataNet, "Data Products", https://www.seadatanet.org/Products#/search

[43] SeaDataNet, "SeaDataNet products catalogue", https://sextant.ifremer.fr/geonetwork/srv/eng/csw-SEADATANET

[44] SIOS, "Data access portal", https://sios-svalbard.org/metsis/search?

[45] SIOS, "Homepage", https://sios-svalbard.org/

Web service Blue-Cloud tailored Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) service endpoint,[46] its data model is Directory Interchange Format (DIF) Standard. Climate and Forecast (CF) Standard Name Table[47] is used for parameters.

3 Blue-cloud brokering approach

Blue Data currently comes from a dozen different infrastructures, platforms, and systems. To provide a common, harmonized entry point for data discovery and access, great heterogeneity must be addressed: the data sources introduced make use of different interfaces (e.g., OAI-PMH, CSW-ISO 2.0. 2, Inventory services, ERDDAP, SPARQL, OpenAPI API, Swagger API) and different metadata/data models (e.g. DIF, ISO 19115, CDI ISO 19115 profile, 5 different JSON-encoded customization models, DCAT/RDFS). It should also be kept in mind that other sources may be added in the near future and that some of those already connected may decide to evolve and/or change their services interface or data model. For example, the EurOBIS BDI replaced its OAI-PMH service with a linked data service during last year. This context highlights the need for a brokering archetypal pattern.

Over the past decades, a brokering approach [13, 14] has been successfully experimented with in various Earth Science initiatives and disciplines [15, 16, 17], to realize systems of systems, such as GEOSS [18], the Ocean Data Interoperability Platform (ODIP) [19, 20], and the World Meteorological Organization (WMO) Hydrological Observing System (WHOS) [21]. The enabling technology adopted in these cases, as well as in Blue-Cloud, is the DAB [22–25]. Developed and operated by the Earth and Space Science Informatics Laboratory (ESSI-Lab) of CNR-IIA as an open-source project available on GitHub[48] the DAB has been continuously advanced over the years through public research funding.

In this context, the broker is a third-party middleware component that plays a crucial role in the digital ecosystem by facilitating data flow from provider systems to consumer systems (e.g., portals, dashboards, models), thereby enabling discovery, access and further data processing [26]. Its use is particularly advantageous in building multidisciplinary systems where each system maintains its own autonomy and specificity, allowing them to evolve according to the needs of their respective communities.

The broker pattern alleviates the burden of both functional and metadata model mediation between different provider systems and consumer systems. This complexity is better managed by a dedicated central component (the broker), reducing the effort required from other system components. Data providers can continue to publish online data through their preferred service interfaces, while data consumers benefit from harmonized portals and standard service interfaces to access harmonized data programmatically.

The DAB is flexible and comprises pluggable components called accessors, which handle metadata mediation (e.g., crosswalks) from data provider models to the harmonized central model. A crosswalk is defined as the mapping of the elements, semantics, and syntax from one metadata scheme to those of another [27], where mapping is the correspondence between instances of one model and instances of another model that represent the same meaning [28]. Profiler components manage metadata mediation from the harmonized central model to the data consumer's required model. By relying on a central harmonized model, the broker approach optimizes the general task of mediation from m data provider models to n data consumer models (which would typically require $m \times n$ mappings) to the more manageable task of implementing $m + n$ components. Thus, only one additional mediation component needs to be implemented when a new system with a new model or protocol joins the system of systems.

3.1 Blue-cloud metadata profile

The harmonized model is a fundamental component of the brokering approach and must be carefully chosen, as it will be the basis for describing the harmonized data. For Blue-Cloud, it has been designed to include mandatory elements essential to the overall Blue-Cloud objectives of discovery, evaluation, access, and use. Additionally, being based on a solid metadata standard, it is rich enough to accommodate detailed information, allowing for the inclusion of precise optional information to further describe the marine collections.

The Blue-Cloud data model is based on the ISO 19115-1 [29] metadata model, which defines over 400 metadata elements. Its encoding is based on ISO 19115-3 [30]. Although the underlying model is capable of hosting very detailed information, which each BDI is encouraged to provide, only a few elements are strictly required or suggested in the Blue-Cloud context:

- **Blue-cloud core elements** (metadata) identifier, title, keyword, keyword Uniform Resource Identifier (URI), bounding box, temporal extent, parameter, parameter URI,

[46] SIOS, "SIOS OAI-PMH service for Blue-Cloud", https://bluecloud-sios.csw.met.no/csw.py?mode=oaipmh

[47] CF Community, "CF Standard Name Table", https://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html

[48] ESSI-Lab of CNR-IIA, "DAB GitHub page", https://github.com/ESSI-Lab/DAB

instrument, instrument URI, platform, platform URI, organization, organization URI, organization role, (metadata) datestamp, (resource) revision date, (resource) identifier
- **Blue-cloud recommended elements** keyword type, cruise, cruise URI, project, project URI

The use of URIs is essential for implementing a linked data approach, where ontologies can be referenced in the records, thereby removing ambiguities in meaning.

Adopting a brokering framework can also be beneficial for improving metadata quality. Once harmonization is complete, automatic components, known as metadata augmenters, can process the harmonized metadata to enhance it – such as completing missing elements by reasoning on the existing content. The Semantic Analyzer introduced in the next section could potentially perform these functions in the future.

3.2 Blue-cloud DAB deployment

Figure 1 shows the Blue-Cloud DAB deployment (in the middle), highlighting the connections with the BDIs on the left side and the consumer components on the right.

One accessor component for each BDI type has been developed and/or plugged in the Blue-Cloud DAB, specifically:

- **EuroArgo accessor**, brokering EuroArgo JSON based API. Each Argo platform is mapped to a dataset collection.
- **ELIXIR-ENA accessor**, brokering ELIXIR-ENA JSON based API. Each study is mapped to a dataset collection.
- **ELIXIR-MGnify accessor**, brokering ELIXIR-MGnify JSON based API
- **CSW-ISO accessor**, brokering both EMODNet Chemistry and SeaDataNet products OGC CSW ISO based services
- **ERDDAP accessor**, brokering both EMODnet Physics, EMSO ERIC and ICOS SOCAT ERDDAP services
- **EurOBIS accessor**, brokering EurOBIS linked data endpoint based on Turtle RDFS
- **OAI-PMH accessor**, brokering SIOS OAI-PMH service based on DIF model
- **ICOS portal accessor**, brokering ICOS Portal SPARQL endpoint
- **SeaDataNet CDI accessor**, brokering SeaDataNet-Open inventory service based on CDI metadata profile of ISO 19115

The Blue-Cloud DAB enables information to flow from the BDIs to the Blue-Cloud client components, specifically:

- **Orchestrator** Developed by MARIS, Responsible for harvesting metadata collections from the Blue-Cloud broker to provide users with a unified search experience through the Blue-Cloud web portal.
- **Semantic analyzer** Developed by the British Oceanographic Data Centre (BODC), the Semantic Analyzer is responsible for performing semantic analysis on the harmonized metadata [31, [32]. Specifically, free text elements are checked against well-known oceanographic ontologies, to identify potential concept URI matches, which are then communicated back to providers to improve the original records.
- **Metadata reports** Developed by CNR-IIA, a web application for visualizing metadata completeness supporting BDI implementation. It displays the total number of records, the percentage of core metadata elements available per BDI and sample core metadata element values.
- **Metadata analysis** OpenSearch data analysis service was connected to the DAB to specifically support the analysis detailed in this article.
- **Discovery test portal** Developed by CNR-IIA, a sample demo portal useful for BDIs to manually perform discovery of Blue-Cloud collections and verify that the mapping is correct.
- **GEOSS connector (not shown)** Developed by CNR-IIA, responsible for publishing Blue-Cloud records to the GEOSS initiative. Although this component was technically successfully tested, the operational dissemination of Blue-Cloud data to GEOSS has not yet officially started.

One profiler component for each Blue-Cloud client has been developed and/or plugged in the DAB taking care of mediation, specifically:

- **CSW-ISO profiler**, used to publish an OGC CSW ISO application profiler compliant service, allowing metadata harvesting through GetRecords requests by the Orchestrator component
- **Terms API profiler**, a JSON based OpenAPI described API to gather statistics about the DAB metadata content. The client can request the list of terms (values) used for a specific metadata element by a specific BDI in its records.
- **OpenSearch API profiler**, used to publish OpenSearch search engine standard, describing the allowed query templates and the JSON or GeoRSS[49] Atom response types
- **OpenSearch upload tool**, this component is used to transform Blue-Cloud records to JSON and upload them to an OpenSearch compliant service, given its endpoint and credentials

[49] OGC, "OGC GeoRSS Encoding Standard", https://www.ogc.org/standard/georss/.

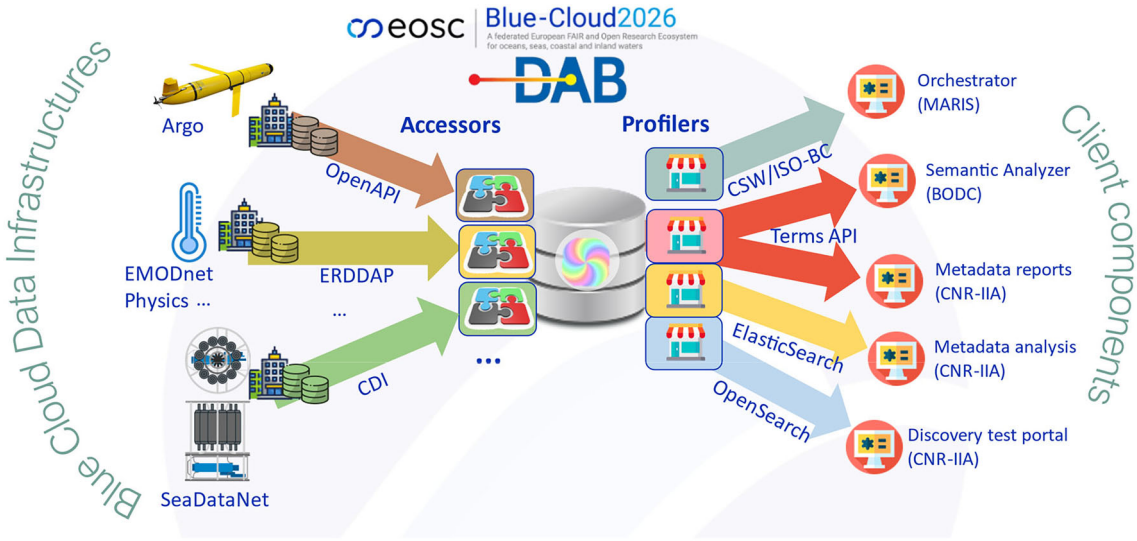


Fig. 1 Brokering approach implementation in Blue-cloud 2026. Blue-cloud DAB enables information to flow from BDIs shown in the left toward Blue-Cloud client components shown in the right

4 Analysis methodology and results

Metadata records provided by the BDIs are continuously harmonized by the DAB to a central database. To analyze them a snapshot, taken on 19 June 2024, was loaded into an OpenSearch service: a RESTful search and analytics suite that simplifies data ingestion, search, visualization and analysis.[50] A custom script was developed to execute queries against the OpenSearch service and generate the results presented in the following sections, enabling reproducibility thereby.

4.1 Metadata quantity

On June 19th, 2024, the platform hosted nearly 70,000 collections. Figure 2 represents the count of collection records shared by each BDI. Although 12 providers are currently contributing, it turns out that a single data provider (ELIXIR-ENA) shares more than half of the records (55.47%). The second-largest contributor, ARGO, shares more than a quarter of the records (28.07%). The third is SOCAT with 11.14%, while the remaining nine providers collectively contribute the rest (5.32%).

There is an obvious imbalance between the providers. This could be due to several factors, such as different levels of aggregation of the data sets. This disproportion could make records from smaller providers more difficult to find compared to those from the larger contributors. Assuming a similar distribution of records across providers, applying general search criteria (such as spatial or temporal extent) would likely return more records from the larger providers. However, this problem is mitigated when applying a thematic query (e.g., a keyword or parameter search) that can only be satisfied by certain data providers. The overall completeness and validity of the metadata also suffer from this disproportion.

4.2 Metadata completeness

As shown in Fig. 3, the core metadata elements are generally well-represented, with some notable exceptions: spatial extent is missing in 35.74% of the dataset collections, project information in 21.5%, instrument details in 14.65%, temporal extent in 8.79%, and platform information in 3.05% of the dataset collections.

The use of URIs remains very limited in metadata values: they are predominantly used for organizations (in 4.43% of the dataset collections), keywords (3.48%), parameters (2.04%), and instruments (1.87%). This limited adoption captures a significant area for improvement to enhance data interoperability and reduce ambiguities at this stage of the project, that should be substantially filled at the end of the project.

Figure 4 highlights the champion providers in the use of URIs. SeaDataNet Open predominantly uses URIs to describe all key elements, utilizing NVS vocabularies [8], the EDMO, and the INSPIRE theme registers.[51] The EurOBIS data provider follows closely, with significant use of concept URIs for keywords (i.e. marine species vocabulary), organizations (i.e. marineinfo.org vocabulary), and instruments

[50] OpenSearch, "Homepage", https://opensearch.org/

[51] INSPIRE, "INSPIRE theme register", https://inspire.ec.europa.eu/theme

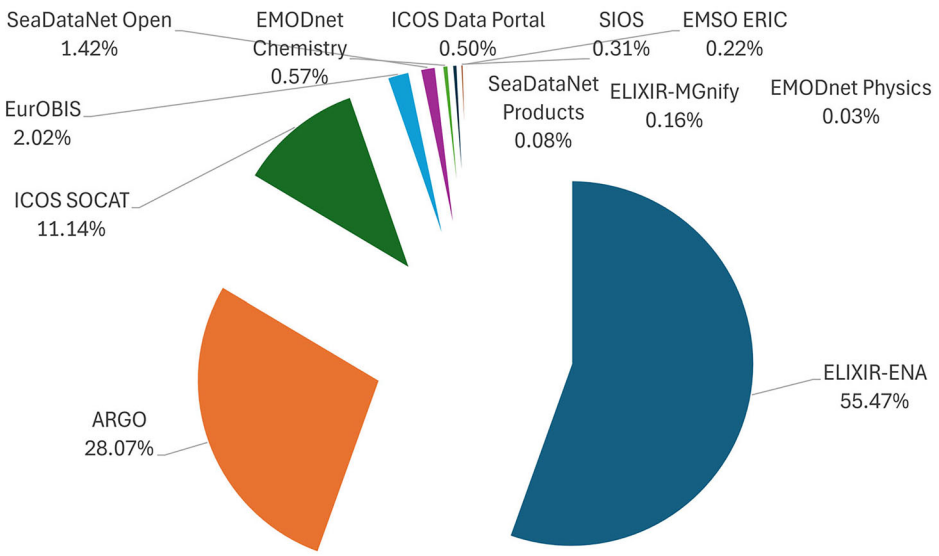


Fig. 2 Percentage of records by BDI

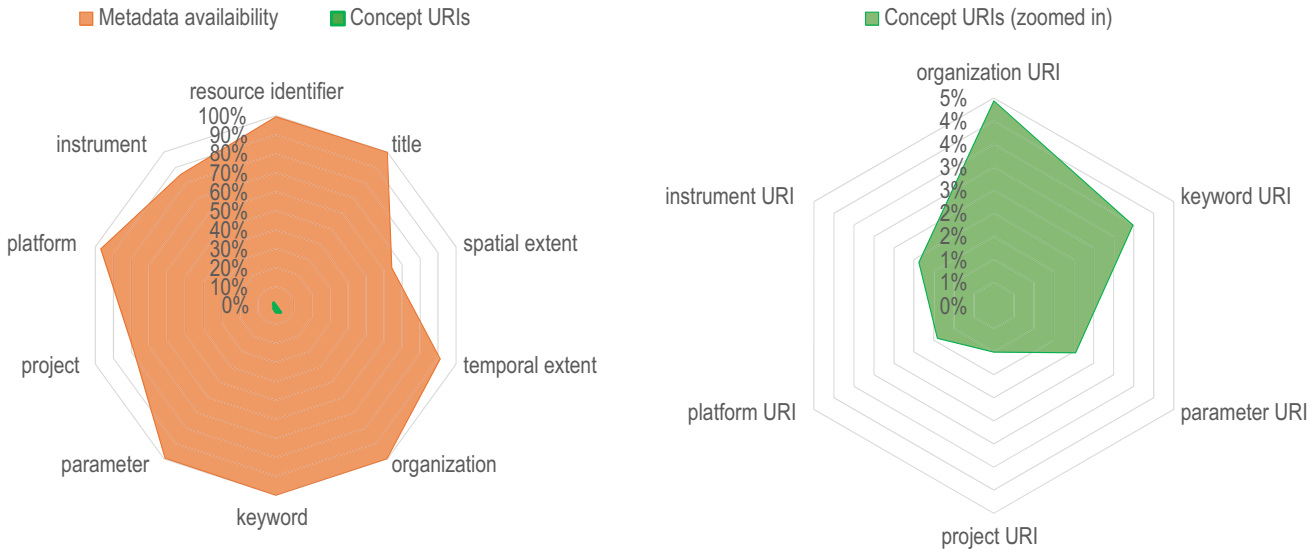


Fig. 3 Availability of core metadata elements. The second spider diagram on the right provides a more detailed view, specifically zooming in on the availability of concept URIs

(i.e. NVS vocabularies). EMODnet Chemistry also employs URIs to describe keywords (i.e. INSPIRE Themes and NVS vocabularies), parameters (i.e. NVS vocabulary P02), and organizations (i.e. EDMO vocabulary). The ICOS Data Portal uses ICOS community vocabularies for organizations and projects. The EDMO vocabulary for organizations is also used by EMSO-ERIC and SeaDataNet products, making it the most used vocabulary, adopted by six data providers. Finally, SeaDataNet products uses NVS vocabularies for parameters and INSPIRE themes for keywords.

Figure 5 analyzes the type of missing items by source. The graph is normalized to improve readability, also highlighting elements with few missing values. In the case of EurOBIS, some elements seem to be missing frequently, but this is probably due to ongoing efforts to improve the metadata model that characterizes the shared data: in some cases, the records have not yet been fully updated, while in other cases the connection to the broker has not yet been fully established. Examining the absolute values, the most significant quantities of EurOBIS missing data pertain to platforms (1,359 occurrences), instruments (1,046 occurrences), and projects (1,359 occurrences).

In terms of volume, however, ELIXIR-ENA appears to have the most records in need of improvement, with 22,691 records missing spatial extents, 4,709 records missing temporal extents, and 11,542 records lacking project information. Further investigation is required, in collaboration with the

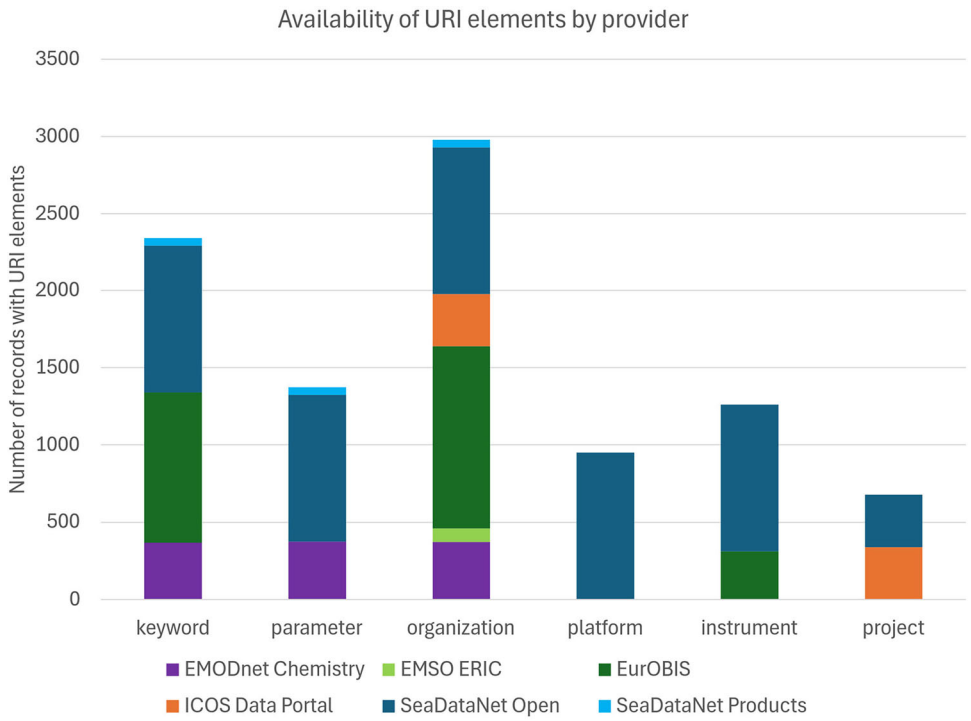


Fig. 4 Linked data elements by provider

provider, to determine the reasons for these low percentages. Since ELIXIR-ENA is focused on a domain that is somewhat different from typical marine observations (nucleotide sequencing), it is possible spatial and temporal information is often not strictly relevant for the study purposes and, as a result, is not typically recorded by the originators.

ICOS SOCAT also has a substantial number of missing instrument values (7,484 records), as well as missing temporal extents (615 records) and spatial extents (615 records). For SIOS, it appears that resource identifiers (e.g., dataset collection DOIs) are currently missing in 211 cases. Common metadata elements such as the missing title could be easily fixed by EMODnet Chemistry (12 records) and Argo (2 records).

4.2.1 Introducing availability and validity indicators

The analysis revealed a limited but notable absence of core metadata elements and a low percentage of core metadata elements available as concept URIs. A set of metadata quality indicators is proposed to address the identified issues with the aim to assess and monitor metadata availability and validity over time:

- **Indicator 1** percentage of records containing the Ω core metadata element, where Ω is one of the following: resource identifier, title, spatial extent, temporal extent, keyword, parameter, organization, platform, instrument, or project
- **Indicator 2** percentage of records containing the Δ metadata element as a concept URI, where Δ is one of the following: keyword, parameter, organization, platform, instrument, or project

Table 1 Availability indicators per metadata element: (1) percentage of documents with elements available and (2) percentage of document with elements available as concept URIs

Ω/Δ	Indicator 1 (%)	Indicator 2
Resource identifier	99.46	N/A
Title	99.93	N/A
Spatial extent	64.26	N/A
Temporal extent	91.20	N/A
Keyword	99.96	3.48%
Parameter	99.31	2.04%
Organization	99.73	4.43%
Platform	96.94	1.41%
Instrument	85.34	1.87%
Project	78.50	1.01%

Table 1 presents the status of the two availability indicators, time being. The aim is to reassess these indicators at the project's conclusion. Two potential targets could be set to halve the gap to completion of Indicator 1 for all elements and to double the values of Indicator 2 for all elements, respectively.

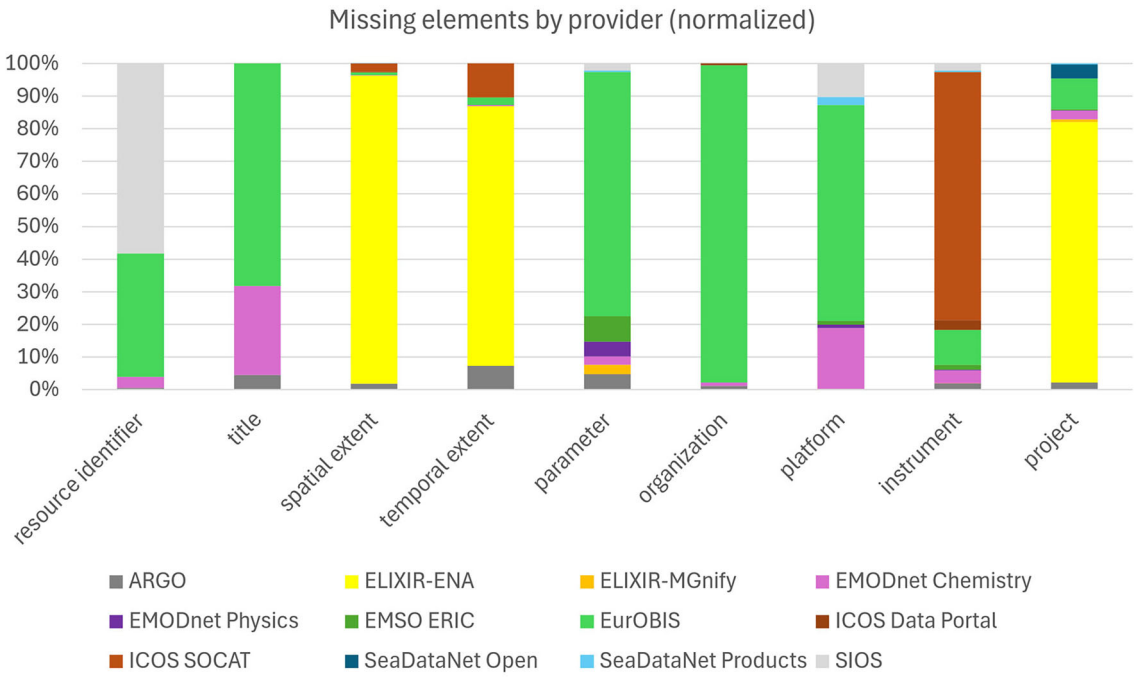


Fig. 5 Missing elements by provider

The analysis also identified issues with the metadata content, specifically with the values of certain metadata elements. For example, invalid bounding box values were detected, such as coordinates outside the allowed range (i.e. $-180 <= \text{longitude} <= 180$ and $-90 <= \text{latitude} <= 90$). In some cases, incorrect positions were published, while in other cases, the semantics was unclear (for example, a value of –99.99 was used by a data provider to indicate a missing value rather than an actual position; this semantics wasn't correctly interpreted by the DAB). Other issues regarded very short resource identifiers and titles, (with length less than four characters) and invalid temporal extents. These issues might stem from errors in the original data publication or in the brokering implementation. The identified issues were analyzed and, when necessary, reported to data providers to improve the original data publication and the implementation of the connection with the broker.

We propose also to achieve other indicators to easily verify targets for ensuring the validity of the content of important metadata elements by the project's end (the goal is to reach 0% for all of them):

- **Indicator 3** percent of records characterized by a Resource Identifier of less than four characters which is invalid
- **Indicator 4** percent of records characterized by a Title of less than four characters which is invalid
- **Indicator 5** percentage of records with invalid Spatial Extent, with latitude values outside allowed range (i.e. -90 and 90 degrees and longitude values outside -180 and 180 degrees)
- **Indicator 6** percentage of records with invalid Temporal Extent, with the end date less than the start date. In general, special attention should be given to temporal ranges before the year 1000 CE (which are unlikely to represent historical data), and after the current year (which are unlikely unless they pertain to model forecasts).

Table 2 Four basic indicators for metadata validity of (3) resource identifier, (4) title, (5) spatial extent and (6) temporal extent metadata elements

Indicator 3	Indicator 4	Indicator 5	Indicator 6
0.21%	0.01%	0.02%	0.02%

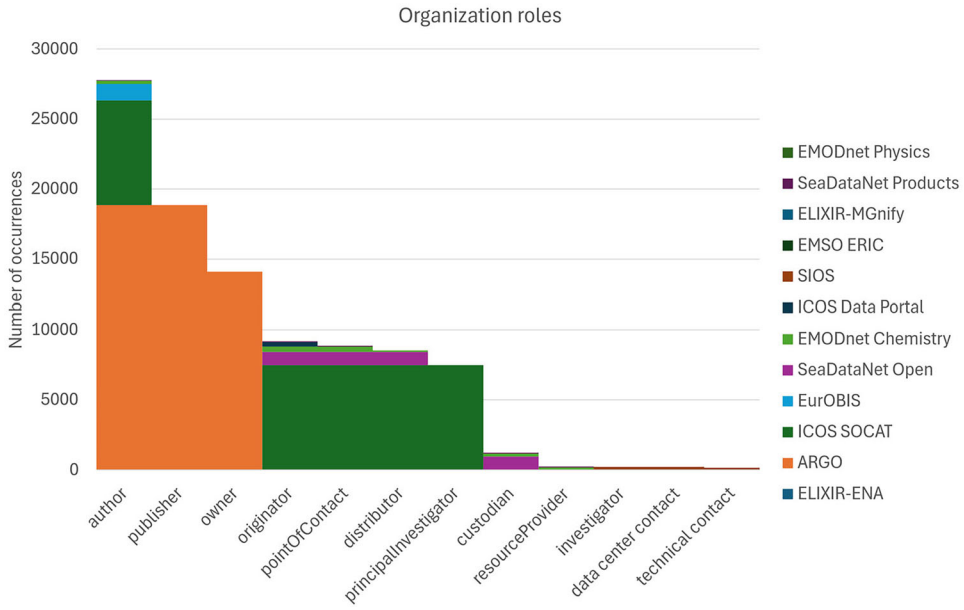
Table 2 shows the status of these indicators, time being. The goal is to reassess them at the project's conclusion, when a significant decrease is expected.

Of course, other indicators could be added as more invalid data types are discovered, along with their test definition.

4.3 Cited organizations

There are 126,348 organizations cited as responsible parties in the analyzed metadata records; each citation can include: an organization name, a URI, and a role. The most cited roles

Fig. 6 Organization roles



(which however are often missing) are distributed according to Fig. 6.

Figure 7 shows the ten most cited organization names. The large number of occurrences is influenced by duplicates, as in the case of "AOML" and "NOAA AOML". To avoid these ambiguities, Blue-Cloud 2026 is putting efforts on metadata curation, in order to reference concept URIs from well-established registries: for example, EDMO is often used in the European marine community, while the Research Organization Registry (ROR) is worth mentioning as a global, community-led registry for research organizations [33, 34].

4.4 Thematic coverage

A valuable practice to characterize the theme of a dataset collection is the use of keyword elements, which must be provided by data providers, as textual information, to facilitate dataset discovery and evaluation. Figure 8 shows the 200 most frequently used keywords, along with their occurrence counts. The colors represent the data provider originating each keyword. The diagram offers insights into the distribution of keywords among providers, highlighting potential areas for improvement, in collaboration with BDIs, to enhance users search experience. The first two significant spikes correspond to the keywords "ELIXIR-ENA" (with 37,270 occurrences) and "ARGO" (with 18,859 occurrences). This can be explained by the fact that each record typically includes a keyword corresponding to the name of the BDI, which is often present in the original metadata but sometimes added by the broker to allow searches based on the provider's name.

Upon closer examination of the most frequently used keywords, a few were found to hold little to no informational content, acting as placeholders such as "missing" (2,469 occurrences) and numerical identifiers such as "31" (8,767 occurrences).

After these initial spikes a plateau is noticeable in the graph, corresponding to keywords from the ICOS-SOCAT BDI. This unusual pattern reflects that all 7,484 ICOS-SOCAT records share the same 154 keywords.

Labeling all records with the same keywords will be ineffective for users trying to select a specific subset of interest. Additionally, if a provider's records have an excessive number of keywords, these records may dominate the overall discovery space, being returned more frequently than others, potentially harming the performance of the entire infrastructure. As noted in SEO practices [35], "using massive numbers of keywords in your content may achieve short-terms gains", but the long-term goal of the Blue-Cloud ecosystem is to improve the performance of user searches, by increasing the relevance of returned records and reducing the number of undesired results. Addressing this issue is therefore a priority.

Further investigation revealed that the BDI provides a single metadata record to describe its entire information content at the BDI level. This record is used as a template for all records in the DAB mapping, which is further modified using metadata at dataset level, such as spatial extent. However, because keywords are not provided at the dataset level, the resulting keywords are derived from the long list present at the BDI level.

The decision to use these BDI-level keywords for all the records was made to benefit the specific BDI with the best

Fig. 7 Top 20 organizations

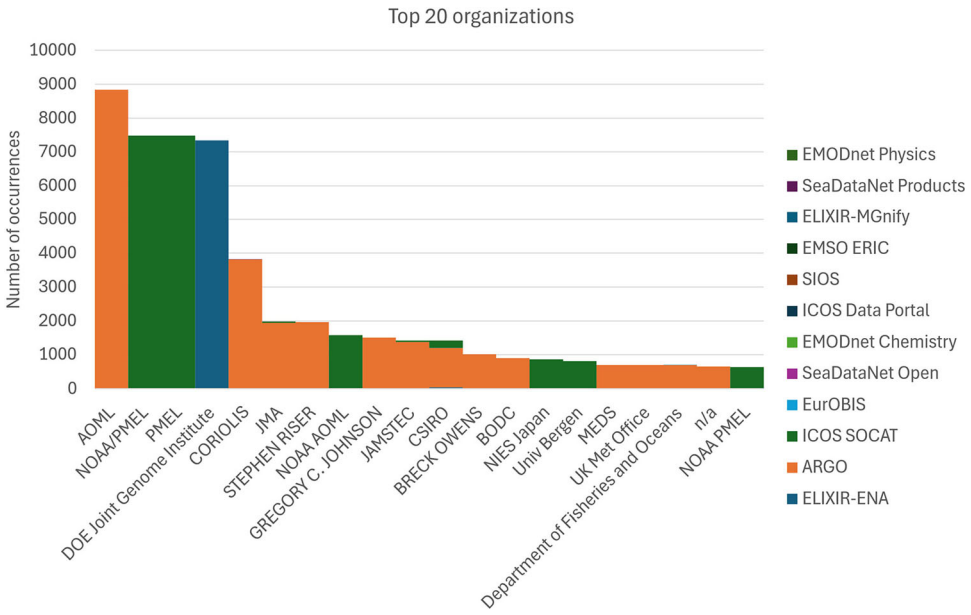
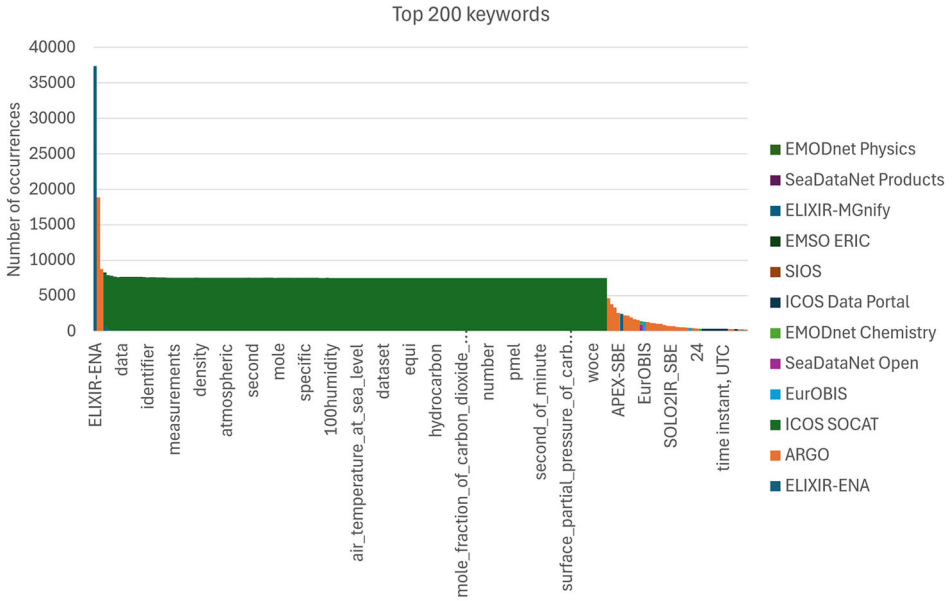


Fig. 8 Top 200 keywords



effort using the available metadata, albeit at the expense of the broader ecosystem. Another option would have been to exclude keywords entirely, which would disadvantage the specific BDI but benefit the ecosystem. Now that the issue has been identified, the best course of action is to contact the BDI, explain the issue and work together on a solution.

To make the keyword analysis more meaningful, a filter was applied to remove selected keywords: data provider names, the "missing" keyword, two digits keywords and all records from the "ICOS-SOCAT" provider. The results, focusing on the top 75 keywords for increased readability, are shown in Fig. 9.

Keywords from Argo dominate the first part of the graph, as expected given that Argo is the second largest provider. These keywords were automatically (and somewhat arbitrarily) extracted by the DAB from other Argo fields (such as sensor model, maker and other identifiers), since Argo does not provide a specific keyword field. After the initial portion of the graph, keywords from different BDIs begin to appear, intermingling with one another. These keywords belong to different categories, including the observation medium (e.g., "sea water"), observed parameters (e.g. "air pressure"), or geographical features like sea names (e.g., "Mediterranean", "Black Sea").

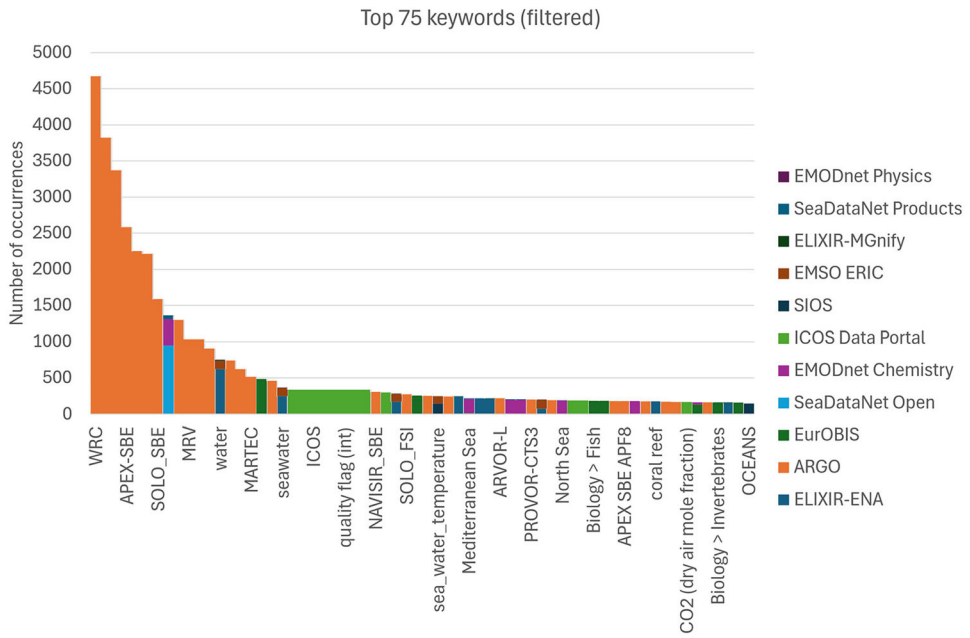


Fig. 9 Top 75 keywords, after a filter to remove selected keywords for increased readability has been applied

One issue highlighted by this analysis is again the presence of syntactic variations between similar terms, which stems from the lack of standardized URIs (e.g., "sea_water", "seawater", "sea water", …). This inconsistency can hinder the effectiveness of keyword searches and data discovery.

4.5 Parameters

The 20 most present parameters are distributed according to Fig. 10. Amongst the most present, "SUBSURFACE PRESSURE", "WOCE flag for aqueous CO2" and "marine metagenome", coming respectively from Argo, ICOS SOCAT and ELIXIR-ENA, which dominates this plot.

4.6 Instruments

The top 20 instruments are shown in Fig. 11. Argo and ELIXIR-ENA equally dominate this plot, respectively with "DRUCK_2900PSIA" and "Illumina HiSeq 2500" instruments. SeaDataNet is the third present BDI, but ranks only 31th with "CTD".

4.7 Platforms

The top 20 platforms are shown in Fig. 12. "ILLUMINA" is the most frequent platform, largely dominating, because of occurring in most of the records of ELIXIR-ENA (the largest BDI). Argo follows with the "APEX Profiling FLOAT". SeaDataNet, the third most prominent BDI in this graph ranks only 11th place with the platform "research-vessel". It is important to note that the list includes both platform categories and specific vessel names, such as "Nuka Arctica" from ICOS Data Portal, highlighting again differences in granularity.

4.8 Temporal coverage

As shown by Fig. 13, the considered dataset collections cover a temporal range starting as early as 1700 with historical data, gradually increasing until the year 2000, when data availability accelerates, reaching a peak in 2016 (10,462 records), 2017, and 2018. Availability then decreases to the current partial year, which has 4,537 dataset collections. Essentially, no dataset collections report future dates (except for two dataset collections that have incorrect dates), confirming that there is no forecast data included beyond today's date.

4.9 Spatial coverage

The map in Fig. 14 is a further result of the analysis and shows the grid of occurrence of the georeferenced data collections that are available with respect to their terrestrial geolocation. The resolution of the map (i.e., 1 × 1 degree) was chosen empirically to provide a general overview of the overall contribution of records with respect to different marine areas.

Each grid square shows the records contributing to that specific area, based on the metadata of the record itself. Brighter areas indicate a higher number of occurrences, between 4 and 1536. As expected, marine records are generally more present than terrestrial records, with particularly predominant coverage over the north-central Atlantic Ocean.

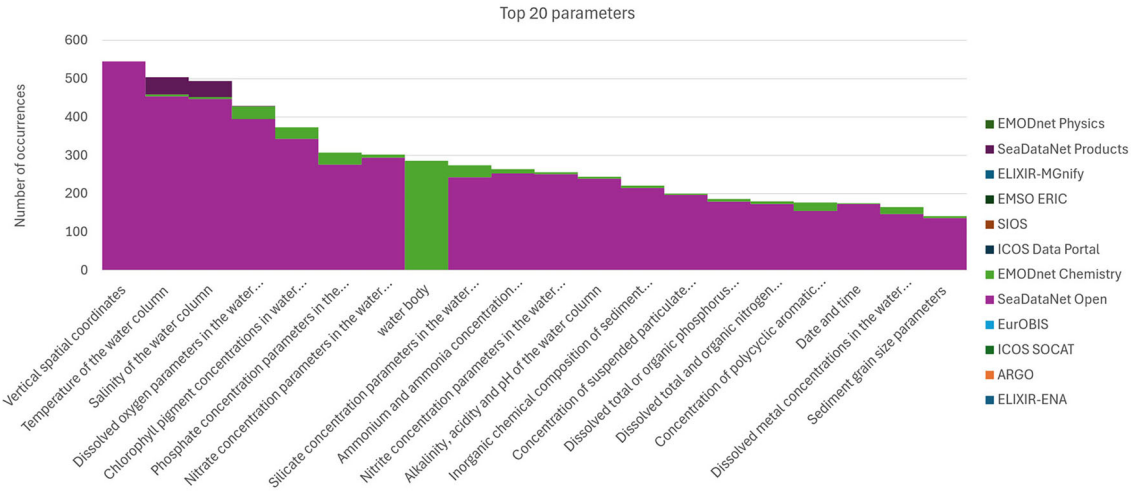
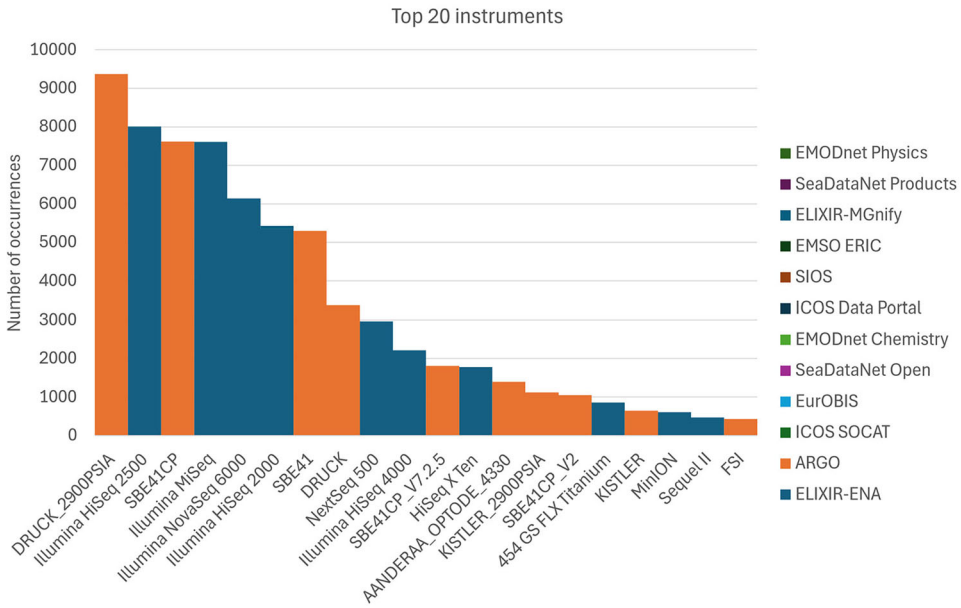


Fig. 10 Top 20 parameters

Fig. 11 Top 20 instruments



Coastal areas are also well represented, which may indicate more coastal monitoring than offshore regions. The presence of any inaccurate land cover in marine data collections could be reduced by using higher resolution, or by using polygons instead of rectangles.

5 Analysis discussion

The analysis performed concerns metadata published by different BDIs, once these have been harmonized, according to the Blue-Cloud marine metadata profile, by means of a brokerage service. The goal is to enable ecosystem users to search the available marine data collections more easily and efficiently—leaving BDIs free to evolve over time.

Previous studies in this area have examined the metadata content of specific systems, such as the following contributions to the GEOSS initiative: NextGEOSS [36], Eurac Research [37], and China satellite data [38]. Additionally, previous research has explored system-of-systems approaches, like the GEOSS Clearinghouse content analysis via the Rubric-Q tool [39] and the exploration of the GEOSS brokering platform content [40, 41], as well as the on-going FAIR metadata assessments by DataONE initiative.[52]

Gaps in Blue-Cloud metadata quality have been identified, particularly regarding the availability of metadata elements and the validity of metadata content. To quantitatively assess

[52] DataOne, "Make your data FAIR", https://www.dataone.org/fair/

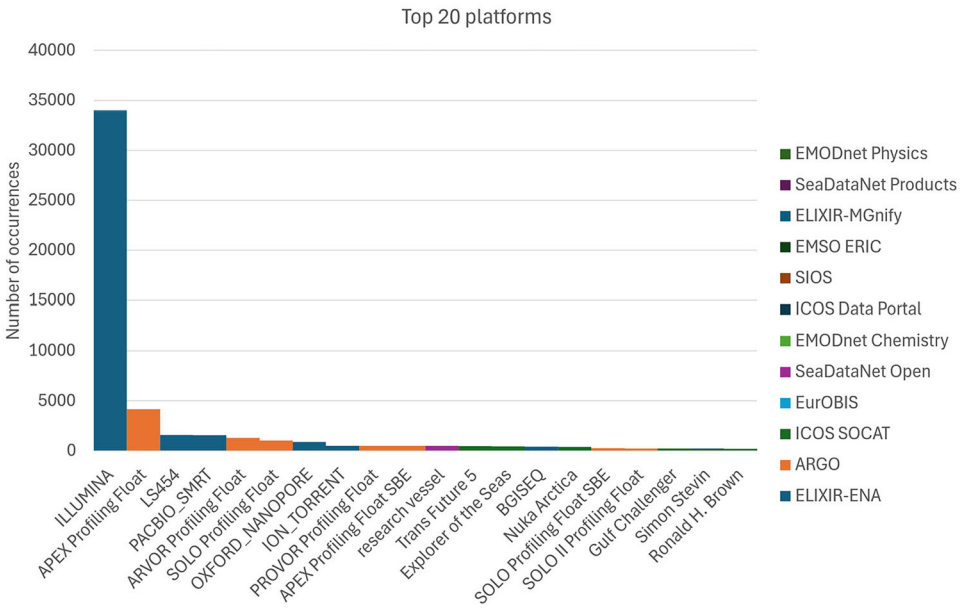


Fig. 12 Top 20 platforms

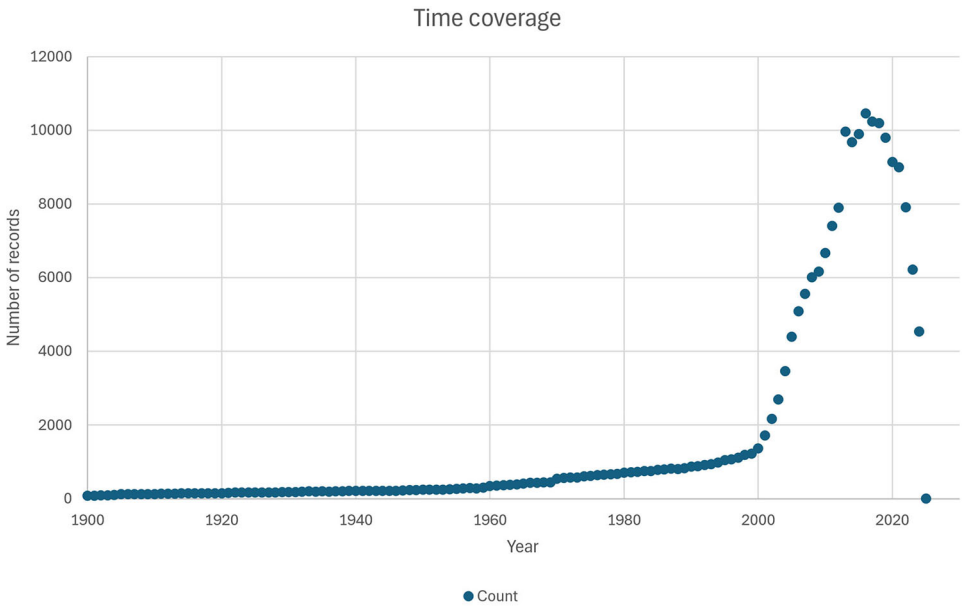


Fig. 13 Temporal coverage: the plot represents the number of records with temporal ranges that intersect each year, from 1900 to 2025. The few historical records prior to 1900 are excluded from the view to better appreciate the trend in recent years

the current state and track progress over time, a set of metadata quality indicators are proposed, drawing on approaches similar to those used in other initiatives, such as the service reports of the network Common Data Form (NetCDF) Attribute Convention for Dataset Discovery (ACDD)[53] and the Key Performance Indicators of WMO Information System (WIS) [42].

Several issues that could hinder effective user searches have been identified, notably the discrepancies across BDIs in data quantity and granularity result in the top two providers (ELIXIR-ENA and Argo) accounting for 83, 81% of the records, overshadowing contributions from other BDIs. Overall metadata quality results are significantly impacted by this disproportion, as the two top providers are not using concept URIs and one of them is missing important metadata elements. Whereas a large use of concept URIs is notable in other BDIs: SeaDataNet Open, EurOBIS, EMODnet Chemistry and SeaDataNet Products.

Missing metadata elements—such as spatial extent (35.74%), project (21.5%), instrument (14.65%), temporal extent (8.79%), platform (3.05%)—limit the effectiveness of user searches and the accurate evaluation of results. In particular, the lack of spatiotemporal localization of the data prevents its use in the case of geographic or transformational

[53] Earth Science Information Partners (ESIP).

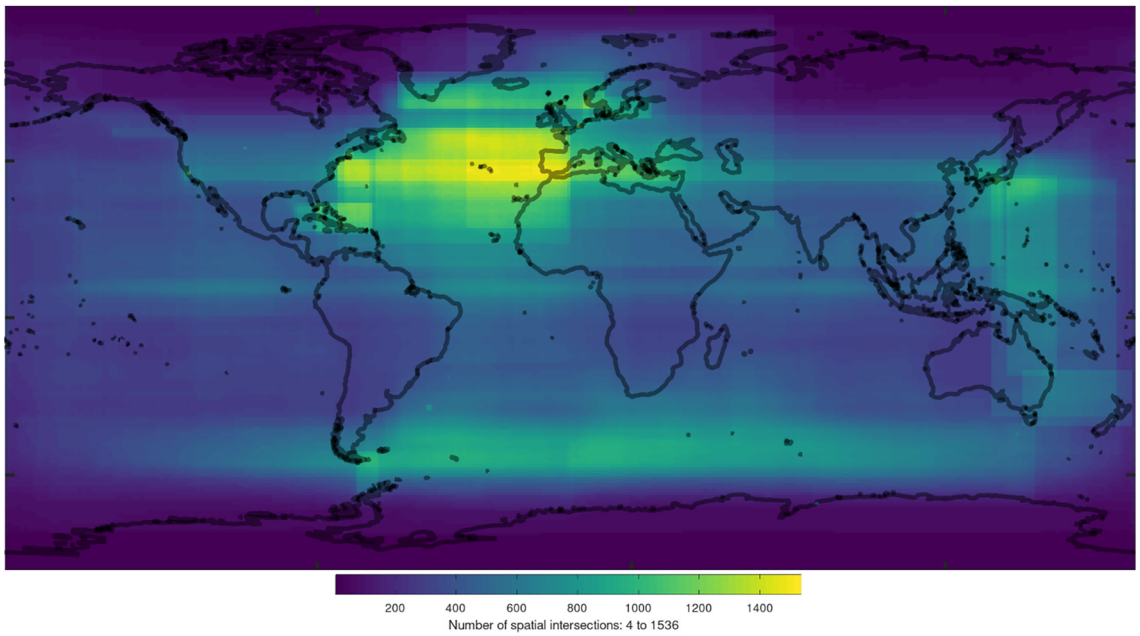


Fig. 14 Spatial coverage: each grid square on the map represents a one-degree resolution area, with the map using EPSG:4326 coordinate reference system (CRS). The color of each square reflects the number of records with spatial extents intersecting that area, ranging from 4 to 1536. Brighter squares indicate higher numbers of intersecting records

analysis of phenomena that may interest society and policy makers.

Alternate spellings and misspellings emerged to be common, largely due to the limited presence of concept URIs for metadata elements, which further hamper effective searches. Metadata curation (a process that improves metadata based on existing content and possibly aided by external vocabularies, ontologies and knowledge bases [43]) has been explored in other works [44, 45], and could be implemented in Blue-Cloud using the Semantic Analyzer. This component provides concept URIs that BDIs can already use to improve metadata publication at the source. The Semantic Analyzer can effectively interact with the DAB API to match existing free text against common marine ontologies. In the future, it could also be used to automatically enhance broker-level metadata.

An additional problem related to metadata content can be seen by observing the unusual distributions of metadata values. A significant example is the distribution of keywords in ICOS-SOCAT BDI, which are numerous and identical across all records. This raises questions about the potential conflict between the goals of data providers-who may use SEO-type techniques [35] to ensure that their records rank well in searches-and the goals of ecosystems such as Blue-Cloud, which aim instead to ensure fair and relevant search results among all providers, returning only the most relevant records to users. Generally, this is a typical example of the belonging-vs-autonomy conflict that each system contributing to an ecosystem must deal with.

6 Conclusions and way forward

The implementation of the marine data broker has enabled unified search of diverse marine data collections, facilitated outreach toward initiatives such as GEOSS, and established a platform for marine metadata analysis and quality assessment that benefits providers.

An information system is only as good as the information it contains. Therefore, metadata quality and related indicators play a crucial role in Blue-Cloud and in this work, particularly regarding metadata availability and validity. Future work will focus on refining and extending these indicators, including the analysis of the distribution of metadata values across providers.

To facilitate this enabling ecosystem process, metadata quality reports have been made available as a service to each BDIs, while invalid content was highlighted through direct communication. Future work will also improve the graphical presentation of the reports to better communicate areas for improvement to providers.

Metadata analysis has already proven valuable in improving the overall quality of metadata within the ecosystem. For example, important corrections were made after invalid values, such as spatial coordinates, were reported.

An important conclusion of the analysis conducted is the recommendation to review the project's conclusion to demonstrate the expected improvements in metadata quality by all providers. This is in line with Blue-Cloud 2026's goal of improving the publication of marine data and its FAIRness. Two primary goals were identified for BDIs: completing the missing metadata elements and using concept URIs from ontologies instead of free text. Another key objective will be to evaluate the effectiveness, from the user's perspective, of the unified (ecosystem-wide) search compared to individual searches conducted at each data provider.

We recognize that achieving these targets will not always be straightforward, as providers often aggregate data from other organizations. Adding missing metadata may require contacting the original data source, which may not always be feasible or may involve information that is lost. In such cases, providing an explanation for the missing metadata element could be beneficial.

The BODC Semantic Analyzer component provides support to data providers, as free-text elements in the harmonized metadata can be evaluated against existing vocabularies and ontologies, with suggested matching concept URIs to enhance metadata quality. The Semantic Analyzer could also be more tightly coupled with the DAB, potentially empowering DAB harmonization through tentative semantics mappings. While this experimental approach could deliver semantically enriched results to the end user, it carries the risk of introducing errors through incorrect mappings.

Documenting the mapping from the BDI to the harmonized model in a more formal manner is another area that could be further investigated in future work, particularly with regard to the openness and FAIRness of the results.

7 Glossary

Term	Definition
ACDD	Attribute convention for dataset discovery
API	Application programming interface
BDI	Blue data infrastructure
BODC	British oceanographic data centre
Broker [13]	An intermediary middleware dynamically implementing a many-to-many interconnection for a client–server framework
CDI	Common data index
CF	Climate and forecast
CMEMS	Copernicus marine environmental monitoring service
CNR-IIA	Institute of atmospheric pollution research of national research council of Italy
Crosswalk [27]	Mapping of the elements, semantics, and syntax from one metadata scheme to those of another
CRS	Coordinate reference system
CSR	Cruise summary report
CSW	Catalogues service for the web
DAB	Discovery and access broker
Dataset [29]	Identifiable collection of data
Dataset collection	Set of datasets sharing the same product specification
DCAT	Data catalog vocabulary
DIF	Directory interchange format
EDMERP	European directory of marine environmental research project
EDMO	European directory of marine organisations
EMBL-EBI	European bioinformatics institute of the European molecular biology laboratory
EMODnet	European marine observation and data network
ENA	European nucleotide archive
EOSC	European open science cloud
EMSO	European multidisciplinary seafloor and water column observatory
ERIC	European research infrastructure consortium
ESIP	Earth science information partners
ESSI-Lab	Earth and space science informatics laboratory
EU	European union
EurOBIS	European node of the international ocean biodiversity information system
FAIR	Findability, accessibility, interoperability, and reusability

Term	Definition
GEOSS	Global Earth observation system of systems
ICOS	Integrated carbon observation system
IFREMER	French national institute for ocean science and technology
INSPIRE	Infrastructure for spatial information in the European community
ISO	International organization for standardization
JSON	Javascript object notation
LOV	Laboratoire d'océanographie de Villefranche
MARIS	Marine information services
Mapping [28]	Correspondence between instances of one model and instances of another model that represent the same meaning
Metadata [29]	Data about data
Metadata element [29]	Discrete unit of metadata
Model [46]	Abstraction of some aspects of a universe of discourse
NOAA	National oceanic and atmospheric administration of the United States of America
NERC	Natural environment research council
NetCDF	Network common data form
NVS	NERC vocabulary server
OAI-PMH	Open archives initiative protocol for metadata harvesting
ODIP	Ocean data interoperability platform
OGC	Open geospatial consortium
OGS	National institute of oceanography and applied geophysics of Italy
PMEL	Pacific marine environmental laboratory
RDFS	Resource description framework schema
ROR	Research organization registry
SDG	Sustainable development goal
SIOS	Svalbard integrated Arctic Earth observing system
SOCAT	Surface Ocean CO_2 atlas
SPARQL	SPARQL protocol and RDF query language
UIB	University of Bergen
URI	Uniform resource identifier
VLIZ	Flanders marine institute
VRE	Virtual research environment
WHOS	WMO hydrologic observing system
WIS	WMO information system
WMO	World meteorological organization
WoRMS	World register of marine species
XML	Extensible markup language

Acknowledgements This research was funded by Blue-Cloud 2026 project from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094227. The authors would like to thank Massimiliano Olivieri for his contributions to the cloud infrastructures configuration and management; Lena Rettori and Eleonora Livi for their assistance as scientific secretaries; Lena Rettori for providing helpful support in proofreading the article; Alexandra Kokkinaki and Gwenaelle Moncoiffé from BODC for the useful discussions and collaboration regarding Blue-Cloud metadata and the Semantic Analyzer. Figure 1 has been designed using icon resources from Flaticon.com FreePik, "FlatIcon", https://www.flaticon.com/

Author contributions E.B. is the main author, producing most of the content. S.N. provided support in shaping introduction and conclusions, R.R. contributed with the BDI section and all authors reviewed and provided useful suggestions.

Funding Open access funding provided by IIA - MONTEROTONDO within the CRUI-CARE Agreement. Horizon 2020 Framework Programme, grant agreement 101094227.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. UN General Assembly: Transforming our world: the 2030 agenda for sustainable development (2015). Accessed on 13 Aug 2024. [Online]. Available: https://www.refworld.org/legal/resolution/unga/2015/en/111816
2. Commission, E. et al., Mission area, healthy oceans, seas, and coastal and inland waters – foresight on demand brief in support of the Horizon Europe mission board. Publications Office of the European Union (2021). https://doi.org/10.2777/054595
3. Commission, E. and D.-G. for R. and Innovation, strategic research and innovation agenda (SRIA) of the European open science cloud (EOSC). Publications Office of the European Union (2022). https://doi.org/10.2777/935288
4. Laney, D.: 3D data management: controlling data volume, velocity and variety. META group research note, 6(70) (2001)
5. Schaap, D.: Blue-cloud 2026–D2.1 existing DD&AS and blue data infrastructures – review and specifications for optimisation report. Zenodo (2023). https://doi.org/10.5281/zenodo.10438757
6. Schaap, D.: Blue-cloud 2026–D2.2 new blue data infrastructures – service analysis report. Zenodo (2023). https://doi.org/10.5281/zenodo.10438608
7. Richardson, L., et al.: MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Res. **51**(D1), D753–D759 (2023). https://doi.org/10.1093/nar/gkac1080
8. Leadbetter, A.M., Lowry, R.K., Clements, D.O.: Putting meaning into NETMAR - the open service network for marine environmental data. Int. J. Digit. Earth **7**(10), 811 (2014). https://doi.org/10.1080/17538947.2013.781243
9. Kokkinaki, A. Moncoiffé, G., and Le Franc, Y.: Fair semantics and the NVS. Bull. Geophys. Oceanogr, 62 (2021)
10. Wilson, C., Robinson, D., and Simons, R. A.: Erddap: providing easy access to remote sensing data for scientists and students. In: International geoscience and remote sensing symposium (IGARSS) (2020). https://doi.org/10.1109/IGARSS39084.2020.9323962
11. Bernot, J. et al.: World register of marine species (WoRMS), WoRMS editorial board (2024). [Online]. Available: https://www.marinespecies.org
12. Schaap, D., Boldrini, E., Nativi, S., Tosello, V., and Fichaut, M.: Ocean data standards volume 8: SeaDataNet common data index (CDI) metadata model for Marine and oceanographic datasets (including SeaDataNet CDI metadata profile of ISO 19115, V12.2.0) (2021). https://doi.org/10.25607/OBP-1961
13. Nativi, S., Craglia, M., Pearlman, J.: Earth science infrastructures interoperability: the brokering approach. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **6**(3), 1118 (2013). https://doi.org/10.1109/JSTARS.2013.2243113
14. Boldrini, E., Craglia, M., Mazzetti, P., and Nativi, S.: The brokering approach for enabling collaborative scientific research (2014). https://doi.org/10.4018/978-1-4666-6567-5.ch014
15. Nativi, S., Mazzetti, P., Santoro, M., Boldrini, E., Manzella, G.M.R., Schaap, D.M.A.: CDI/THREDDS interoperability in the SeaDataNet framework. Adv. Geosci. (2010). https://doi.org/10.5194/adgeo-28-17-2010
16. Salas, F.R., Boldrini, E., Maidment, D.R., Nativi, S., Domenico, B.: Crossing the digital divide: an interoperable solution for sharing time series and coverages in Earth sciences. Nat. Hazards Earth Syst Sci (2012). https://doi.org/10.5194/nhess-12-3013-2012
17. Boldrini, E., Luzi, D., Nativi, S., Pecoraro, F.: Harmonising CERIF and INSPIRE metadata models to support multidisciplinary data sharing. Int. J. Metadata Semant. Ontol.Semant. Ontol. (2015). https://doi.org/10.1504/IJMSO.2015.070831
18. Nativi, S., Mazzetti, P., Santoro, M., Papeschi, F., Craglia, M., Ochiai, O.: Big data challenges in building the global earth observation system of systems. Environ Model Softw.Softw. **68**, 1–26 (2015). https://doi.org/10.1016/j.envsoft.2015.01.017
19. Glaves, H. M.: Developing a common global framework for marine data management (2016). https://doi.org/10.4018/978-1-5225-0700-0.ch003
20. Pearlman, J., Schaap, D., and Glaves, H.: Ocean data interoperability platform (ODIP): addressing key challenges for marine data management on a global scale. In: OCEANS 2016 MTS/IEEE Monterey, OCE 2016 (2016). https://doi.org/10.1109/OCEANS.2016.7761406
21. Boldrini, E., Nativi, S., Pecora, S., Chernov, I., Mazzetti, P.: Multi-scale hydrological system-of-systems realized through WHOS: the brokering framework. Int. J. Digit. Earth **15**(1), 1259–1289 (2022). https://doi.org/10.1080/17538947.2022.2099591
22. ESSI-LAB of CNR-IIA: Discovery and access broker (DAB) community edition. GitHub (2021). Accessed on 13 Aug. 2024. [Online]. Available: https://github.com/ESSI-Lab/DAB
23. Nativi, S., Bigagli, L., Mazzetti, P., Boldrini, E., and Papeschi, F.: GI-cat: a mediation solution for building a clearinghouse catalog service. In: Proceedings of the international conference on advanced geographic information systems and web services. GEOWS 2009 (2009). https://doi.org/10.1109/GEOWS.2009.34
24. Nativi, S., Bigagli, L., Mazzetti, P., Mattia, U., and Boldrini, E.: Discovery, query and access services for imagery, gridded and coverage data a clearinghouse solution. In: International geoscience and remote sensing symposium (IGARSS) (2007). https://doi.org/10.1109/IGARSS.2007.4423731
25. Nativi, S., Bigagli, L.: Discovery, mediation, and access services for earth observation data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **2**(4), 233 (2009). https://doi.org/10.1109/JSTARS.2009.2028584
26. Santoro, M., Mazzetti, P., Nativi, S.: Virtual earth cloud: a multi-cloud framework for enabling geosciences digital ecosystems. Int. J. Digit. Earth **16**(1), 43 (2023). https://doi.org/10.1080/17538947.2022.2162986
27. ISO: ISO/IEC TR 20943–5:2013 information technology — procedures for achieving metadata registry content consistency. Accessed on 14 Aug. 2024. [Online]. Available: https://www.iso.org/standard/55026.html
28. ISO: ISO/TS 18876–1:2003 industrial automation systems and integration — integration of industrial data for exchange, access and sharing. Accessed on 14 Aug. 2024. [Online]. Available: https://www.iso.org/standard/33701.html
29. ISO: ISO 19115–1:2014 geographic information — metadata, part 1: fundamentals. Accessed on Aug. 14 2024. [Online]. Available: https://www.iso.org/standard/53798.html
30. ISO: ISO 19115–3:2023 geographic information — metadata, part 3: XML schema implementation for fundamental concepts. Accessed on 20 Aug. 2024. [Online]. Available: https://www.iso.org/standard/80874.html
31. Kokkinaki, A., Moncoiffé, G., Habgood, D., Boldrini, E., and Car, N.: The semantic analyser. In: International conference on marine data and information systems - proceedings volume, S. Simoncelli, M. Vernet, and C. Coatanoan, Eds., Bergen: Miscellanea INGV (2024). https://doi.org/10.13127/MISC/80
32. Kokkinaki, A., Moncoiffé, G., Krijger, T., Boldrini, E., Thijsse, P.: D2.3 - FAIR-EASE semantic brokerage service. Zenodo (2024). https://doi.org/10.5281/zenodo.10606930
33. Gould, M.: ROR and organizational identifier interoperability in publishing systems (2023). https://doi.org/10.14293/s2199-sspam23-01030

34. Meadows, A.: Are you ready to ROR? an inside look at this new organization identifier registry. The scholarly kitchen. Accessed on 20 Aug. 2024. [Online]. Available: https://scholarlykitchen.sspnet.org/2019/12/04/are-you-ready-to-ror-an-inside-look-at-this-new-organization-identifier-registry/
35. Shenoy, A., and Prabhu, A.: Introducing SEO (2016). https://doi.org/10.1007/978-1-4842-1854-9.
36. Roncella, R., et al.: Publishing NextGEOSS data on the GEOSS platform. Big Earth Data **7**(2), 413 (2023). https://doi.org/10.1080/20964471.2022.2135234
37. Roncella, R., et al.: Publishing Eurac research data on the GEOSS platform. Big Earth Data **7**(2), 428 (2023). https://doi.org/10.1080/20964471.2023.2187659
38. Roncella, R., Zhang, L., Boldrini, E., Santoro, M., Mazzetti, P., Nativi, S.: Publishing China satellite data on the GEOSS platform. Big Earth Data **7**(2), 398 (2023). https://doi.org/10.1080/20964471.2022.2107420
39. Zabala, A., et al.: Rubric-Q: adding quality-related elements to the GEOSS clearinghouse datasets. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **6**(3), 1676 (2013). https://doi.org/10.1109/JSTARS.2013.2259580
40. Craglia, M., Hradec, J., Nativi, S., Santoro, M.: Exploring the depths of the global earth observation system of systems. Big Earth Data **1**(1–2), 21 (2017). https://doi.org/10.1080/20964471.2017.1401284
41. Boldrini, E., Nativi, S., Hradec, J., Santoro, M., Mazzetti, P., Craglia, M.: GEOSS platform data content and use. Int. J. Digit Earth **16**(1), 715–740 (2023). https://doi.org/10.1080/17538947.2023.2174193
42. Infrastructure and information systems WMO Commission for observation, WIS metadata key performance indicators. Geneva (2022). Accessed 14 Aug. 2024. [Online]. Available: https://meetings.wmo.int/INFCOM-2/InformationDocuments/INFCOM-2-INF06-3(2)-WIS-METADATA-KPI_en.docx
43. Mazzetti, P., et al.: Knowledge formalization for earth science informed decision-making: the GEOEssential knowledge base. Environ Sci PolicySci Policy (2022). https://doi.org/10.1016/j.envsci.2021.12.023
44. Habermann, T.: Connecting repositories to the global research community: a Re-curation process. J. Esci. Librariansh **12**(3), 12 (2023). https://doi.org/10.7191/jeslib.739
45. Chong, S.S., Schildhauer, M., O'brien, M., Mecum, B., Jones, M.B.: Enhancing the FAIRness of Arctic research data through semantic annotation. Data Sci. J. (2024). https://doi.org/10.5334/dsj-2024-002
46. ISO: ISO 19109:2015 geographic information — rules for application schema. Accessed on 14 Aug. 2024. [Online]. Available: https://www.iso.org/standard/59193.html