algorithms

MDPI

Project Report

Data Management and Data Services in Large Collaborative Projects—DiverSea Experience

Vassil Vassilev [1,*], Georgi Petkov [1], Boris Kraychev [1], Stoyan Haydushki [1], Stoyan Nikolov [1], Viktor Sowinski-Mydlarz [2], Ensiye Kiyamousavi [1], Nikolay Shivarov [1] and Denitsa Stoilova [1]

[1] GATE Institute, Sofia University, 5 James Bouchier Blvd, 1164 Sofia, Bulgaria; georgi.petkov@gate-ai.eu (G.P.); boris.kraychev@gate-ai.eu (B.K.); stoyan.haydushki@gate-ai.eu (S.H.); stoyan.nikolov@gate-ai.eu (S.N.); ensiye.kiyamousavi@gate-ai.eu (E.K.); nikolay.shivarov@gate-ai.eu (N.S.); denitsa.stoilova@gate-ai.eu (D.S.)
[2] School of Computing and Digital Media, London Metropolitan University, 166-220 Holloway Road, London N7 8DB, UK; sowinskiw@londonmet.ac.uk
* Correspondence: vassil.vassilev@gate-ai.eu

Abstract

Collaborative projects under the Horizon Europe Framework Program of the European Union typically involve a large number of partners from multiple countries. Data-centric projects, among them, often require integration of disparate data source formats and collection methods, leading to complex data management architectures and policies. This article is an extended version of an article presented at the 1st International Conference on Big Data Analytics and Applications (BDAA'2025). It explores design decisions, organisational principles, and technological solutions to address these challenges by focusing on data integration of data sources and the hybridisation of data services. This experience was gathered while working on **DiverSea**, a project dedicated to the analysis of biodiversity dynamics along European coastlines—ranging from the Black Sea to the Mediterranean and the North Sea. While grounded in established technologies, the project's takeaways offer valuable insights for environmental data projects across aquatic, terrestrial, and atmospheric domains.

Keywords: data space services; data, metadata and ontologies; semantic search; multimodal chatbots; marine biodiversity

1. Introduction

Many collaborative projects funded under the EU's Horizon Europe include significant information processing, which in turn requires robust data management architectures and policies. This article is an extended version of an article presented at the 1st International Conference on Big Data Analytics and Applications (BDAA'2025) [1]. It presents our experience in this area acquired while working on **DiverSea** [2], one of several major European projects dedicated specifically to monitoring and analysing the state and dynamics of biodiversity and associated ecosystems in coastal seas around Europe (related initiatives include **Marco Bolo** [3] and **OBAMA-NEXT** [4]).

DiverSea is a relatively large project with more than 18 partners from 15 European countries. Its focus is on the analysis of the dynamics of marine biodiversity along the shores of Europe from the North-West shelf of the Black Sea to the Mediterranean from Turkish Waters along the Hellenic Volcanic Ark, and across the Adriatic and Balearic Seas, further along the East Atlantic Coast of Portugal, and from there all the way up to the North Sea. These data come from separate case studies, conducted by teams from different

Academic Editor: Sergey Y. Yurish

Received: 1 January 2026
Revised: 12 February 2026
Accepted: 13 February 2026
Published: 15 February 2026

countries with catchment areas in close proximity to partners' locations. Although it is by no means exhaustive, it is representative enough to establish a methodology for data processing in such projects.

The project relies on the work of both domain expert teams, conducting case studies, and technical teams, working on data management and data analysis tasks. The work is organised in a number of work packages (see Figure 1). Our team is responsible for overall data management with several tasks separated into work package 2 (**WP2**). It bridges the empirical case studies in work package 1 (**WP1)** with the work on data analysis in work package 3 (**WP3)**, which performs data analysis using statistical methods and machine learning, and work package 4 (**WP4)**, which evaluates the impact of the environment on biodiversity dynamics through system dynamics simulation and factor analysis.

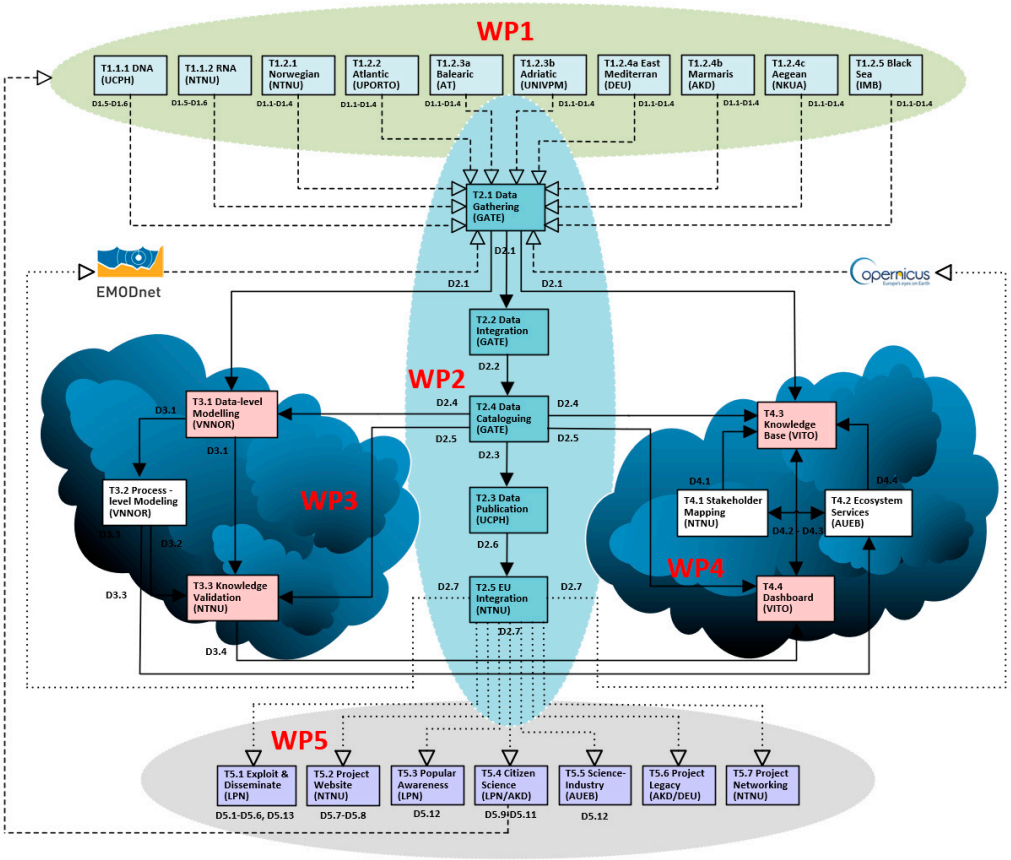


Figure 1. Data management structure of the DiverSea project.

The spatial and temporal distribution of the data in such projects, often characterised by multiple dependencies, requires a complex methodology that combines an elaborate data management strategy with modern data analysis methods. From a software engineering point of view, the DiverSea project relies on both classical software technologies, such as databases and web-based systems, as well as recent developments, such as data spaces, cloud platforms, and services. To this, we can add the potential of Artificial Intelligence (AI) in its multiple recent reincarnations, including ontological modelling and semantic technologies, machine learning, reinforcement learning, transformers, chatbots, and natural language processing. Although our experience with most of them on the project is very positive, we have also encountered some inherent limitations and organisational discrepancies, which have led to choices not always in line with expectations. We believe this experience is valuable not only for marine research and innovation projects such as DiverSea but also for other data-centric environmental projects, which motivated the writing of this report.

2. Research Strategy and Design Considerations

In data-centric projects, data management includes multiple data processing operations that start with the collection of raw data from source systems and end with the interpretation of consolidated, integrated, and fully analysed data. In DiverSea, for example, the focus is on biodiversity dynamics along Europe's coasts, from the Black Sea in the south-east, through the entire Mediterranean coast from east to west, and all the way up to the North Sea in the North-West corner of the continent. The data in such projects typically come from separate case studies, conducted by field teams with catchment areas in proximity to the partners' locations. Due to the high diversity of data sources, formats, and granularity of the data, ranging from eDNA samples to LIDAR scans, as well as the variety of collection modes and communication protocols, it is often tempting to frame data management policies around the concept of Big Data, even if the volume of data is not particularly large. This was exactly the case in the DiverSea project, where the concept of a Data Space [5] was considered the backbone of the data management approach. However, as we will see later, this is not always the best option for a project of this kind for a variety of reasons.

2.1. Distributed vs. Centralized Architectures for Data Management

There are several architectural solutions for Big Data management, which lead to different organisation of the data processing workflows and impose different requirements on the technical staff involved in data processing within each partner organisation.

Centralised Architecture: Based on full centralisation of the analytical operations and rooted in early *data warehousing* [6]. Heavily used in many business projects at the end of the last century, and further elaborated in digital twins architectures [7], with many applications outside business. Light technical requirements are placed on data providers, while heavy expertise is required at the central coordinating site.

Distributed Architecture: Full decentralisation of data processing with an equal role for all data providers. This modern solution often involves complex systems such as data spaces [5] and blockchains [8] and requires significant technical capacity and dedicated resources from each participant.

Hybrid Architecture: Combining elements of the previous two, this approach is typically simpler than a fully distributed architecture and more flexible than a fully centralised one, allowing distribution of data consumers' responsibilities while lowering the technical requirements for data providers.

The solution implemented in DiverSea is a hybrid one. Although there are clear benefits to a more modern, fully distributed architecture, especially for projects at a pan-European level, the leading role in such projects, as a rule, is typically assumed by non-technical organisations. This can shift the focus toward data generation while somewhat neglecting data management and analysis based on recent statistical and machine learning methodologies. In the context of marine biodiversity, for example, it would be of strategic advantage to have a single project that brings together the three complementary initiatives—DiverSea, OBAMA-NEXT, and MARCO-BOLO, and establishes a common Data Space of European Sea and Ocean Biodiversity similar to the Copernicus Data Space Ecosystem (https://www.sentinel-hub.com/explore/copernicus-data-space-ecosystem/ (accessed on 12 February 2026)). This would allow the combining of complementary data and data services, supporting more elaborate data management strategies and more valuable data analytics pipelines.

2.2. Data and Metadata vs. Data Usage and Knowledge

The data-centric projects start with gathering raw data from field studies. The most important artifact from a data management perspective at this stage is the metadata description, which serves as the entry point for developing the data models of the project's data space. Following the common understanding that metadata is information about the data, however, it is often overlooked that metadata is not only about the structure and content of the underlying data but is also a source of additional information that can inform the data space architecture (see Table 1).

Table 1. Some important metadata parameters.

Granularity	Data Sources	Transfer Mode	Transport Protocols
Records	IoT, Computing Devices, Networks	One-off	Memory sharing, Parameter Passing
Messages	Emails, Messages, Logs, Alerts	One-off	MQTT, AMQP, SMS, SOAP
Artefacts	Documents, Sound Recordings, Images, Videos	One-off	FTP, HTTP, MTP, MPC
Datasets	Exports, CSV files	One-off, Batch	FTP, HTTP
Streams	Dynamic URLs, APIs	Continuous	native to the broadcaster
Files	Static URIs, Data Files	One-off	FTP, HTTP, WebDAV, SCP
Repositories	Databases, Data marts, Data lakes	Batch	native to the repository

Depending on the nature of data as specified in the metadata, the data model can differ substantially—it can be purely relational, object-relational, or object-oriented, tree, graph, or file-based. For well-structured datasets, the most appropriate approach is a purely relational model since it is easily represented in standard relational databases, but when the data comes from sensors, drones, or satellites, a more suitable format is an object-relational, object-oriented, or graph-based one. Completely unstructured data, such as images and videos produced through recording, scanning, or photography, is naturally kept in separate files. Finally, in the case of a very large volume of data, popular big data repositories such as Apache Hadoop or Apache Cassandra [9,10] can be used instead. This data can also be complemented by external data from public repositories such as EMODnet (https://emodnet.ec.europa.eu (accessed on 12 February 2026)).

2.3. Single Unified vs. Multiple Specific Data Models for All Case Studies

The choice of a suitable alternative is dictated by the potential need for cross-site analysis. A single unified model is simpler to prepare, and its maintenance can also be simplified by automation. However, it may contain redundant information because metadata varies from case study to case study. On the other hand, multiple specific models may be optimised for each case study, but this requires more development effort and complicates maintenance.

2.4. Physical vs. Logical Model of the Data and Its Use

The primary focus of data management in both the classical data warehousing tradition and more recent data space architectures is the data itself. There are well-established standards for data modelling, which directly translate into the physical implementation of the data repositories. This also supports the subsequent two-stage development of

software systems, such as resource planning systems, decision support systems, operation automation systems, etc. However, the physical design of the repositories does not support the subsequent data analysis, since the physical relationships do not always map directly to logical dependencies. Although the initial efforts to develop data space are typically led by the domain specialists collecting the data, some input from the data analysts can make the metadata richer and provide hints about its potential use. Metadata can feed directly into the physical data model, but it does not translate directly into the domain ontology to provide information about partitioning, taxonomic classification, and dependencies across potential uses of the data. An ontological model of the data space, on the other hand, allows the addition of two important logical layers to the data model: conceptual abstraction, which can easily be modelled using a taxonomic classification of both data and operations, and use cases, which can support additional logical data processing and form a body of knowledge on its own.

3. DiverSea Data Space

Unlike the data spaces in the sense of IDSA [5], the methodology for developing DiverSea data space is based on practical experience rather than on a prescribed roadmap or formal blueprint. The reason is pragmatic and will be discussed later, but nonetheless, the successful outcomes make us confident that it may be helpful in other large projects in which the domain expertise is focused on data gathering, while the technical solution is heavily dependent on resource and data availability.

The development process is illustrated in Figure 2. The process goes through three different phases in a streamlined manner, from a specification of the data to the development of the data space, and finally to its elaboration by adding data services. While in the first phase we have a single task, the second phase progresses incrementally in two parallel streams—physical and logical, so the authors could worked on several tasks independently during the third phase. In this section, we will briefly review the main tasks.

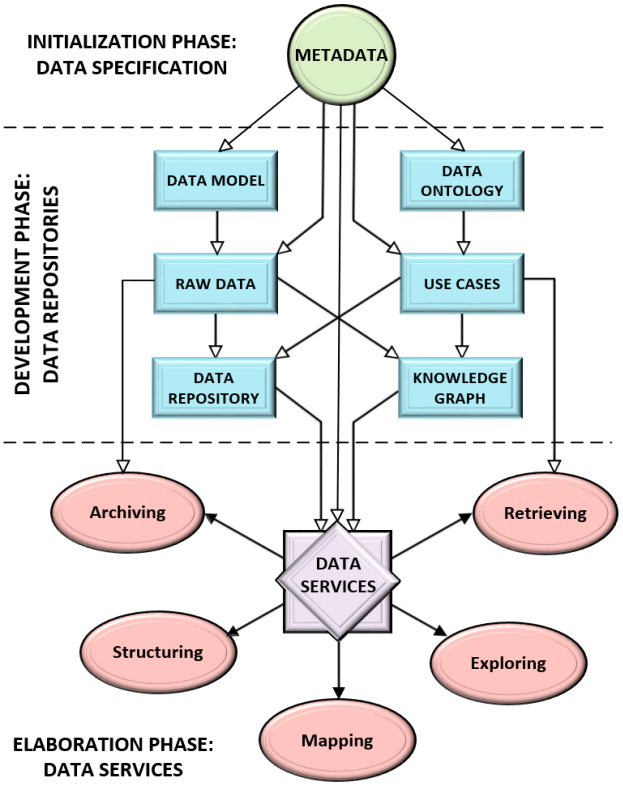


Figure 2. Data space and data services roadmap.

3.1. Gathering Metadata

This is the very first task that initiated the process of developing the data space by collecting metadata from the field case studies. In DiverSea, it was organized by distributing a template to capture both the informational content and the technical parameters of the data generated by each separate case study team.

3.1.1. Information Content

Location: region, area, place, coordinates, etc.

Supplier: partner, ISPs, NGOs, public, industrial, commercial, etc.

Timeline: date, interval, periods, etc.

Content: seagrass, phytoplankton, fish, sea mammals, water, sediments, transport routes, touristic activities, etc.

Origin: sensor data, statistical data, scans, water samples, diagrams, photogrammetry images, satellite images, maps, spatial models, etc.

Related data: from the same source, from a different partner, case study, project, or public repository, etc.

Potential analysis: correlation, recognition of patterns, identification of trends, prediction, impact analysis, etc.

3.1.2. Technical Parameters

Metadata format: text description, SQL DDL, ER model, UML model, JSON model, RDF ontology, etc.

Data Format: text documents, text messages, binary files, CSV files, SQL files, JSON files, XML files, Sound files, Image files, LIDAR files, etc.

Data Source: databases, data files, Web services, programmatic APIs, etc.

Volume: Kilobytes, Megabytes, Gigabytes, Terabytes

Mode of transfer: one-off file upload, external repository import, online streaming, periodic batches, etc.

Access rights: with granted permission, project limited, research purpose only, registered users, public, etc.

Sharing: private, licensed, commercial, public, etc.

Standards: internal, community, professional, national, international, etc.

In addition, the partners were requested to provide data samples to ensure an unambiguous interpretation of the information provided. This information proved invaluable for many of the subsequent tasks. To keep pace with evolving requirements, we updated the metadata midway through the project to reflect the natural evolution and to avoid frequent changes in the data space design, resulting from shifts in field practices.

3.2. Modelling of the Data Space Ontology

Ontological modelling plays an important role in the process of developing data spaces due to their complexity and the necessity to incorporate the profiles of data consumers alongside the logical model of the data. As such, ontologies provide a richer representation than the traditional logical models used by database designers as a step towards creating the physical models. In the DiverSea project, the task was naturally split into two phases. First, the data management team modelled the data ontology in collaboration with the data suppliers, using only the metadata. At a later stage, the data ontology was expanded in collaboration with the data consumers to incorporate information about potential analytical use cases.

The ontology of DiverSea data space was developed using Protégé (https://protege.stanford.edu/ (accessed on 12 February 2025)). Currently, it contains around 600 concepts

described with more than 80 attributes. These concepts are classified into eight taxonomic hierarchies and are related to more than one hundred relations. Figure 3 shows a fragment of the taxonomy of essential ocean and biodiversity variables used to describe the state of biodiversity and associated ecosystems [11].

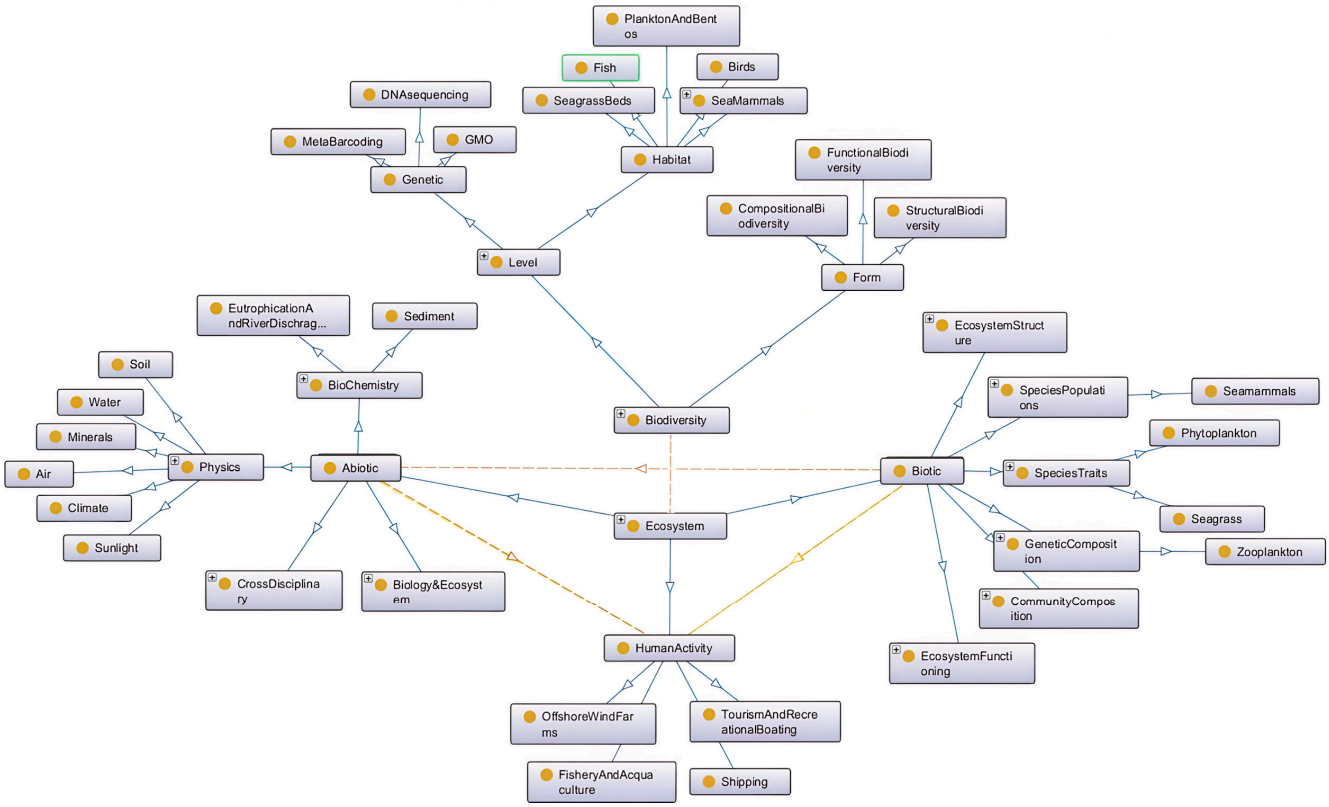


Figure 3. Ontology of the essential variables (EOV, EBV, and Human Activities).

3.3. Developing the Physical Data Model

The physical model of DiverSea data space was developed using metadata in a direct translation. The data management team was prepared to utilise multiple data storage models to accommodate project data relational, sequential, associative, parallel, etc.; after analysing the metadata, however, it quickly realised that all data is purely relational, can be collected entirely offline, and does not require Big Data facilities for processing. This significantly simplified the physical design and subsequent implementation of the data repository.

The model is relatively small due to the strict focus on sea biodiversity and its dynamics. It consists of around 80 tables, structured into groups according to different criteria: by content, format, type, granularity, reference, and standards (see Figure 4). Due to this simplicity, the data management team decided to implement identical data models across all project case studies to enforce unification and support potential cross-location data analysis.

It is also possible to develop the model as a direct translation of the data ontology instead, but since the data is purely relational, this would add unnecessary complexity by requiring the mapping of the taxonomic hierarchies of the ontology into the flat relational structure of the data model. The only adjustment was to use the taxonomic structuring of the data ontology to group the tables on the top level, which improves the mapping between the ontology and the schema without affecting the physical representation.

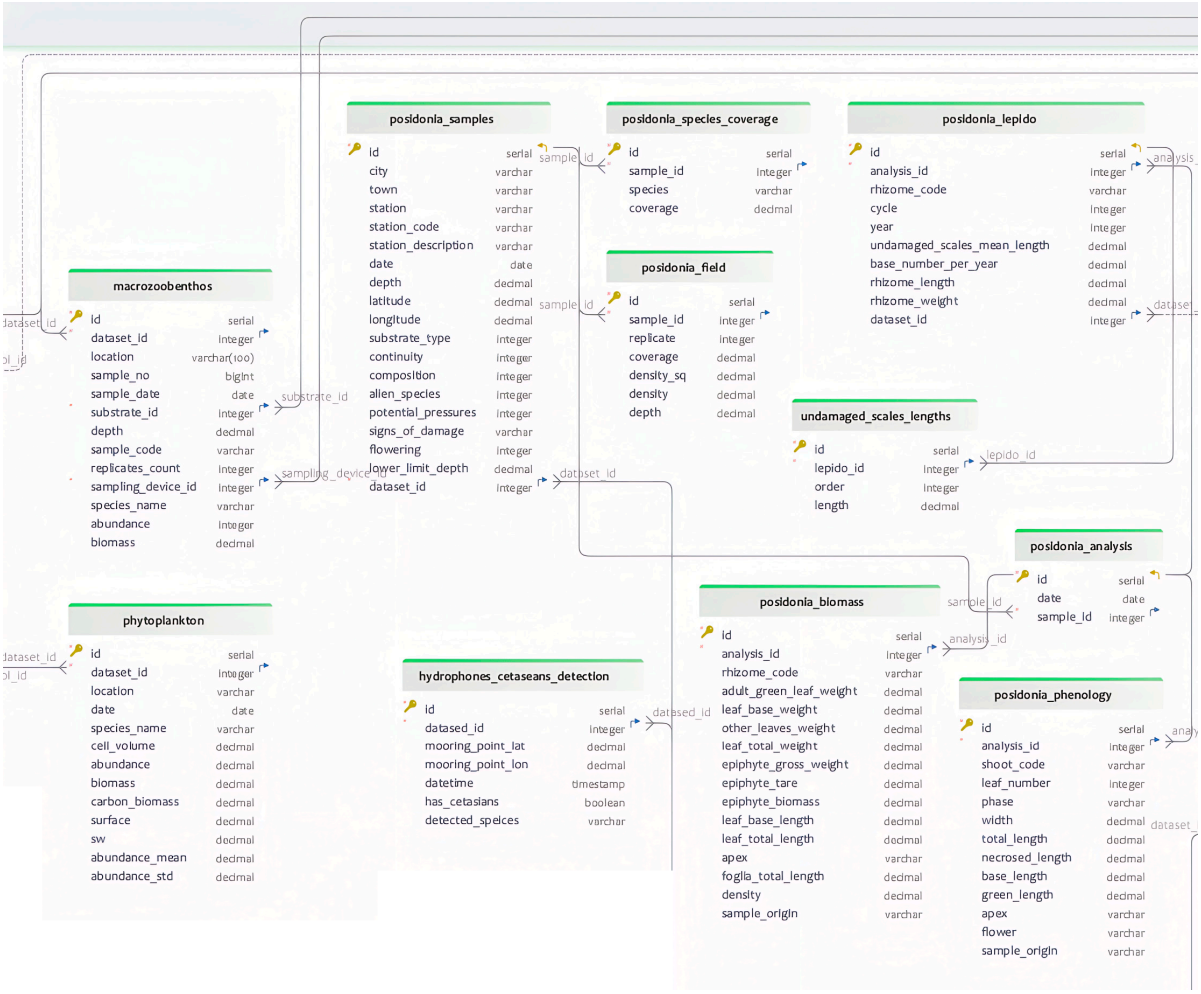


Figure 4. Physical model of biotic data (fragment of the relational model).

3.4. Implementing Data Space Repositories

Since the data model of the DiverSea data space is strictly relational and the volume of expected data is modest, the data management team adopted a streamlined, robust, and cost-effective solution for the implementation of the data repositories. We used PostgreSQL as a repository for storing structured data after preliminary pre-processing, and a web-accessible file system for storing unstructured data in its raw format. Both repositories are maintained centrally and maintain a uniform structure across the case studies, which facilitates maintenance and reduces architectural complexity.

3.5. Validation

While validation is not strictly part of the development cycle, it is more critical than the usual standard quality assessment due to its scale and the high stakes of potential failure. In the case of DiverSea, the initial validation was conducted by both data providers and internal data consumers, while the subsequent validation of the updated data space will involve external data consumers as well.

4. Data Space Services

Data spaces are only as valuable as the services they support. The concept of *service-oriented architecture* [12] inspired the development team to facilitate access to and utilization of data from the DiverSea data space via a set of services. Although these services employ diverse technologies, they provide complementary access for various users and tasks.

Currently, five different data services are maintained, and they cover the entire spectrum of project needs for environmental and biodiversity data: *Raw Data*, *Structured Data*, *Data Source Mapping*, *Chatbot Exploration*, and *Semantic Data Search*. While the first two services use the standard interfaces of the repository for raw data (an Internet drive) and the repository for structured data (a PostgreSQL Database Management System), the other three are bespoke services developed by our team after the Perception-Cognition-Action pattern.

4.1. Raw Data Repository Service

Following a design decision approved by DiverSea partners, the data space incorporates two distinct versions of the data generated by the case studies: original data in its raw format, and curated data structured within a relational database. The raw data is currently stored on a cloud-based file system and is restricted to read-only access without any modification This restriction safeguards the original data from unauthorized or accidental modifications. Datasets were gathered from data providers offline; due to the manageable volume of data, manual ingestion was preferred over automated network transfers. Figure 5 illustrates the structure of raw data collected by partners in the Balearic Islands during the monitoring of sea mammals around the archipelago.

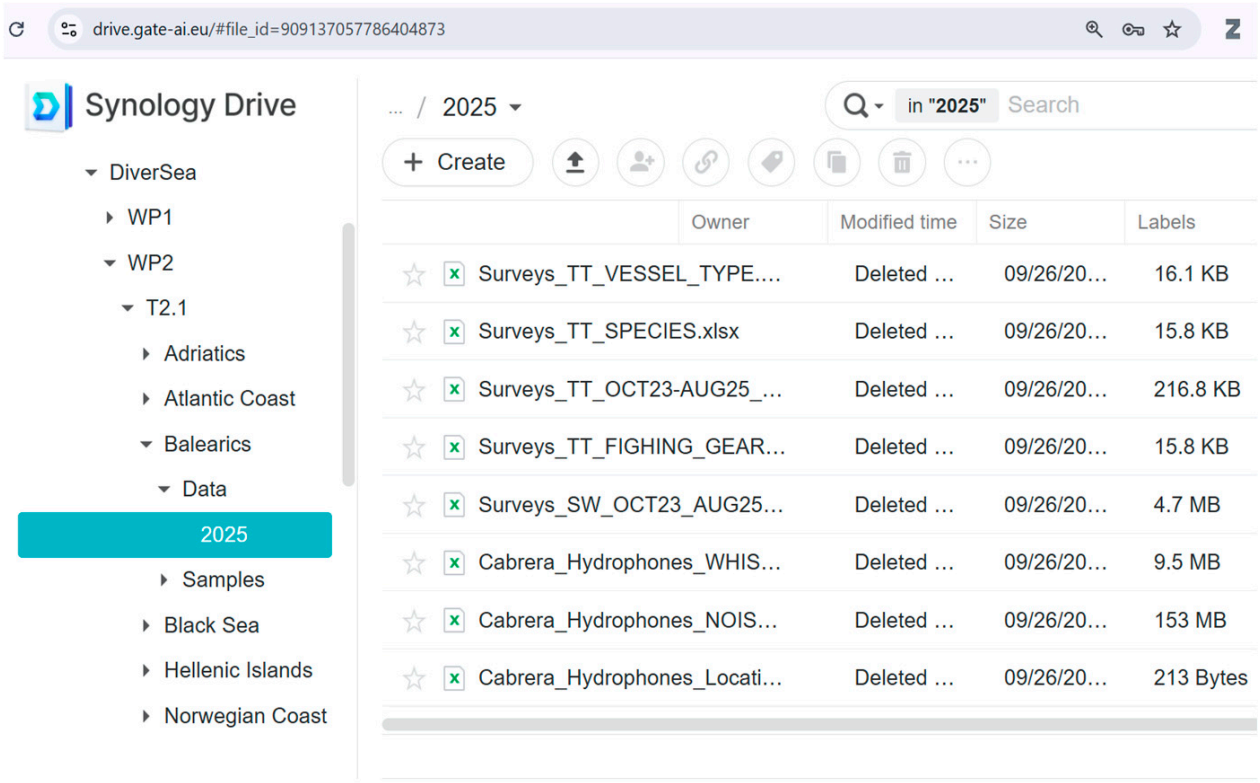
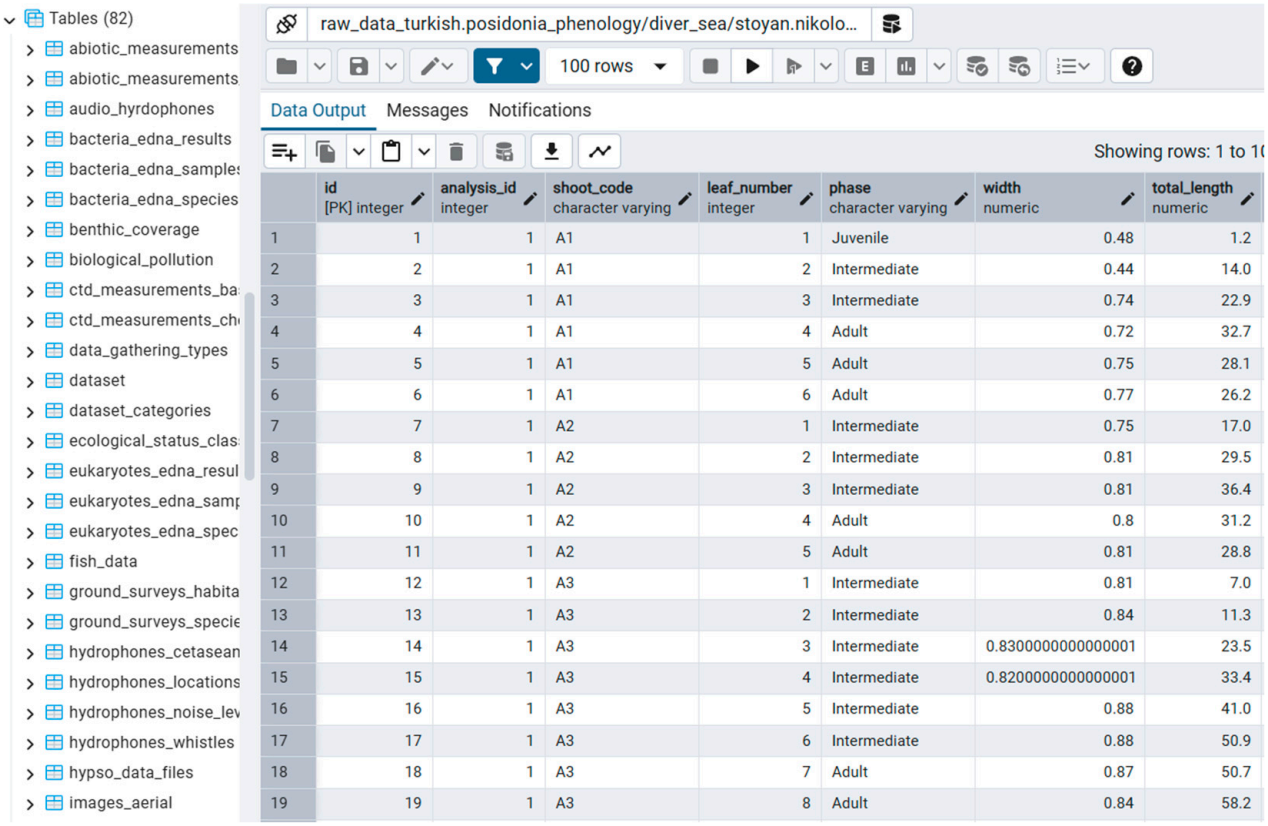


Figure 5. Raw data repository on the internet (Balearic Islands case study).

4.2. Structured Data Storage Service

Before persisting case study data in a structured format, it undergoes curation to meet specific structuring, formatting, and quality requirements. Due to the seasonal acquisition of raw data, the conversion into a relational format is performed periodically and manually, although for larger and more regular data streams, this process could be automated. Figure 6 illustrates the schema and records for Posidonia data from the East Mediterranean coast of Turkey. In the case of structured data, authorised personnel may export the data for local data analysis, simulation, or prediction.



	id [PK] integer	analysis_id integer	shoot_code character varying	leaf_number integer	phase character varying	width numeric	total_length numeric
1	1	1	A1	1	Juvenile	0.48	1.2
2	2	1	A1	2	Intermediate	0.44	14.0
3	3	1	A1	3	Intermediate	0.74	22.9
4	4	1	A1	4	Adult	0.72	32.7
5	5	1	A1	5	Adult	0.75	28.1
6	6	1	A1	6	Adult	0.77	26.2
7	7	1	A2	1	Intermediate	0.75	17.0
8	8	1	A2	2	Intermediate	0.81	29.5
9	9	1	A2	3	Intermediate	0.81	36.4
10	10	1	A2	4	Adult	0.8	31.2
11	11	1	A2	5	Adult	0.81	28.8
12	12	1	A3	1	Intermediate	0.81	7.0
13	13	1	A3	2	Intermediate	0.84	11.3
14	14	1	A3	3	Intermediate	0.8300000000000001	23.5
15	15	1	A3	4	Intermediate	0.8200000000000001	33.4
16	16	1	A3	5	Intermediate	0.88	41.0
17	17	1	A3	6	Intermediate	0.88	50.9
18	18	1	A3	7	Adult	0.87	50.7
19	19	1	A3	8	Adult	0.84	58.2

Figure 6. Structured data repository on the internet (Turkish Waters case study).

4.3. Data Source Mapping Data Service

For both internal and external users, it is advantageous to visualize the available data by location. The most effective method for this is mapping the datasets geospatially; DiverSea Mapping Data Service fulfills this requirement for all datasets that have been processed and stored in the relational repository. Figure 7 illustrates the conceptual scheme of the service, while Figure 8 presents the visual output produced by the service.

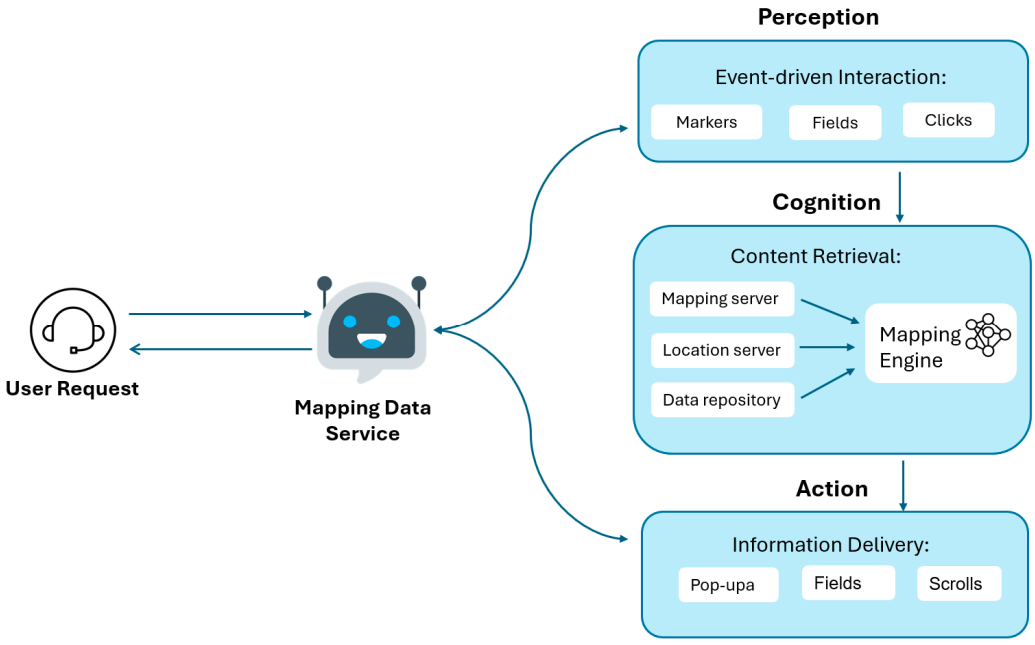


Figure 7. Conceptual scheme of DiverSea mapping data service.

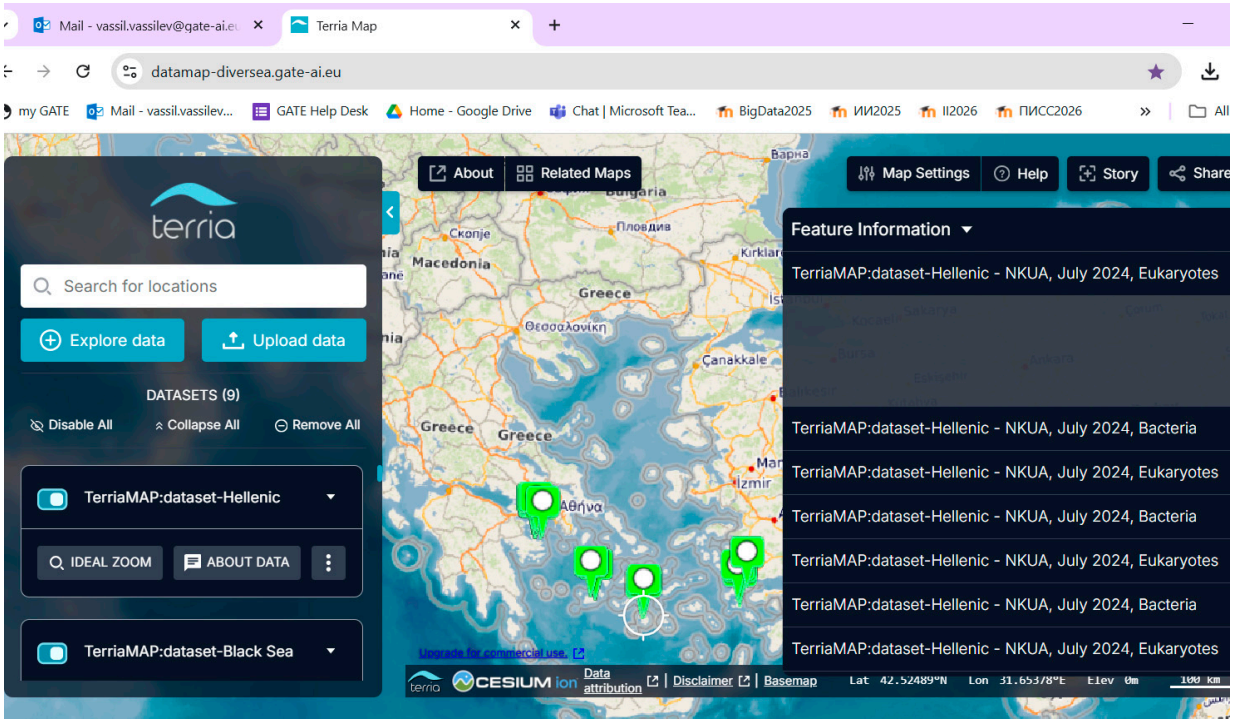


Figure 8. Data source locations on the map (Hellenic Volcanic Ark case study).

The service extracts metadata associated with the datasets and displays it on a dynamic map with global coverage (the green marks point at the data source locations). The cartographic base used for this purpose aligns with the EMODnet visualisation service. Additional technical details regarding the implementation of this service are provided in Appendix A.1.

4.4. Chatbot Exploration Data Service

Recent advances in language technologies, including Natural Language Processing (NLP) and Large Language Models (LLMs), enable users to interact with complex systems using natural language rather than template- or command-driven interfaces. In the context of data spaces, LLM-based assistants can combine deterministic access to curated project resources (via retrieval) with controlled natural-language generation, improving accessibility while preserving traceability to underlying records.

The chatbot exploration service couples' conversational interaction with map-based exploration of the DiverSea data space. Users submit requests as text or speech. Spoken input is transcribed via Speech-to-Text, and the current map state (e.g., viewport, selected area, or clicked coordinates) is captured from the Cesium viewer. The service then interprets the request to determine the target entity type (e.g., institution, event, meeting, activity, or location) and formulates a retrieval query over the project exploration index. The top-ranked records are used as grounding context for the LLM (Qwen3-32B served via vLLM), which generates (i) a concise response and (ii) a structured set of geospatial actions (e.g., pan/zoom, highlight features, open pop-ups) executed by the front-end in Cesium. The final output is presented as text and, optionally, synthesised via Text-to-Speech to support multilingual interaction. Implementation details and supported interaction operations are summarised in Appendix A.2.

Figure 9 presents the conceptual architecture of the chatbot data service. Figure 10 illustrates chatbot-assisted exploration of a project case-study site on the East Atlantic coast of Portugal, demonstrating how conversational queries can trigger map navigation and

context-based access to project data. Additional technical specifications are provided in Appendix A.2.

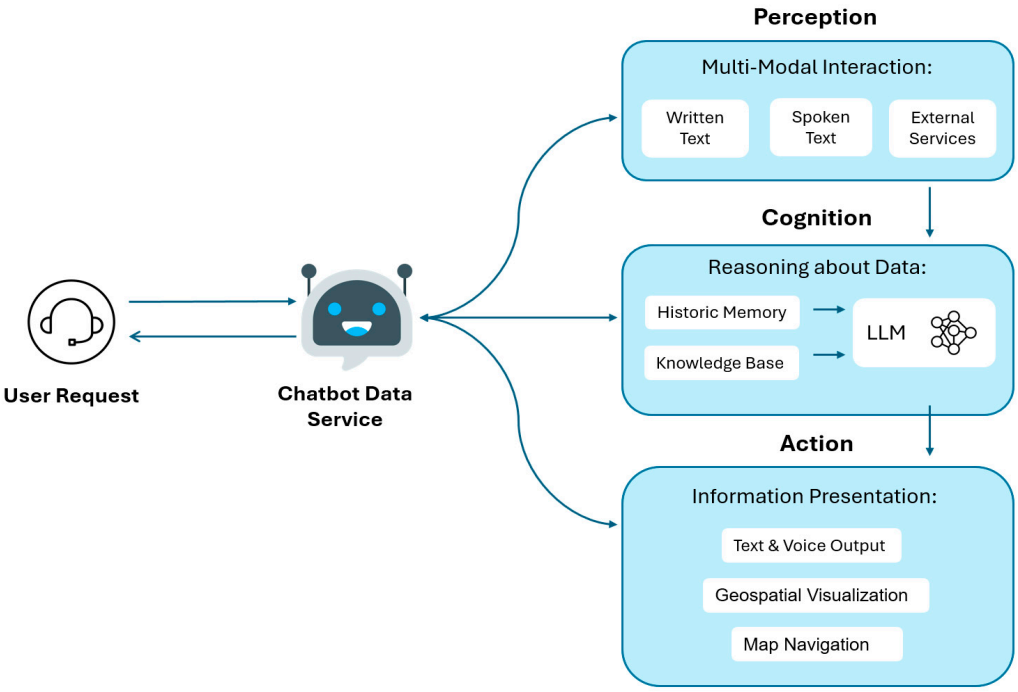


Figure 9. Conceptual architecture of the DiverSea chatbot data service.

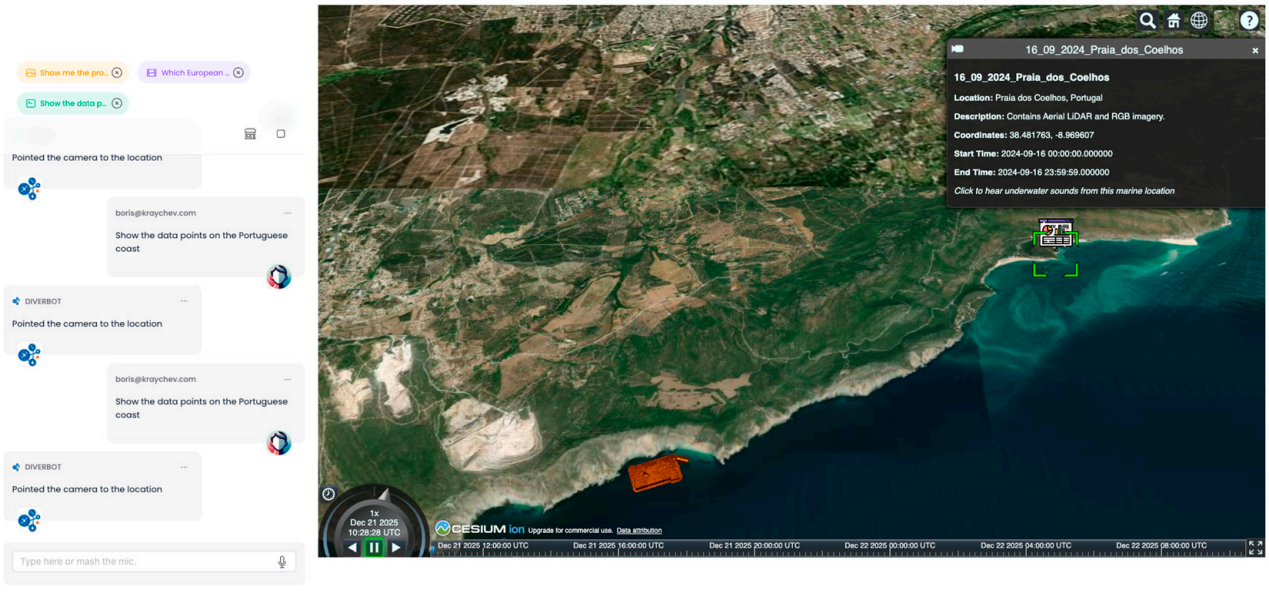


Figure 10. Chatbot–assisted exploration of the data space (East Atlantic Coast case study).

4.5. Semantic Search Data Service

Metadata available within the data space acts as a natural starting point for identifying appropriate data for analysis, simulation, prediction, and other possible uses. However, the integration of multiple data sources across various case studies complicates this retrieval process. Furthermore, cross-area data analysis requires consistent querying across data housed within different repositories. The difficulty arises from the fact that the classical syntactic search methods rely on exact string matching or lexical similarity; consequently, they need to define search parameters in advance.

To address these limitations, the semantic search service employs an expanded ontology that integrates not only the logical data model but also the model of potential use cases for data analysis, simulation, and prediction. Figure 11 illustrates the conceptual diagram of the service, while Figure 12 demonstrates its use in the use case for identifying correlated datasets within the environmental and biodiversity repositories for the North-West shelf of the Black Sea.

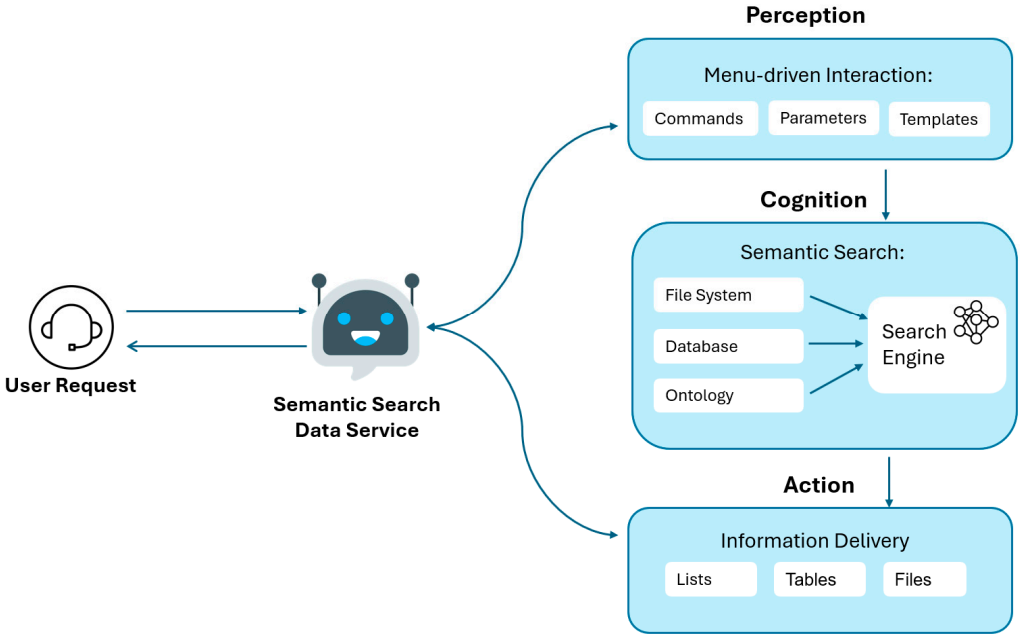


Figure 11. Conceptual architecture of the DiverSea semantic search data service.

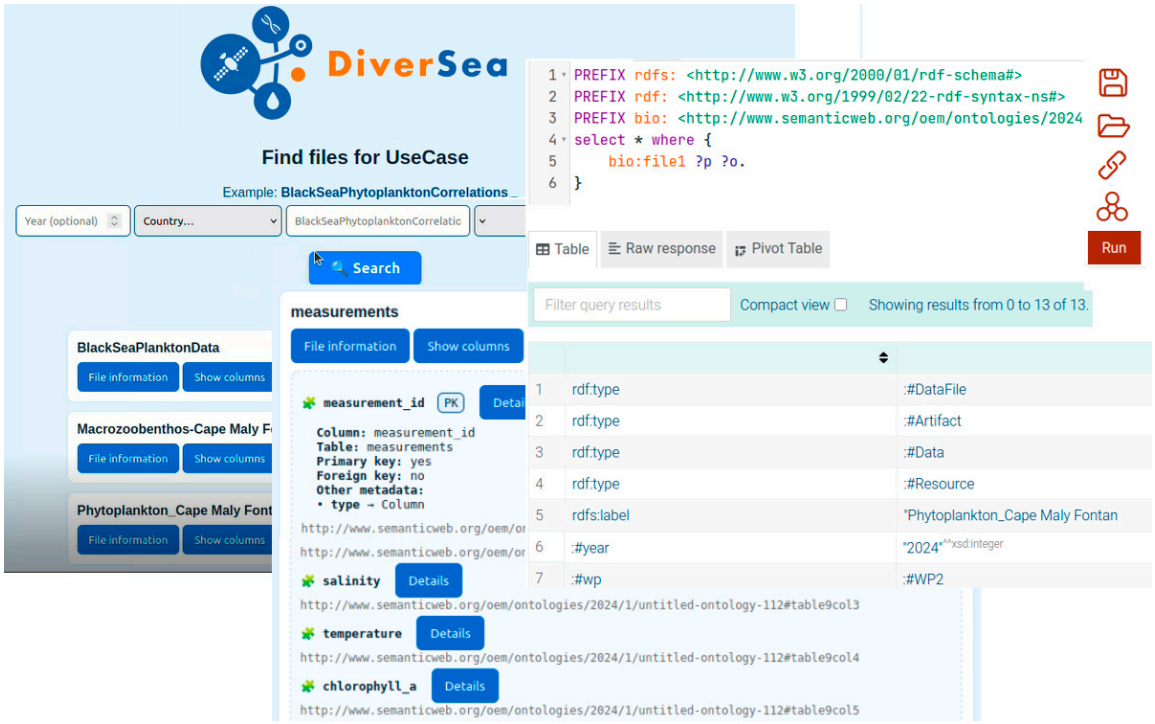


Figure 12. Semantic search service interface and query (North-West Black Sea case study).

The top-left screen in Figure 12 shows the main interface of the data service. Through this interface, users can search for data files and tables using multiple parameters such as year, geographic location, and associated use case. The interface supports advanced

operations over datasets, including set-based operations (intersection, union, and difference) applied to the collections of files associated with two selected use cases. This enables users to compare and explore overlaps and distinctions between datasets originating from different scientific scenarios. The top-right figure presents the underlying ontology stored in GraphDB. The ontology models data files, tables, parameters, and their semantic relationships, forming the backbone of the semantic search functionality. The service executes a SPARQL query against the ontology to retrieve relevant information, infers logically the relations which exist within the data, and enables navigation across the linked physical data repository for retrieving the information using the query language of the physical repositories, in this case SQL for the structured data and the commands of the directory services of the underlying operational system for raw data. The bottom figure illustrates the result of a completed search.

After executing the query, the user receives detailed information about the retrieved files/tables, including column-level metadata. For each column, semantic annotations and links to related columns are displayed. When a column is semantically connected to other columns, the interface provides direct links that allow users to open and explore these related columns in new tabs, effectively enabling graph-based navigation without requiring prior knowledge of graph databases. This way, the service implements true semantic search beyond the syntactic string-matching mechanisms, which allows for incorporating abstract classification and conceptual dependencies among different parameters of the query. We are also working on physical downloads of the found data. Technical specification regarding the implementation of the semantic search service is provided in Appendix A.3.

5. Conclusions and Future Work

This article presents the methodology and the progress achieved by the DiverSea project data management team in applying it for developing the project data space. It allows integration of data from multiple data providers to support multiple consumers for the purpose of monitoring, analysis, and prediction of the dynamics of biodiversity in European marine ecosystems, spanning from the Black Sea in the South-East through the entire Mediterranean and by the Eastern Atlantic all the way to the North Sea. The team implemented a centralised system for handling all data gathered from the case studies and established data services enabling different stakeholders to utilise the data for varied purposes—analysis, simulation, and prediction. The project continues for another two years, but the solution is already operational and has been successfully validated by both data providers and data consumers. During the project's remaining tenure, the team will continue field data acquisition while systematically preparing curated datasets for publication in the large European repository of marine data, EMODnet.

During the project, the data management team faced various challenges related to data content as well as the project governance, which necessitated the refinement of some originally proposed solutions. The adoption of a robust and flexible architecture proved critical, allowing the data space to evolve into its current form, which effectively addresses the project's requirements. At the same time, the expectations for implementing more advanced solutions, such as a proper data space in the sense of IDSA, were not met. Nevertheless, the data management team believes the methodology illustrated here provides a valuable framework for similar large-scale collaborative data initiatives.

6. Lessons and Recommendations

The lessons derived from the data management aspect of the DiverSea project centre on early design stage/architectural decisions, which are often interrelated, mutually dependent, and structurally complex:

Distributed vs. Centralized Data Architectures. While distributed architectures offer greater autonomy to participants, they impose significant overhead regarding communication and synchronisation. Conversely, while centralised architectures are often perceived as restrictive, they frequently facilitate tighter integration across operational levels and more efficient resource utilization. The optimal solution lies in providing enhanced technical support for each participating data provider and data consumer while introducing a kind of central management authority.

Classical vs. Modern Data Technologies. Legacy data technologies, such as backend databases and frontend interfaces, provide a robust foundation for deploying modern solutions. Conversely, knowledge graphs, machine learning, and multi-modal interactions contribute significant value to data utility. This necessitates a hybrid approach that pairs reliable, high-fidelity data provisioning with sophisticated, albeit evolving, analytical services.

Big vs. Small Project Consorts. Small projects operate within constrained budgets, data volumes, and timeframes, which may constrain their scalability. However, large-scale projects introduce significant overhead in terms of coordination, personnel engagement, and administrative management. Consequently, enhancing the funding of smaller, agile teams may often yield greater efficiency than allocating massive resources to oversized consortia.

Substantial vs. Limited Project Resources. Large-scale projects and extensive consortia require extended timelines and greater investments. This inherent inertia can be mitigated better if projects are phased to allow for incremental progress. In the context of data-centric projects, this implies that certain work streams can be executed in parallel instead of sequentially, thereby improving synchronization through external milestones and reinforcements.

The predominant practice in the development of joint European projects, unfortunately, complicates the selection of optimal architectural and organizational alternatives. Designing a well-targeted, balanced, and strategically aligned project is inherently challenging; furthermore, the overhead of deliberation, synchronisation, and multi-team coordination often renders successful execution prohibitively difficult. Mirroring the methodological revolution in software engineering sparked by Agile principles, the management of large-scale collaborative data spaces necessitates a shift toward greater flexibility and adaptive governance.

Author Contributions: Conceptualisation, methodology, and design, V.V., B.K. and G.P.; software development, B.K., E.K., V.S.-M. and N.S.; data curation and data management, S.H. and S.N.; original draft preparation, V.V.; review and editing, G.P., B.K., V.S.-M. and D.S.; visualisation, V.V., V.S.-M. and B.K.; project administration, V.V. and D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Research Executive Agency (grant number 101082004) under HORIZON Research and Innovation Actions programme, supporting the project "Integrated Observation, Mapping, Monitoring and Prediction for Functional Biodiversity of Coastal Seas" (DiverSea), (Call: HORIZON-CL6-2022-BIODIV-01).

Data Availability Statement: In accordance with the open data policy adopted by the funding organisation, the data produced within the DiverSea project is being released incrementally for public use. Currently, it is available in several formats via the DiverSea data services, which are maintained by the GATE Institute at Sofia University. The spatial distribution of the data can be explored at https://datamap-diversea.gate-ai.eu/ (accessed 30 December 2025), while the curated datasets are accessible at https://datasets-diversea.gate-ai.eu/ (accessed 30 December 2025). For internal consortium use, data in its raw format is stored at https://datafiles-diversea.gate-ai.eu/ (accessed 30 December 2025). Upon conclusion of the project, high-value datasets will be integrated into the European Marine Observation and Data Network (EMODnet) repository.

Acknowledgments: The GATE Institute team wishes to express its gratitude to all colleagues within the partner institutions of the project "Integrated Observation, Mapping, Monitoring and Prediction for Functional Biodiversity of Coastal Seas" (DiverSea) for their support of our work on data management. Special thanks are extended to the Project Coordinator, Murat Ardelan, and the work package leaders—Renato Mendes, Jean-Luc de Kok, and Bob O'Hara—whose encouragement, strategic advice, and ongoing support were instrumental in shaping the data management methodology and validation policy. Furthermore, we thank the case study leaders for their continuous cooperation in providing field data and their vital role in the iterative validation of the DiverSea data space.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

EOV	Essential Ocean Variables, a set of physical measurements for assessing the state and variability of oceans and seas.
EBV	Essential Biodiversity Variables, a set of standardized biological measurements for assessing the changes in biodiversity across time, space, and biological level.
eDNA	Environmental DNA obtained from the environment by sampling, typically from the water and air.
LIDAR	LIght Detection And Ranging, a remote-sensing technology that uses laser beams to measure precise distances and movement in an environment in real time.
IDSA	International Data Space Association advocates a distributed model for data spaces, maintains standards, and provides support for implementing such concepts.
AI	Artificial Intelligence, as a branch of Computer Science, has been under intensive development during the last ten years in both anthropomorphic and non-anthropomorphic settings.
NLP	Natural Language Processing is an AI direction that deals with the automation of processing of both spoken and written languages.
ML	Machine Learning is an AI direction that provides the methodology, methods, and tools for intelligent data analysis.
LLM	Large Language Model, a generative transformer trained with machine learning algorithms on texts, which can be fine-tuned for specific tasks.
SQL	Structured Query Language for popular databases, such as PostgreSQL. Combines SQL DDL (Data Definition Language) and SQL DML (Data Manipulation Language).
SPARQL	Semantic Protocol and Resource Query Language, a standard semantic technology providing access to many graph databases.

Appendix A

Appendix A.1. Data Sources Mapping Service

This appendix provides additional technical details regarding the implementation of the data sources mapping service, which visualizes the locations of DiverSea case study data sources via a real-time, live geospatial interface. The architectural foundation of the service integrates three prominent geolocation and visualization platforms: GeoServer, Terria, and Cesium.

Table A1. Mapping service protocol.

Operation	Parameters	Description	Tools
Spatial navigation	Data source, Geolocation Coordinates	Obtaining geographic location coordinates for the chosen data source by clicking the map	JavaScript, Terria, Cesium,

Table A1. *Cont.*

Operation	Parameters	Description	Tools
Request formulation	Database, Dataset	Query formulation, connecting to the server, accessing the repository, and executing a sequence of SQL commands	GeoServer, PostGIS, PostgreSQL
Information retrieval	Case study, Time, Institution, Content, Statistics, etc.	Retrieving the metadata using SQL DDL and dataset statistics using SQL DML, and combining it with the location information	PostgreSQL, PostGIS, GeoServer
Information presentation	Current information as retrieved at the moment of requesting	Presenting the retrieved information on the map next to the data source location in a separate pop-up window	3D Geospatial Viewer, Cesium, Terria, JavaScript

Appendix A.2. Chatbot Exploration Service

This appendix provides additional technical details regarding the implementation of the chatbot service, which facilitates user-friendly data exploration by providing context information on relevant events, organizations, and locations. The service supports both spoken and written interaction in multiple languages. The nexus of the service is a Qwen3-32B Large Language Model (LLM) integrated with Natural Language Processing (NLP) frameworks for Speech-to-Text (STT) and Text-to-Speech (TTS) conversion.

Table A2. Chatbot communication protocol.

Operation	Parameters	Description	Tools
Request formulation	Project context, Search space, Natural language	Map navigation, Spoken and written text in natural language, Multi-lingual support	React, Whisper, Cesium
Spatial navigation	Geolocation coordinates	Obtaining the geographic location coordinates from the map	Cesium
Information retrieval	Institutions, Events, Meetings, Activities, Locations, Histories	Keyword-driven query analysis, Search space contextualization, Query formulation	Qdrant, vLLM
Information presentation	Geospatial actions, Spoken text, Written text	The information is represented visually on the map, accompanied by a written textual description	React, Kokoro, Cesium

Appendix A.3. Semantic Search Service

This appendix provides additional technical implementation details regarding the semantic search service, which enables logic-based querying of the data space by leveraging a dedicated ontological model encompassing both data relationships and the analytical use cases of the project.

The engine of the service is the GraphDB triplestore, which manages the project's ontology and semantic relationships.

Table A3. Semantic search protocol.

Operation	Parameters	Description	Tools
Query specification	Identification, Grouping, Splitting, Classification, Relation	Interactive formation of the search query using drop-down menu options and search strings using only a standard Web browser	JavaScript
Query generation	Language, Destination	Automatic parsing and translation of the queries with parameters into the query language of the destination (in this case, SPARQL)	Python 3
Query execution	as in the original formulation	Search for raw data files, database tables, and columns by traversing the knowledge graph	GraphDB, PostgreSQL
Information presentation	as retrieved	List the findings, described using location, name, type, place, time, etc.	JavaScript
Data download	as linked	Exports the data from the data space repository, following the links to it in the ontology	SQL DML

References

1. Vassilev, V.; Petkov, G.; Kraychev, B.; Haydushki, S.; Sowinski-Mydlarz, V.; Nikolov, S.; Shivarov, N. Above, On, and Bellow the Surface: Data Services in Large Collaborative Projects. In *Proceedings of the 1st International Conference on Big Data Analytics and Applications (BDAA'2025)*; IFSA: Indianapolis, IN, USA, 2025; pp. 63–69.
2. Integrated Observation, Mapping, Monitoring and Prediction for Functional Biodiversity of Coastal Seas. Available online: https://www.ntnu.edu/diversea (accessed on 9 August 2025).
3. Marco-Bolo: MARine COastal BiOdiversity Long-Term Observations: Strengthening Biodiversity Observation in Support of Decision Making. Available online: https://marcobolo-project.eu/ (accessed on 9 August 2025).
4. OBAMA-NEXT: Observing and Mapping Marine Ecosystems—Next Generation Tools. Available online: https://obama-next.eu/ (accessed on 9 August 2025).
5. International Data Spaces Association. Reference Architecture Model. Available online: https://internationaldataspaces.org/offers/reference-architecture/ (accessed on 12 February 2026).
6. Kimball, R.; Ross, M. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modelling*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2013.
7. Khan, R. *Digital Twin & Digital Development's Handbook*; Independently Published, 2023; ISBN 979-8397986014.
8. Bashir, I. *Mastering Blockchain: Inner Workings of Blockchain, from Cryptography and Decentralised Identities*, 4th ed.; Packt: Mumbai, India, 2023.
9. Anjomshoaa, A.; Caceres, S.; Wolff, C.; Baún, J.C.P.; Karvounis, M.; Mellia, M.; Athanasiou, S.; Katsifodimos, A.; Garatzogianni, A.; Trügler, A.; et al. Data Platforms for Data Spaces. In *Data Spaces. Design, Deployment and Future Directions*; Curry, E., Scerri, S., Tuikka, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; pp. 43–64.
10. Vassilev, V.; Sowinski-Mydlarz, V.; Gasiorowski, P.; Radu, S.; Nakarmi, S.; Hristev, M.; Baghaeishiva, R.; Bali, T. Building A Big Data Platform using Software without License Costs. In *Open Source Horizons-Challenges and Opportunities for Collaboration and Innovation*; Castro, L., Ed.; Intechopen: London, UK, 2024; pp. 29–52.
11. Muller-Karger, F.; Miloslavich, P.; Bax, N.; Simmons, S.; Costello, M.J.; Pinto, I.S.; Canonico, G.; Turner, W.; Gill, M.; Montes, E.; et al. Advancing Marine Biological Observations and Data Requirements of the Complementary Essential Ocean Variables (EOVs) and Essential Biodiversity Variables (EBVs) Frameworks. *Front. Mar. Sci.* **2018**, *8*, 373058. [CrossRef]
12. The Open Group. SOA Source Book. Available online: https://collaboration.opengroup.org/projects/soa-book/ (accessed on 11 December 2005).