D2.4. DIVERSEA DATA SPACE (INITIAL)

Funded by the European Union

DiverSea - HORIZON-CL6-2022-BIODIV-01-01
Grant Agreement 101082004

PROJECT	
Project Number	101082004
Project Acronym	DiverSea
Project Name	Integrated Observation, Mapping Monitoring and Prediction for Functional BioDiversity of Coastal SEAs
Call	HORIZON-CL6-2022-BIODIV-01
Topic	HORIZON-CL6-2022-BIODIV-01-01
Type of action	HORIZON-RIA
Service	REA/B/03
Project starting date	01 Sep 2023
Project duration	48 months

DELIVERABLE INFORMATION	
Number:	NAME: D2.4 DiverSEA Data Space (Initial)
Responsible Partners	**GATE-SU**
Contributors	Vassil Vassilev, GATE Stoyan Haydushki, GATE Boris Kraychev, GATE Georgi Petkov, GATE Emil Hristov, GATE Denitsa Stoilova, GATE Renato Mendes, UPORTO
Type	DATA — data sets, microdata, etc
File	PDF
Revision	29th December 2024
Due Date	31/12/2024
Submission Date	29/12/2024
Dissemination	PU - Public

DOCUMENT HISTORY		
Revision Date	Modification	Author
14.11.2024	v.0.1 Deliverable structure	Denitsa Stoilova
21.11.2024	v.0.2 Review of the structure by projects partners	Renato Mendes
21.11.2024	v.0.3 Contributions from WP1 partners to Annex 5	

28.11.2024	v. 0.4 Contributions to Chapters 1, 2, 3, 4, 7, Annex 1 and Annex 2	Vassil Vassilev
07.12.2024	v.0.5 Contribution to Chapter 5 and Annex 3	Stoyan Haydushki
07.12.2024	v.0.6 Contributions to Chapter 6	Emil Hristov, Denitsa Stoilova
08.12.2024	v.0.7 Contributions to Chapter 1 and Annex 1	Boris Kraychev
08.12.2024	v.0.8 Contributions to Chapter 1 and Annex 1	Georgi Petkov
18.12.2024	v.0.9 Integration of various chapters, sections and graphs	Vassil Vassilev
20.12.2024	v.1.0 Final review	Denitsa Stoilova

CLASSIFICATION	
Draft/Final	Final
Confidential/Restricted / Public - Link	Public

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This document reports on Deliverable 2.4 „*DiverSEA Data Space (initial*)“, part of Work Package WP2 “*DiverSEA Data Space: Data integration, harmonization, cataloguing and compatibility with standards” of project “DiverSea – Integrated Observation, Mapping, Monitoring and Prediction for Functional BioDiversity of Coastal SEAs*“.

DiverSEA Data Space is a distributed resource developed to support the project’s primary goal of advancing new methods for monitoring biodiversity and analyzing the dynamics of the coastal ecosystems along the shores of Europe. This effort leverages cutting-edge computer-based technologies and integrates applications directly with field ecosystem studies, aligning with professional, European and global standards in the domain. The key objectives of the initial version of the data space are:

1. to establish a common understanding of data processing methodologies among project partners,
2. to develop a unified conceptual model for integrating data from diverse field case studies conducted in WP1, and
3. to provide the technical foundation for analyzing data through statistical methods, data analytics and impact analysis in WP3 and WP4 by developing a unified database for storing raw data collected in the WP1 case studies.

The major milestones achieved in this deliverable are the development of a common vocabulary of terms, creation of an ontology of the data space, design of a data model for the databases for storing raw data and the establishment of a file catalogue for sharing files within the project.

DiverSea Data Space serves as a central resource in the project, playing a crucial role in connecting work packages WP1, WP3, WP4 and WP5: WP1 acts as a data provider, contributing field study data, WP3 and WP4 rely on the data space for analytics, utilizing the collected, stored and catalogued data, and WP5 facilitates communication between the work packages following the flow of data through the project.

In the next version of the data space, its scope will be extended to include restructured and further processed data, making it ready for sharing with both internal and external stakeholders, tailored to their specific objectives. From this perspective, the initial version of DiverSea Data Space represents an intermediate step toward empowering researchers, service providers, business operators and decision-makers at local, national and international scale. It provides an invaluable resource for informed decision-making and competent action to promote the prosperity and health of EU citizens.

1. Introduction

1.1. Overview of Work Package 2 and its Place in DiverSea Project

Work Package 2 (WP2) plays an integral role in the DiverSea project, acting as a source, destination, and common ground for activities across other project work packages. WP2 focuses on managing the data generated, processed, and utilized within different tasks of these work packages (see Fig. 1). The primary data sources for WP2 are the regional case studies conducted by project partners in various European seas - the Black Sea, the Sea of Marmara, the East Mediterranean, the Aegean Sea, the Adriatic Sea, the Ballearic Sea, the East Atlantic Coast of Portugal, the Baltic Sea and the North Sea. Together with the other two technical work packages (WP3 and WP4), WP2 forms the technological backbone supporting project activities across these regions. Additionally, WP2 serves as an external interface to the common European maritime resources in both incoming and outgoing directions by publishing data created within DiverSea and loading public data from repositories such as EmodNet and Copernicus.

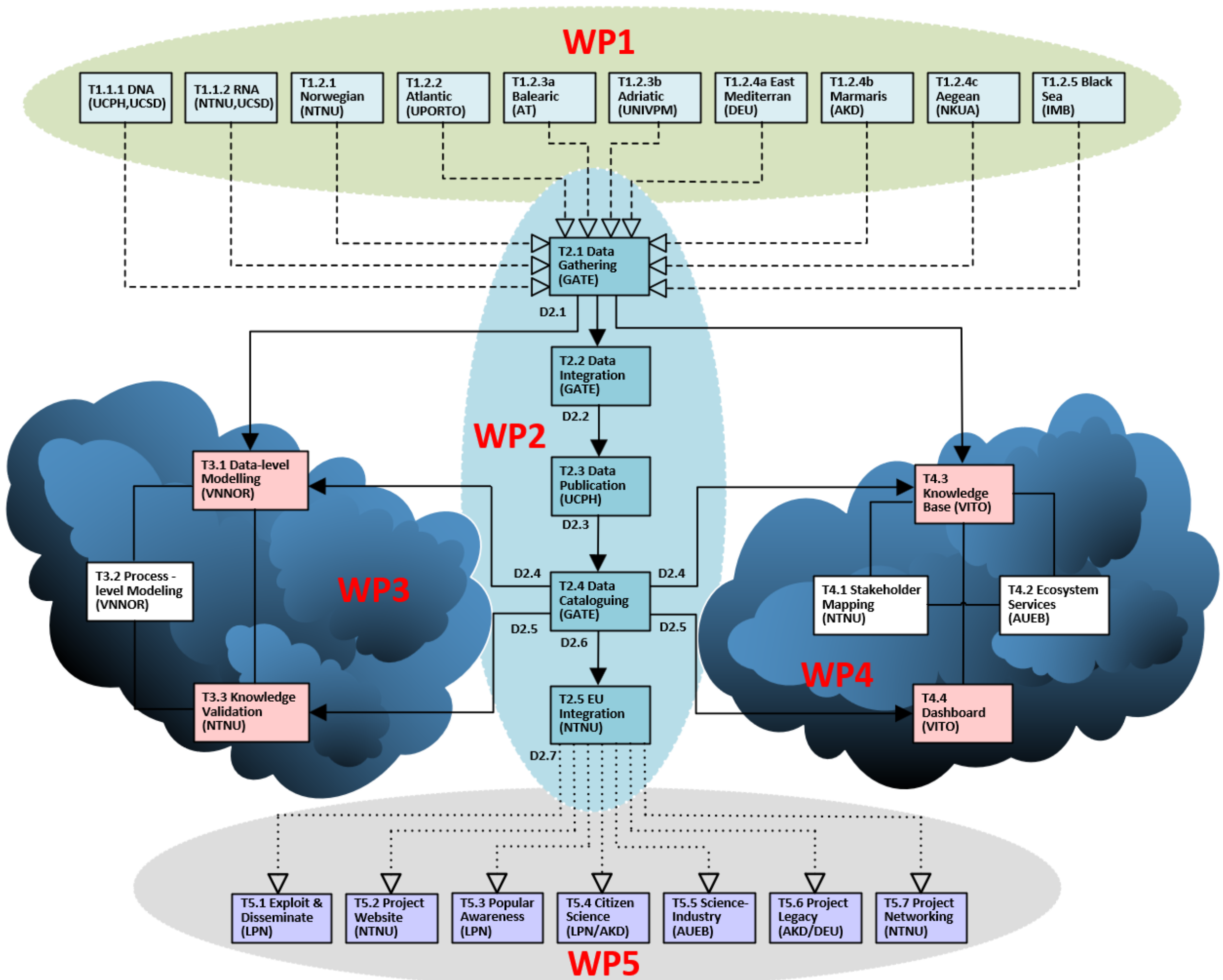


Figure 1. Project Work Packages and Dependencies

1.2. The concept of Data Space and its role in monitoring the sea biodiversity

The technical concept of a *Data Space*[1] is a successor of the older concept of a *Data Warehouse*[2], widely used in industry to refer to the technical organization of distributed data for centralized analysis. An essential difference between the two concepts is that while the data warehouse assumes a physical collection of the distributed data in one centralized place for its analysis, the data space is a purely logical concept and does not require the data to be collected, while maintaining transparency of its distributed and heterogeneous nature. The distinctions can be summarized as follows:

- different technical infrastructures required (data warehouse database vs. data space file system);
- different data models implemented in them (data warehouse data scheme vs. data space ontology);
- different modes of data processing supported (centralized vs. decentralized).

Although both concepts have their own merits and can create a basis for the analysis of distributed data, neither is appropriate for DiverSea for several reasons:

Data Warehouse: Easier to build, but best suited for more comparable and structured data. This technology is widely used in corporate businesses as it is well-suited for business intelligence. **DiverSea** data across regions is highly diverse (ranging from dolphin sound recorded by underwater drone sensors to phytoplankton areas photographed by satellites, to eDNA sequencing of water samples).

Data Space: Provides more opportunities for data integration and data analysis of heterogeneous data. This technology is being increasingly adopted in B2B contexts because it is more suitable for managing complex interconnected supply chains. Unfortunately, it requires significant resources to develop and maintain centrally, and also additional resources to implement locally. DiverSea could serve as a follow-up to several European maritime projects, such as OBAMA-NEXT and Marco-Bolo.

Since WP2 is responsible for preparing data for analysis using statistical and machine learning (ML) methods in WP3 and for factor analysis using system dynamics methods in WP4 at start of the project we explored opportunities to build suitable data space to support them. In consultation with the partners during the first and the second annual project meetings, we agreed to develop an architecture, which is more federated than distributed, with data that is physically unified rather than logically integrated.

---

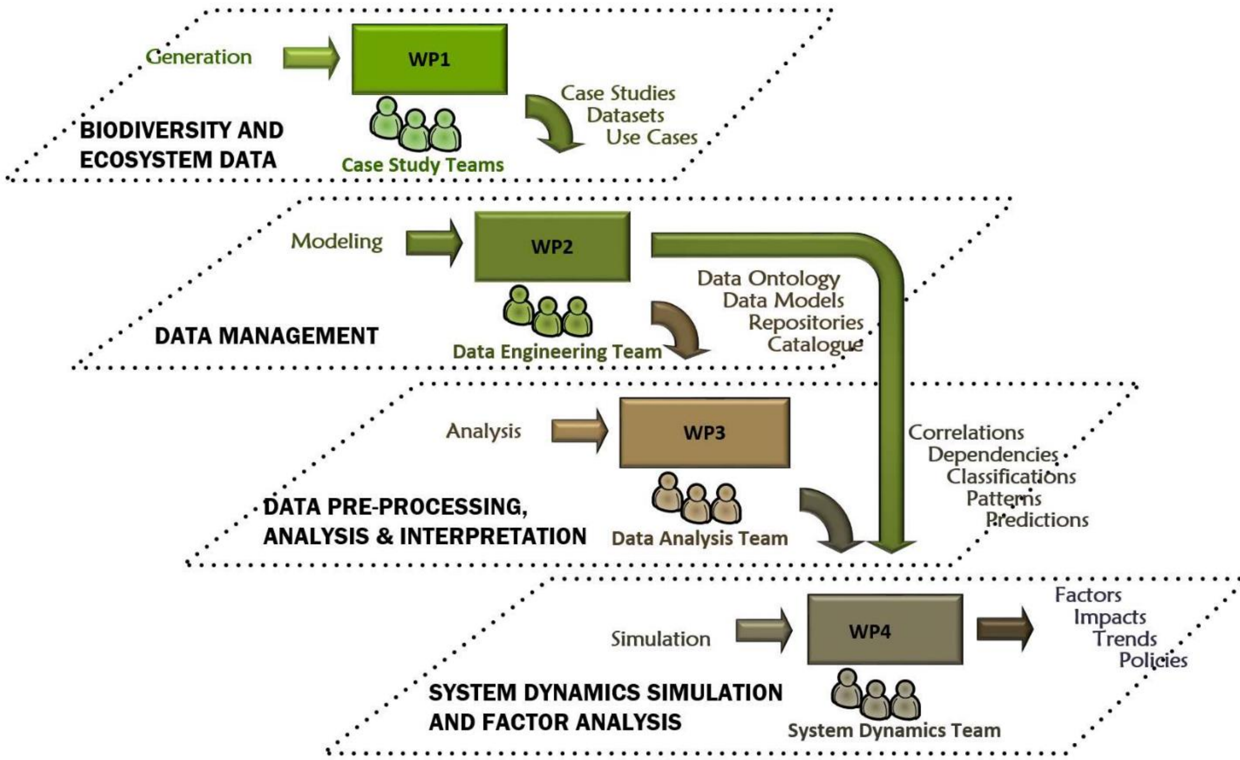
[1] Curry, Edward (2020), Curry, Edward (ed.), "Dataspaces: Fundamentals, Principles, and Techniques", Real-time Linked Dataspaces: Enabling Data Ecosystems for Intelligent Systems, Cham: Springer International Publishing, pp. 45–62, doi:10.1007/978-3-030-29665-0_3, ISBN 978-3-030-29665-0

[2] AWS, Inc. "What is a Data Warehouse?" [Internet]. Available from https://aws.amazon.com/what-is/data-warehouse/ [Accessed 01/12/2024]

1.3. Methodology of Working with the Data Space

The methodology for working with data in DiverSea project is illustrated in Figure 2.

Figure 2. Operational scheme of the working methodology



Since the DiverSea data space comprises a common logical resource rather than a physical repository, access to the data is indirect, via the registration of data location, instead of direct, using drivers to the physical repositories containing the actual data. The common *Data Model* of the DiverSea data space provides the data unification, needed as a prerequisite for the analytical work. It is based on a common *Data Ontology*, which specifies the details about the available data in DiverSea data space. Throughout the project work this ontology will be updated and extended to accommodate new developments supporting the project objectives.

The logical conceptualization of the methodology for working with the DiverSea data, which will take place in WP3 and WP4, is incorporated into DiverSea data ontology. It will also incorporate the description of the preparatory operations which will be executed within WP2 for integration and standardization purposes. The technical details for accessing the data in its final form will be provided in the *Data Catalogue* that will be developed during the next phase of the project within WP2 under Tasks 2.3 and 2.4.

The analysis of the state of the marine ecosystem, its dynamics, and the factors determining them will rely not only on the data created within the DiverSea project but also on the historical data collected by the partner organizations prior to the project and the data stored in public repositories like **EmodNet** and **Copernicus**. This leads to a distributed scheme of hybrid ownership of the data in the *Data Repositories* of DiverSea dataspace, where some of the data will be local, some of it will be remote. The role of an integrator of the data space will be fulfilled by the Data Catalogue, to be developed for the second version of the data space. It will serve as a one-stop data store not only within DiverSea but also for the international community of maritime researchers.

1.4. Outcomes of Deliverable D2.4

As a result of the design considerations discussed previously, **DiverSea** data space provides a logical unification of data resources from the various case studies conducted in WP1 rather than a physical integration of the data management workflow. As such, it is a set of unified datasets, supporting multiple independent regional analytics, rather than a single fully integrated and synchronized data space, supporting multiple centralized data analytics. This concept simplifies both the storage of the data generated in WP1 and its preparation for further processing in WP3 and WP4. At the same time, it allows much more direct support for the publication of generated data and its subsequent integration with the European repositories such as **EmodNet** and **Copernicus** as outlined in tasks T2.3 and T2.4 of WP2.Deliverable D2.4 presents the first version of the DiverSea data space by providing several artefacts:

- Metadata of existing datasets available to project partners that are involved in the regional case studies of WP1 at the beginning of the project (data from public repositories and previously generated historical data available at the partners).
- Metadata of existing datasets generated from regional case studies of WP1 during the first year of the project operation. This metadata is embedded in the data ontology of DiverSea data space which is documented in this deliverable.
- Datasets containing raw data generated in the regional case studies of WP1. Their metadata will be registered in the data catalogue developed under Tasks T2.3 and T2.4 in the next project phase, while the processed datasets will be integrated with **EmodNet** and **Copernicus** in future phases.
- An integral data ontology for logical unification of the diverse data generated in WP1 case studies. This ontology includes operations for processing raw data, preparing it for analysis in WP3 and WP4.

- Unified data model for raw data generated by each WP1 case study, designed for local, regional or public use. In the next project phase, this model will be registered in the data catalogue for broader, common use.
- Database scripts for creating local databases capable of accommodating physically raw data for local, regional, or public use. These scripts have been used to create a central repository at GATE Institute of Sofia University, which coordinates WP2 activities. This repository will be used for experimentation and for support of regional initiatives such as the Argonaut Corridor (Black Sea-Sea of Marmara-Aegean Sea-Adriatic).

In this report, we have also included a *vocabulary of terms* as a separate artifact (see Annex 1). This vocabulary serves as a bridge between WP2 and the technical work packages handling project data (WP1, WP3 and WP4). It also lays the groundwork for the communication protocol to be developed in WP5. Given the diverse terminology used by different research communities, this vocabulary facilitates mapping and unification for avoiding ambiguity. Furthermore, the vocabulary of terms serves as the starting point for developing the DiverSea data ontology and is also incorporated in it.

2. Case Studies Meta Data

The metadata plays a crucial role in designing the data space on both conceptual and physical levels.

2.1. Historic and Public Datasets

Historic and public datasets are relevant to the datasets gathered during the case studies because they can support various data analytics and impact assessments. Because of this, at the start of the project we developed a metadata template to describe these datasets, which were available to the partners (see Table 1). Subsequently, it will facilitate analytical tasks in WP3 and WP4, such as classification, finding patterns, and time series prediction.

Table 1. Template for Historical and public data available to the Case Studies (see Annex 4 for the information)

TEMPLATE FOR GATHERING META-DATA										
Data (region/ provider)	**Type of data** (sensor data, photographs, documents, messages, diagrams, etc.)	**Source Format** (binary, text, structured, hypertext and 2D/ 3D, etc.)	**Location & size** (databases, data files, Web services, APIs, etc.)	**Transfer mode** (one-off upload, repository import, online streaming, periodic batches, etc.)	**Meta-data** (text-based, SQL DDL, ER model, UML model, ontologies, etc.)	**Standards** (published models, ontologies, professional standards, etc.)	**Origin** (partners, ISPs, NGOs, industrial, commercial, public, etc.)	**Ownership** (private, licensed, commercial, public, etc.)	**Use** (private, restricted, public, etc.)	**Availability** (expected availability in the case of new data only)
DNA/ UCH,UCSD										
RNA/ NTNU,UCSD										
Norwegian/ NTNU										
Atlantic/ UPORTO										
Balearic/ AT										
Adriatic/ UNIVPN										
East Med/ DEU										
Marmaris/ AKD										
Aegean/DE U NKUA										
Black Sea/ IMB										

2.2. Case Study Datasets

The metadata of the Case Study datasets, used to build the data model, was based on information obtained from the partners working on them. For this purpose we developed a template and sent it to all teams working on Case Studies as shown below:

Template for collecting additional information about biodiversity data generated in WP1 case studies (see Annex 5 for data)

Case Study No.:
(the separate regional case studies are fixed – see file WP1 Case Studies.doc, attached)
Dataset No.:
(each partner case study may have several datasets; for describing more copy the template)

1. **Geographic Location:**
(*identify the geographic area covered by the data to understand the research context and spatial boundaries*.)
2. **Classification** (*specify the type of data included - sediments, pollution, biochemical, eDNA, plantation, species, audio, visual, video, human activities such as tourism, transport, industry, etc.*)
3. **Economic Sector** (*Identify the economic sector(s) related to the dataset to understand the potential use and relevance of the study*)
4. **Data source** (*describe the technical aspect of the data - underwater measurements, aerial scans, photogrammetric images, videos, analytically generated data, statistical distributions, etc.*)
5. **Dataset formats** (*text descriptions, numerical measurements in value/vector/matrix format, images, time series, audio recordings, videos, numerical distributions, geolocation distributions, spatial distributions, temporal distributions, etc.*)
6. **Major data descriptors** (*point at the unique identifiers, essential attributes, and, whenever possible – the full attribute list with data types*)
7. **Data Samples** *(provide couple of data samples in an additional file)*
8. **Regularity of data collection** (*how the data is collected - one-off, over specific period, daily/seasonally/annually, streamlined in real-time)*
9. **Data Quality and Limitations** (*Assess the data quality, including noise, completeness, accuracy, and potential presence of biases.)*
10. **Data management practices and tools** (*detail any available data models, metadata descriptions, physical memory storage, repository software, data sharing policies, etc.*)
11. **Relevant standards and professional guidelines** *(for classification, modelling, cataloguing, publishing, storing, sharing and exchanging data pertinent to the dataset)*
12. **Quantitative description** (*estimate the number of records, the necessary amount of memory per record in bytes, kilobytes or megabytes, and total volume of the dataset in kilobytes, megabytes or gigabytes*)
13. **Potential data services** which might be provided to external stakeholders for research purposes (*consider raw datasets, filtered, smoothened and enriched data, aggregated and accumulated historical data*)
14. **URL for Download** (when available)

2.3 Case Study Data Samples

Although the WP1 case studies began at different periods during the first year of the project, to validate the design of the data space and to obtain realistic overview of the data under Task 2.1 the team of WP2 collected from partners data samples containing real data. Particularly important for the design were samples of eDNA/rDNA files, as their metadata was in R programming language and required pre-processing which was carried out by the data engineer. Additionally, some data samples, such as binary files containing images and sound, were kept intact since their metadata was embedded within the files themselves. As a result, the database design, informed by the metadata descriptions, was adjusted to reflect the structure and content of these files.

The three sources of data were subsequently reconciled and, as a result, one unified data model was created that forms the core of the DiverSea data space. It became the basis for the database design which was selected to accommodate raw data obtained from various case studies. The data model for the second version of the data space, however, will be developed in collaboration with WP3 and WP4. This collaboration is essential to address specific pre-processing requirements of the data, which may lead to adjustments in the structure and content of the datasets, ensuring they are suitable for the purpose of the intended analysis.

3. Ontology of DiverSea Data

Ontology as a philosophical term refers to the world „as it is“ —an objective reality reflecting the external world. In a technical context, ontology refers to a model of the world „in a digital format“ — an objective reality shaped by our subjective understanding of the external world. From this point of view the ontological models articulate individual and professional knowledge of the perceived world and social practices in the real world. These models enable mutual understanding and facilitate collaboration by providing a shared conceptual framework. This principle underpins our use of the ontological model as the logical foundation for the DiverSea Data Space. It allows specialists from diverse domains – biologists, physicists, chemists, computer scientists, engineers, social scientists and economists – to collaborate effectively. DiverSea Data Space ontology serves as a logical model covering the initial version of the Data Space defined in WP2.However, its scope already extends beyond these initial specifications.

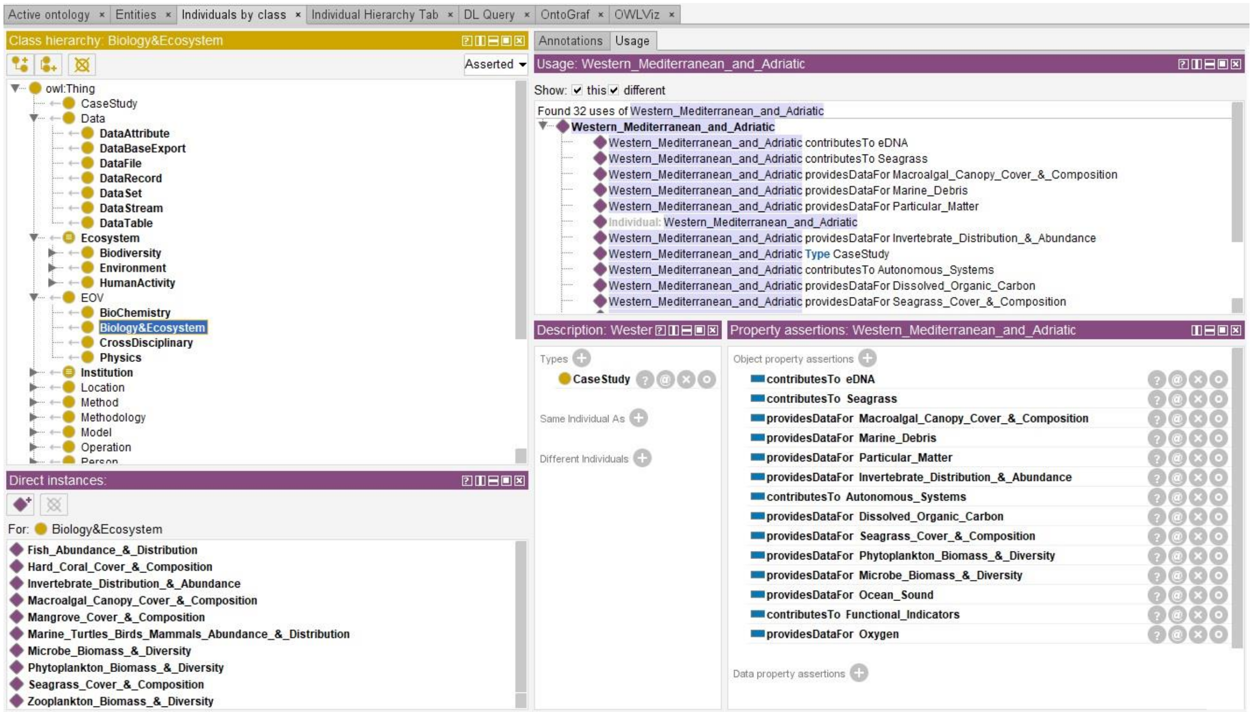


Figure 3. View of the ontology within the console of Protégé

The data ontology has been developed according to the methodology of logical modeling using the standard semantic modeling languages using Protégé and is available in a variety of standard serialized formats: as an .RDF file, as an .OWL file and as a .JSON file (see Annex 2). Figure 3 shows a view of a fragment of the ontology within the Protégé dashboard.. Additionally, Figure 4 provides a visualization of fragment of the taxonomy of concepts used in the ontology.

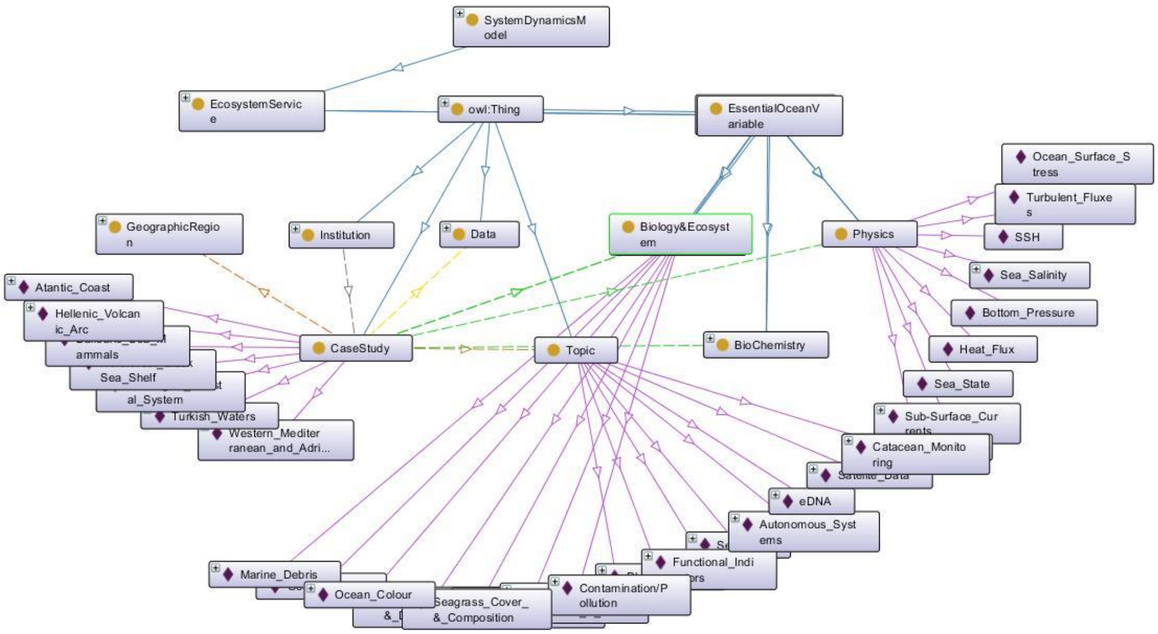


Figure 4. Diagram of a fragment of the ontological taxonomy of concepts

3.1 Concepts and their Taxonomies

The starting point for developing the ontology was the vocabulary of terms (see Annex 1) which was developed in collaboration with marine researchers, data analysts, specialists in system dynamics and impact analysis. Domain-specific information was provided by WP1 partners during a series of joint meetings, supported by presentation materials describing the project's case studies. Table 2 lists the key concepts used to build the ontology, organized in taxonomies reflecting their domain-specific classifications, along with corresponding instances.

Table 2. Concepts with object instances in DiverSea ontology

GROUPING AND TAXONOMIC CLASSIFICATION OF ONTOOGICAL CONCEPTS				
Group	**Purpose**	**Classifications**	**Classes**	**Instances**
Project	Management Perspective	Consortium Work Package Role Task	Consortium Dissemination Stakeholder CaseStudy Deliverable Topic Team	ARGONAUTS DIVERSEA_WP6 EC BALLEARICS DATASPACE PHYTOPLANKTON TEAMGATE
Domain	Professional Perspective	Engineering Science Business Media Governance	DataEngineering Biology Tourism Magazine Law	DATABASE_MANAGEMENT MARINE_BIOLOGY SEA_DIVING BIODIVERSITY_DATA_JOURNAL (BG) CONVENTION_OF_WETLANDS (NO)
Ecosystem	Context Perspective	Abiotic Biotics Human Activity	Wetlands Marine Mammals Offshore Wind Farming	BLANKEN_BOG (NO) RISSO_DOLPHINS (SP) PRINCESS_ELISABETH_ZONE (B)
Biodiversity	Content Perspective	Levels Forms	Comp, Structural, Functional Genetic, Habitat, Species	PHOTOSYNTHESIS POSIDONIA
Area	Localization	Physical Digital Societal	GeographicRegion InternetSite Civilization	EASTERN_MEDITERRANEAN HTTPS://WWW.EEA.EUROPA.EU EUROPEAN
Resource	Means available for working on the project	Methodology Method Model Technology Tool Standard	DataCollection Bioinformatics SystemDynamics Biological Hardware MarineBiodiversity	SATELLITE_IMAGING METABARCODING STRESSOR WASTEWATER_TREATMENT eDNA_SAMPLING_KIT LAB_TESTING_AND_CALIBRATION
Data	Information Content	Self-contained Fragment	DB Export, Stream, Set, File Collection, Record, Attribute	*See the datasets in the appendices*
Institution	Social establishments	Government Agency	Directorate Tour Operator	DG_ENVIRONMENT (EC) BENAGIL_BOAT_TOURS (PT)

		Society Company Centre Institute University	Conservation Society Biodiversity Experts Diving Centre Research Institute Maritime University	AGDENIZ_KORUMA_DERNEGI (TR) NIRAS A/S (DK) CENTRO_SUB_MONTE_CONERO (IT) INSTITUTE_OF_MARINE_BIOLOGY (UA) WORLD_MARITIME_UNIVERSITY (SE)
Person	Individuals	Academic Technical Public	Researcher Diver Tourist	*Removed for privacy protection*

3.2 Qualitative and Quantitative Descriptions

The qualitative and quantitative descriptions of the concepts and their instances are modelled using attributes called "data properties" in Protégé terminology. List of the essential data properties used for this purpose is shown in Table 3. Protégé supports two types of data for these properties: 1) standard numerical data types available in the modeling languages for quantitative descriptions, 2) string data types used for qualitative descriptions, such as names, identifiers, labels, and lists of strings.

Table 3. Data Properties (attributes) for describing the concepts in DiverSea ontology

ATTRIBUTES OF ONTOLOGICAL CONCEPTS				
Group	**Purpose**	**Classifications**	**Data Properties**	**Data Types**
SELECTORS	Provide ways to identify, refer to or select the concepts and individual objects	Identifier	ID Email Name Index Abbreviation	String String String Integer String
DESCRIPTORS	Describe qualitatively or quantitatively the state of the individual objects or collections	Static Dynamic	Volume Dimensions Size Weight Quantity Material Shape Color	Decimal Decimal Decimal Decimal Integer String String List
LOCATORS	Locate relatively or absolutely in a list, on a map or in coordinate system the objects	Physical Digital	Postal Adress Coordinates URL IP Address Memory Address	Text Numerical Pair String String Hexadecimal
ASSESSORS	Assess the state of the individual objects or collections	Living Working	Health Age Security Safety Scalability Sensitivity	String Integer Text Text Text Text
PROCESSORS	Process physically or digitally artefacts	Software Language Mode	Service Markup Transportation Protocol Format	String List List List List

3.3 Conceptual Dependencies

Conceptual dependencies such as associations, structural aggregations, and functional dependencies, are modelled using object properties in Protégé terminology. They are used for structuring the terminology by establishing various relationships, including associations, aggregations, compositions, temporal ordering, causal chaining, and functional mappings. Each relationship can be restricted by domain and range, cardinality and optionality, and a predefined functional indexing scheme which offering additional flexibility to the model. This aspect of the ontology development is very important because it guides the development of a unified data model so that effectively supports a wide range of analytical tasks.

The logics behind the ontological modelling, the *Description Logic*, also facilitates the classification of relationships into taxonomies, simplifying the process of building the data ontology. Although relationships are often defined at purely object level, their grouping into more abstract categories enhances technical convenience and flexibility.

Table 4 presents the key groups of conceptual relations, including their classification, structural restrictions and functional characterisation. Currently the data ontology of DiverSea data space offers a detailed conceptualization of data management, the focus of WP2. However, it can naturally extend beyond the data to model the analytical operations, pipelines and workflows for analyses involving statistical methods and machine learning, system dynamics simulation and factor analysis. At present, the ontology encompasses nearly 500 concepts, organized into 10 partitions and 40 taxonomies, described using 65 different attributes and 70 different object properties. It is a living organism that will evolve as the project progresses. It aims to cover also the final version of the data space, as well as the methodology for analyzing marine ecosystems and the biodiversity dynamics specified in WP3 and WP4. By the end of the project, the ontology of the data space will encapsulate the knowledge acquired into a comprehensive knowledge base, which can be used to support further research and innovation in marine biodiversity monitoring and coastal ecosystem management.

Table 4. Object properties (conceptual dependencies) in DiverSea ontology

STRUCTURAL AND FUNCTIONAL DEPENDENCIES BETWEEN ONTOLOGICAL CONCEPTS					
Group	**Purpose**	**Classifications**	**Relationships**	**Domain**	**Range**
Project	Stating the different players, roles and case studies in project	CaseStDependence TaskDependencies WorkDependencies ProjectRoles	supplyingData nextOperation sharingWork expert	CaseStudy Operation Institution WorkPackage	Data Operation Work Leader
Domain	Case Studies Research informs	NaturalScience	sourceFrom featureOf	Data Data	CaseStudy Ecosystem
Ecosystem	Relating the parameters of the ecosystem		determinedBy affectedBy endangeredBy variableOf	Seagrass SeaMammals Fish EOV	Sediment Tourism&Boating Fishery&Aquacult Ecosystem
Biodiversity	Identifying ecosystem		biodiversityOf variableOf	Biodiversity EBV	Ecosystem Biodiversity
Area	Physical and Technical localization	Geography	connectedWith population localization	SeaRegion Area Area	Straight Biodiversity Ecosystem
Resource	Describing resources on a level of abstraction		useIn ImplementedBy basedOn supportFor	Methodology Technology Method Tool	Method Method Model Technology
Data	Precise data description	DataPartitioning DataProfiling	aggregatedIn generatedBy originatesFrom storedIn	DataAttribute Data Data Data	DataRecord Physical Physical Digital
Institution	Organization of the work	Structural Functional	divisionOf directorOf partnerRole worksAt involvedIn	Person Institution Institution Person Institution	Institution Person Institutional Role Institution Project
Person	Human resources		personRole	Person	Personal Roles

4 Technological Framework for Data Management

4.1 Data Platform

The DiverSea Data Space, both in its initial version for managing raw data from project case studies and its final version for supplying pre-processed data for biodiversity dynamics analysis, requires substantial technical support during development, testing and deployment phases. Although all partners possess technical infrastructures capable of providing some support, a centralized resource is needed to coordinate the work with data. For this purpose, as a partner responsible for WP2, GATE Institute has offered its own infrastructure as a central project resource for collecting (T2.1), pre-processing and storing (T2.2) and cataloguing of the project data (T2.4) as a preparation for its analysis in WP3 and WP4 and before integration of the datasets with the European Marine repositories like **EmodNet** and **Copernicus** (T2.5).

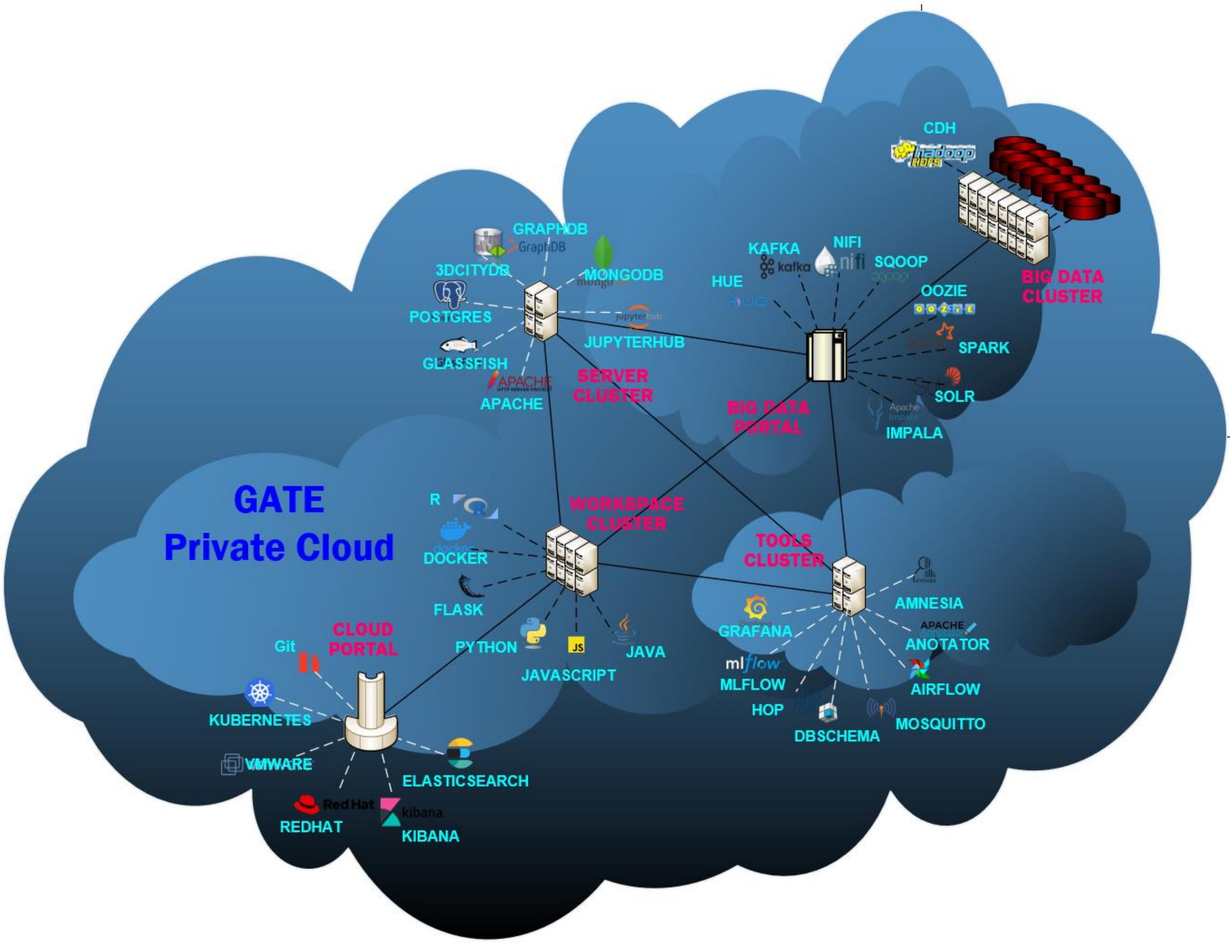


Figure 5. GATE Institute Data Platform

GATE Data Platform offers ample storage capacity and computational power to meet all the requirements of the DiverSea data space. This state-of-the-art infrastructure, designed and entire implemented as a Private Cloud (see Fig. 5), was funded by the Bulgarian Government[3] and established with the support of EU[4] . It allows working both locally (within the network of GATE) and remotely (from any Internet-connected location). The platform is equipped with multiple software tools which support the entire process of managing and processing of a variety of data formats along the data lifecycle – *collection*, *transportation*, *pre-processing*, *storage*, *post-*

[3] Operational Programme Science and Education for Smart Growth under Grant Agreement No. BG05M2OP001-1.003-0002-C01
[4] H2020 WIDESPREAD-2018-2020 TEAMING Phase 2 under Grant Agreement no. 857155

processing, *analysis*, *reporting*, *interpretation* and *archiving* (see Table 5). All of them operate in server mode, providing essential data services needed by the project teams in WP1-WP4.

Table 5. Software for Data Processing available on GATE Data Platform

Operations	Description	Software products
Pre-processing	Processing data at source, in transition & before analysis	**Annotator** (annotating), **Amnesia** (anonymizing), **SemTUI** (enrichment)
Ingestion	Transporting files, repository exports, messages & streams	**Mosquito** (messages), **Kafka** (streams & messages), **NiFi** (streams & files), **Hop** (all)
Persistence	Storage in file, hypertext & binary format	**PostgreSQL** (SQL), **3DCityDB** (GML), **MongoDB** (JSON), **GraphDB** (RDF), Git (files)
Post-processing	Transformation, analysis & integration	**Python** and **R** (data pipelines), **Java** (operations), **JavaScript** (presentations)
Big Data Management	Storing, retrieving, mapping & searching Big Data	**Hadoop** (large volumes of data)
Big Data Analysis	Analysing Big Data using statistics, ML & RL methods	**Spark** (analysis of large distributed data),
Reporting	Data interpretation, explanation & reporting	**Kibana** (statistics), **Grafana** (diagrams), **Node.js** (hypertext), **Jupyter** (reports)
Visualization	Data diagramming, 3D mapping & 3D rendering	**ESRI**, **Terra** and **QGIS** (geolocation and geospatial data)

4.2 Modeling and Design Tools

GATE project team working on WP2 also uses additional tools for implementing data services. They include data modeling, system design, software development and service deployment. These tools are used in autonomous or in client mode of operation on the individual workstations which can connect seamlessly to GATE cloud.

Modeling plays an essential role in the development of DiverSea Data Space. It is used to model the conceptual ontology of the data space, the data collected and stored in it, the workflows for processing the data, as well as the corresponding operations performed on it. In this section, we review the tools used for modeling the data on logical and physical level– the ontology modeling tool Protege and the database modeling tool DBSchema.

Protégé

Protégé is a free graphic editor developed by Stanford Research Institute[5] that supports all three standard semantic modeling languages maintained by the World Wide Web Consortium (W3C): RDF, OWL and SWRL[6]. By strictly adhering to these standards, ontological models created in Protégé can be deployed to any compliant graph databases, including **GraphDB**, which is available on GATE Data Platform. Despite its ease of use, **Protégé** is a powerful tool which supports multiple formats for storing the models – XML, JSON, LaTeX and Turtle. It is also a versatile tool, allowing to formulate and execute queries using the built-in SPARQL interpreter, to conduct logical inference using several popular reasoners and supports several housekeeping operations, including consistency checking and visualizations of the ontological model.

DBSchema

DBSchema is a free, universal database design and management tool which supports structured, semi-structured and unstructured data formats for all popular relational, object-relational and object databases using standard formats such as SQL and JSON, as well as for many Big Data repositories and distributed file systems using non-standard data formats[7]. It provides all the necessary capabilities needed to build the physical data models and generatecode for the three main data repositories on GATE Data Platform – PostgreSQL, MongoDB, Hadoop.

4.3 Data Repository

Final destination for the datasets generated during WP1 case studies should be European marine and biodiversity data repositories, such as Copernicus and EmodNet. However, before integration with these repositories, the data requires temporary local storage for collection, aggregation, accumulation and pre-processing. This process is coordinated closely with WP3 and WP4, where the project data will be analyzed. As a partner responsible for data management, GATE Institute relies on several repositories, available on GATE Data Platform, which provide all the necessary tools and resources for this purpose.

This section describes briefly some of the available repositories on GATE Data Platform since the preliminary analysis of the data indicated that most datasets are either in a purely relational format (structured collections of data records) or completely unstructured (binary data files). Therefore, the primary need is for standard relational database systems and standard file systems. In addition, since no data arrives in real-time, there is no need for data repositories designed for fast-growing data or transactional systems.

[5] https://protege.stanford.edu/
[6] https://www.w3.org/2001/sw/wiki/Main_Page
[7] https://dbschema.com/

PostgreSQL

The most advanced free database management software[8], supporting data management in flat relational format - the predominant data format used by DiverSea partners - is used by GATE Institute to implement the central data repository for the DiverSea data space. This repository stores raw data produced by the WP1 case studies. Partners working with biodiversity data are able to implement their own local copies of the database using the scripts developed within WP2 (see Annex 2).

MongoDB

MongoDB is a widely-used, highly scalable NoSQL database management system ideal for storing fast-growing data typically collected and ingested in real time[9]. While **MongoDB** is a commercial database, it offers a free community edition and can be used locally. Real-time data, such as continuous sensor measurements, is not currently considered in the DiverSea WP1 case studies. However, the community edition of **MongoDB** is already available on GATE Data Platform. If future expansions involve real-time data collection, **MongoDB** would be an excellent choice for storing such data.

GraphDB

GraphDB is a leading semantic graph database software for storing ontological models and performing semantic data analysis using extensive build-in functionality, including a driver for interacting with popular generative AI software like ChatGPT[10]. Developed and marketed by Ontotext, a company that sponsors and partners with GATE Institute, the full commercial version of GraphDB has been donated to GATE. As part of WP2 GATE Institute has developed an ontology of DiverSea data space (See Annex A3), which is stored in GraphDB on GATE Data Platform and soon will be available for remote access to all project partners.

Hadoop

Hadoop is a distributed file storage system, originally developed from early Google concepts, specifically adapted for managing Big Data. Despite being one of the oldest systems of such kind, Hadoop remains highly effective due to its mature architecture and the wide range of supporting tools, making it suitable for storing and managing Big Data both at rest and in motion. GATE Data Platform includes a commercial version of Hadoop from Cloudera, Inc., which is fully compatible with the open-source version of Apache Software Foundation[11].

[8] https://www.postgresql.org/
[9] https://www.mongodb.com
[10] https://graphdb.ontotext.com/
[11] https://hadoop.apache.org/

Currently, the metadata collected from the WP1 case studies does not suggest data volumes exceeding the capacity of relational databases like PostgreSQL (for static data) or NoSQL systems like MongoDB (for real-time data). However, if the volume of generated data significantly increases in later stages, it would be possible to accommodate it in the large storage space available on GATE Data Platform using Hadoop's enormous storage capacity (approx. 5 petabytes). Additionally, Hadoop's ecosystem supports distributed data analytics tools like Spark[12], enabling comprehensive data analysis directly within the Hadoop environment.

4.4 Data Catalogue

An essential requirement for collaborative work with shared data is having a single point of registration; to address this need, WP2 includes a dedicated task for cataloguing the data generated in the WP1 cases studies (Task T2.4). Since all data in the project case studies is available in files for data catalogue, the WP2 team responsible for this task considers the use of a repository for storing files on a dedicated file server. The chosen tool is Git[13].

Git

Git is a free, distributed version control system widely known for powering the popular file repository platform for collaborative work on the Internet, GitHub. Apart from its role as a file server, supplying datasets for common use, Git also facilitates building, sharing and deploying code for data processing. This capability is especially valuable in large collaborative projects, where tasks such as data collection, storage, analysis, simulation and factor analysis are performed by different partners. In the case of DiverSea, Git supports collaborative work across WP1, WP2, WP3 and WP4 where data is used for various purposes by different teams.

For this purpose, Git has been installed on GATE Data Platform, and the files from the appendices of this report have already been placed there. The organization of the file repository, as outlined in T2.4, is scheduled for the second half of the project's second year, after the complete collection of raw data from the case studies. It will be used to persist the data for the first version of the DIverSea data space which is expected to be completed by the first half of the next year and populated with all relevant datasets containing raw data from the case studies. Until the data catalogue is ready the files from the Annexes will be accessible via the public account on the file server of GATE Institute, reachable over the Internet using a standard Web browser at: https://drive.gate-ai.eu (password DiverSea).

[12] https://spark.apache.org/
[13] https://git-scm.com/

5 Data Model of the Data Space

The first version of DiverSea data space is dedicated to managing the raw data, which has been generated while working on the selected case studies, without preparing it for the subsequent analysis in WP3 and WP4. The data model of the second version of the data space will also account the results of the raw data pre-processing and its preparation for analysis, which may differ significantly from the data model of the raw data. This means that the data model of the first version will change in the next phase, when the analysis begins, and will be updated in collaboration with WP3 and WP4, the primary consumers of the data.

5.1 Design Considerations

Since data has been arriving while the case studies are still ongoing, the data model of the first version of DiverSea data space, created under WP2, reflects both the expectations and the realities of the datasets collected from partners. Expectations were gathered from partners at the beginning of the year, while additional data samples collected later helped unify the datasets and produce a more realistic model. This process will naturally continue as new data arrives.

The data generated in the case studies is entirely offline, indicating that it is not linked to real-time data sources or cloud-based systems which limits the data model to static data without the need for streaming. Analysis of the provided data samples and metadata from the selected case studies revealed that the data is entirely structured, either organized in a consistent and predictable format with well-defined rows and columns or stored in binary files containing pictures and sound recordings. This led to further simplification of both the data model and the implementation of the repository needed for storing the raw data:

- the files are stored in a file directory within the data space repository, implemented as a virtual drive;
- structured data is stored in a relational database using the popular open-source software PostgreSQL DBMS.

The data models created for the SQL repository of DiverSea data space are used to generate the SQL DDL for implementing the database repository. Access to the repository over the Internet requires a VPN client to meet higher security requirements, but the file system for storing binary files was created on a virtual drive in the GATE private cloud and is publicly accessible over the Internet using a standard web browser.

5.2 Historical Data and Public Data relevant to the case studies

Based on the metadata obtained from the partners in the beginning of the project the following data sources and data formats of the datasets are available to them from previous work or from public sources (see Annex 4 for the file locations containing this metadata):

- **Atlantic Coast**

Data sources: Aerial camera survey and laser scanning; Underwater camera surveying, measurements with eDNA Prototype filter with subsequent onshore analysis and measurements for temperature, conductivity and pressure; **EMODnet**

Data formats: Numerical measurements in matrix format, 3D point cloud, image mosaics.

- **Balearics**

Data sources: Hydrophone acoustic samples; Visual observations; Boat data; **EMODnet**; Copernicus; Marinetraffic.com

Data formats: Numerical measurements in matrix format.

- **Black Sea**

Data sources: Underwater measurements of temperature, salinity, depth, etc. and field samples analyzing plankton, benthos, macrophytes and fish.

Data formats: Tables with numerical measurements; tables with numerical measurements in value format and qualitative categorizations.

- **Hellenic Islands**

Data sources: DNA extractions from water/sediment; Preserved water samples; Satellite data

Data formats: Abundance matrices of ASVs / Taxonomy; Numerical measurements

- **Norwegian Coast**

Data sources: Underwater CTD measurements; Species observation

Data formats: CSV, NetCDF

- **Turkish Waters**

Data sources: Photogrammetric and satellite images; Underwater acoustic measurements; Underwater images and video; Underwater measurements of temperature, salinity, depth, etc. and DNA extraction from sea water and sediment samples.

Data formats: Images, vector & raster GIS layers, spatial distributions; Text descriptions and numerical measurements in matrix format; Videos, images, tables with text descriptions; Numerical measurements in matrix format; Lists of species.

This data may be used in WP3 for predictive analytics alongside the current data generated in the WP1 case studies. To support this, it will undergo pre-processing tailored to the needs of the analysis and will be included only in the repository for the second version of DiverSea Data Space. This will enable publication and integration of DiverSea data with the European repositories such as **Copernicus** and **EmodNet**.

5.3 Case Studies Current Data

The data sources and formats of the datasets generated in each WP1 case study are as follows (see Annex 5 for the file locations containing this metadata):

- **Adriatic**

Data sources: Underwater measurement of temperature, salinity and depth; eDNA water and sediment samples; Underwater images and video; Laboratory analysis of sediment and biological samples.
Data formats: Numerical measurements in matrix format.

- **Atlantic Coast**

Data sources: Aerial camera survey and laser scanning; Underwater camera surveying, measurements with eDNA Prototype filter with subsequent onshore analysis and measurements for temperature, conductivity and pressure.
Data formats: Numerical measurements in matrix format, 3D point cloud, image mosaics.

- **Balearics**

Data sources: Hydrophone acoustic samples; Visual observations; Boat data.
Data formats: Numerical measurements in matrix format.

- **Black Sea**

Data sources: Underwater measurements of temperature, salinity, depth, etc. and field samples analyzing plankton, benthos, macrophytes and fish.
Data formats: Tables with numerical measurements; tables with numerical measurements in value format and qualitative categorizations.

- **Hellenic Islands**

Data sources: DNA extractions from water and sediment samples; Preserved water samples; Satellite data
Data formats: Abundance matrices of ASVs / Taxonomy; Numerical measurements

- **Norwegian Coast**

Data sources: eDNA lab analysis
Data formats: R files

- **Turkish Waters**

Data sources: Photogrammetric and satellite images; Underwater acoustic measurements; Laboratory analysis of preserved water and sediment samples; Underwater images and video; Underwater measurements of temperature, salinity, depth, etc. and DNA extraction from sea water and sediment samples.
Data formats: Images, vector & raster GIS layers, spatial distributions; Text descriptions and numerical measurements in matrix format; Videos, images, tables with text descriptions; Numerical measurements in matrix format; Lists of species.

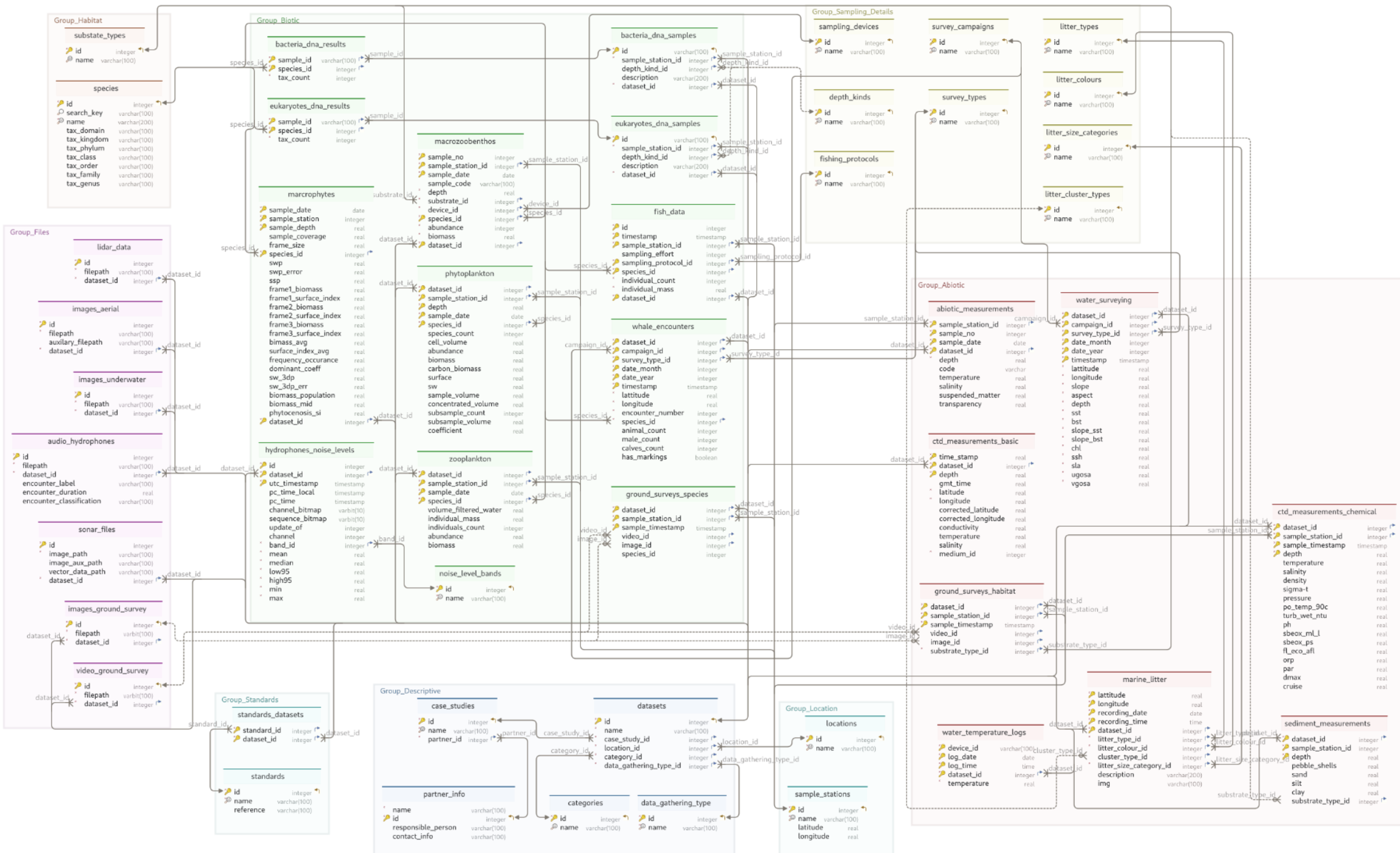


Figure 6. Integrated Entity-Relationship (ER) model for designing the database

5.4 Data Space Data Model

The data model of the DiverSea data space was developed using the metadata gathered from all partners involved in the case studies and was refined using the data samples they provided. The model was created using DBSchema database design software. In this section, we present the model through a series of ER diagrams, which is the standard way for illustrating relational database design. The data model is divided into several parts which group the metadata from different perspective, so that it matches the data ontology presented earlier.

The complete data schema of the model is shown in Figure 6, followed by detailed diagrams of the most important fragments of the model.

Case Study Dataset Model

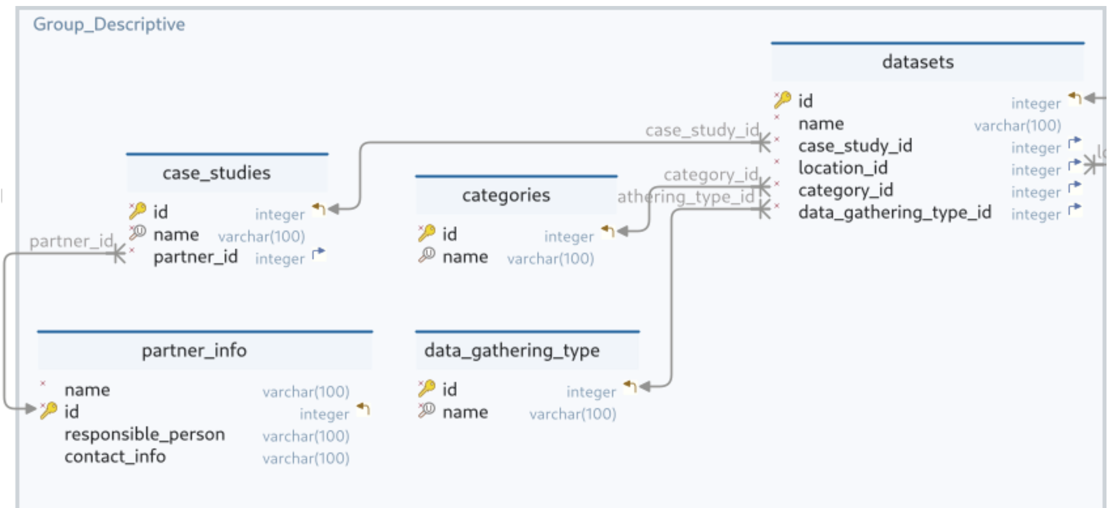


Figure 7. ER model of the descriptive information about the datasets from WP1 case studies

Abiotic Data Model

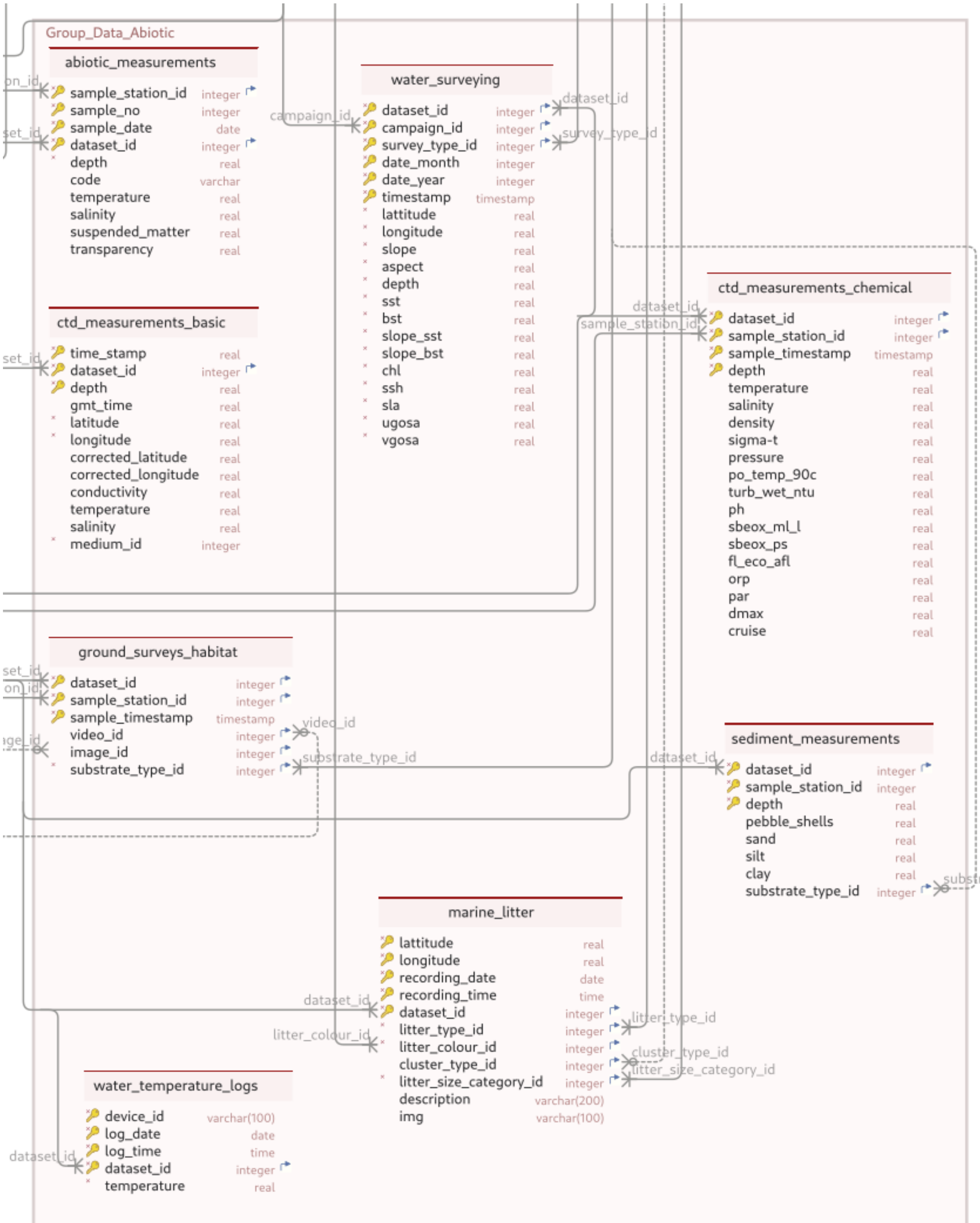


Figure 8. ER Model of the abiotic variables in the different case studies

Biotic Data

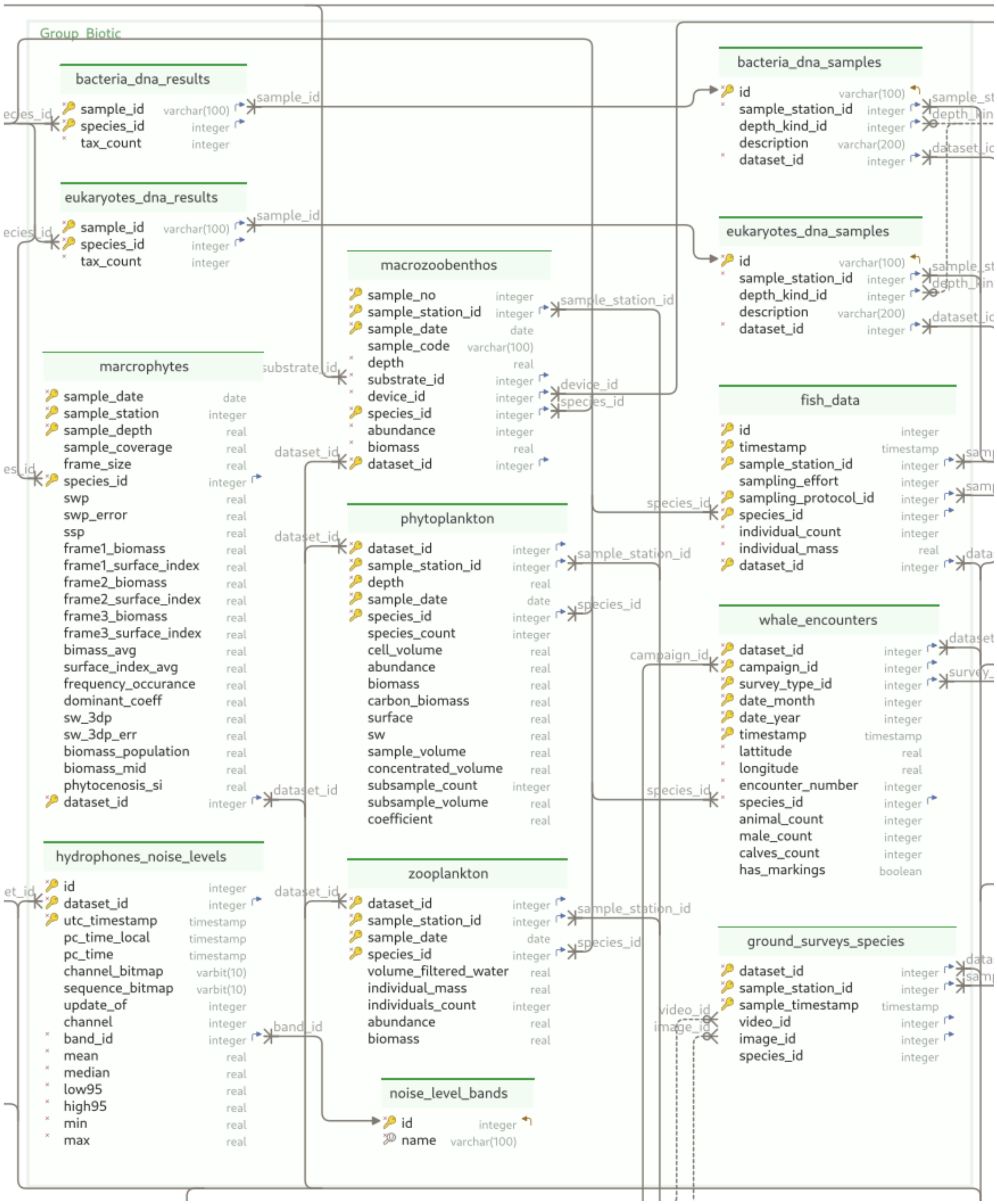


Figure 9. ER Model of the biotic variables in the case studies

Sampling Details

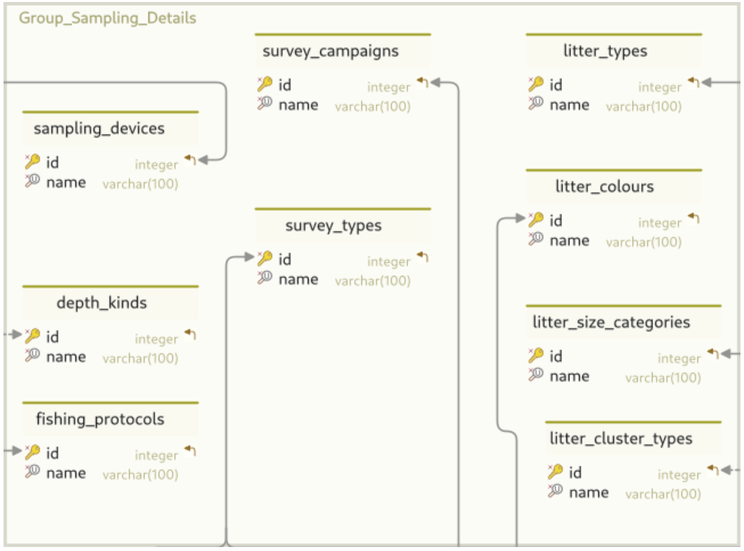


Figure 10. ER Model of the information about sampling procedures used for obtaining data

Auxiliary Data

In addition to the above fragments, the integrated model also contains components for describing the habitat, the ecosystem and the standards which govern the data collection across the case studies.

From the integrated model, an SQL DDL script for the data repository was generated. It was used to create the PostgreSQL database that stores the raw data generated in the WP1 case studies. This database is expected to expand over time and may require incremental additions, which will be made in alignment with the core model presented here.

6 Interoperability and Standards

Interoperability and the adoption of standardized data practices are fundamental to the success of the DiverSea project. As part of Task 2.5: Integration with EU Data Platforms and Systems, this chapter outlines the frameworks, standards, and protocols that ensure the data collected in the DiverSea project are interoperable, accessible, and reusable across multiple platforms and stakeholders. This will affect the design of the second version of DiverSea Data Space which will accommodate datasets prepared for analysis in WP3 and WP4 and to be more valuable for the community of researchers and other stakeholders, it will adhere to the international standards.

The rationale for adopting these standards is clear: marine biodiversity, oceanographic, geological, and remote sensing datasets are highly diverse and complex. To unlock their full potential, these datasets must be integrated, shared, and analyzed seamlessly across regional and international platforms. Compliance with well-established standards ensures that DiverSea data can support scientific research, environmental management, and policy development on a global scale.

By aligning with EU data frameworks and internationally recognized standards, DiverSea enhances the quality, reliability, and usability of its data. The mapping of datasets to appropriate standards, policies, and platforms will be common across all WP1 Case Studies. This approach strengthens collaboration among marine research communities, supports long-term data preservation, and facilitates data integration with global and regional marine data initiatives.

6.1 Standards and protocols for data cataloguing, sharing and integration

The DiverSea Data Space considers employing a comprehensive set of internationally recognized standards and protocols to ensure datasets are properly catalogued, shared, and integrated. These standards have been selected based on their relevance to different types of data, their widespread adoption in the marine research community, and their ability to facilitate seamless data exchange across platforms.

ISO 19115 is considered as the primary standard for documenting geospatial metadata. This standard provides a detailed structure for describing the content, quality, and lineage of geospatial datasets. Geospatial datasets such as bathymetric surveys, UAV imagery, side-scan sonar data, and remote sensing data may be catalogued using ISO 19115. This approach is intended to facilitate integration with platforms like EMODnet, SeaDataNet, and GOOS (Global Ocean Observing System), enhancing the discoverability, traceability, and interoperability of geospatial data. To fully ensure compatibility, DiverSea will prioritize compliance with SeaDataNet/Cloud common vocabularies, metadata formats, and data transport standards, as explicitly required in Task 2.5 of the Grant Agreement.

For biodiversity data, the DiverSea Data Space considers adopting *Darwin Core* due to its consistency in documenting species occurrence, taxonomy, and ecological information. This standard supports datasets such as eDNA analyses, species distribution surveys, and underwater visual census. By using Darwin Core, datasets are expected to be compatible with global biodiversity platforms like GBIF and OBIS, enabling large-scale biodiversity assessments and facilitating international data integration. Furthermore, DiverSea considers integrating datasets with long-term environmental monitoring infrastructures such as eLTER, LifeWatch, ERIC, and EMSO to support wide accessibility and interoperability.

To ensure compliance with European spatial data infrastructure policies, the DiverSea project aims to align with the INSPIRE Directive. This directive promotes interoperability of spatial datasets across the EU and is relevant for datasets involving marine habitats, remote sensing imagery, and environmental monitoring. By adopting INSPIRE standards, data integration with platforms such as EMODnet and the Copernicus Marine Environment Monitoring Service (CMEMS) is anticipated, supporting cross-border collaboration and policy development.

For oceanographic data, the DiverSea Data Space considers using MEDATLAS to standardize datasets such as CTD profiles, sediment core analyses, and water quality measurements. MEDATLAS supports data consistency and quality control, making it suitable for sharing data through platforms like SeaDataNet and EMODnet. This approach is expected to ensure that oceanographic datasets are reliable, comparable, and suitable for long-term analysis and monitoring.

NetCDF (Network Common Data Form) is under consideration for managing multidimensional datasets, particularly those involving real-time and time-series data. Potential applications include Argo float measurements, hydrophone recordings, and sensor array data. NetCDF is known for its efficiency in storing, retrieving, and sharing large datasets and may be used for integration with real-time platforms such as GOOS and CMEMS.

To facilitate the distribution of geospatial data online, the DiverSea project considers adopting OGC standards such as Web Map Service (WMS) and Web Feature Service (WFS). These standards are essential for delivering datasets like bathymetric maps, habitat surveys, and species distribution data through interoperable web services. Integration with platforms like EMODnet and SeaDataNet is expected to enhance real-time data accessibility and visualization.

Other standards and platforms have also been reviewed during the development of the DiverSea Data Space. These include EML (Ecological Metadata Language) for ecological metadata, Dublin Core for general-purpose metadata, and WMO Core Metadata Profile for meteorological and oceanographic data. CF Conventions (Climate and Forecast) are also considered in conjunction with NetCDF for climate-related datasets. Each of these standards has been evaluated based on their relevance and utility for specific data types within the DiverSea context. The selection of these standards aims to ensure data quality, integration capability, and long-term preservation. By focusing on widely adopted standards, the DiverSea project seeks to maximize compatibility with global platforms and initiatives, making datasets valuable to a diverse range of users, including researchers, policymakers, and conservationists.

6.2 Compliance with EU data frameworks

The DiverSea project partners consider ensuring compliance with key EU data frameworks to enhance data accessibility, interoperability, and reusability. One of the guiding principles for this compliance is the adoption of the FAIR Data Principles, which aim to make data Findable, Accessible, Interoperable, and Reusable. To achieve findability, datasets may be assigned globally unique and persistent identifiers (DOIs). Accessibility will be promoted by making datasets available through open-access platforms such as EMODnet, GBIF, and SeaDataNet. Interoperability will be ensured by adhering to standards like ISO 19115 Darwin Core, and INSPIRE, which facilitate seamless integration across platforms. Comprehensive metadata and clear licensing policies will support data reusability, enabling datasets to be applied in new research contexts. In addition to aligning with INSPIRE and FAIR principles, the DiverSea Data Space will ensure integration with the Biodiversity Information System for Europe (BISE) and contribute data to the European Nucleotide Archive (ENA) for genetic diversity datasets, as outlined in Task 2.5.

The INSPIRE Directive is being considered to ensure that spatial datasets meet harmonized specifications for metadata, data themes, and services. By aligning datasets related to geospatial mapping, remote sensing, and environmental monitoring with INSPIRE standards, DiverSea aims to support integration with European data infrastructures and facilitate cross-border collaboration.

Integration with major European platforms such as EMODnet for marine and geospatial data, SeaDataNet for oceanographic data, and GBIF and OBIS for biodiversity data is a key goal. These platforms are expected to enhance the usability, accessibility, and long-term preservation of DiverSea datasets. Where applicable, DiverSea will utilize APIs provided by GBIF and other relevant platforms to facilitate automatic data submission and retrieval, ensuring seamless data integration.

6.3 Conclusion of the Mapping Across Case Studies

The mapping exercise across the WP1 case studies indicates that the most relevant standards for the DiverSea project are ISO 19115 for geospatial metadata, Darwin Core for biodiversity data, INSPIRE for spatial data infrastructure, and MEDATLAS for oceanographic data. The most commonly referenced platforms are EMODnet for diverse marine datasets, SeaDataNet for oceanographic data, and GBIF and OBIS for biodiversity data.

To fully adhere to the requirements of Task 2.5, DiverSea considers prioritizing integration not only with EMODnet, SeaDataNet, GBIF, and OBIS but also with platforms such as Copernicus Marine Service, GEOSS, GOOS, and the European Space Agency's Earth Observation Program. This approach will ensure comprehensive data sharing and accessibility.Given these findings, the DiverSea Data Space considers prioritizing these widely adopted standards and platforms for further steps in data integration. This approach is intended to streamline data harmonization, enhance interoperability, and ensure that datasets remain accessible, reusable, and valuable for future research, environmental management, and policy-making efforts.

7 Roadmap for Implementation and Deployment of the Data Space

This deliverable (D2.4) reports on the design of the initial data space of the DiverSea project, which accommodates the raw data produced or already available to the teams working on the WP1 case studies. While it represents the culmination of a year's work and the production of numerous artefacts, , the initial data space is not yet ready for analyzing marine biodiversity and the ecosystem dynamics as needed in the two other technical work packages WP3 and WP4. Similarly, making the data available to the case study teams, enabling its use for citizen science, or supporting the broader research community will occur in later stages, according to the roadmap for implementation and deployment as described below.

Step 1. Validation of the Initial Data Space. Before opening the data space for use within the project, it will be validated with the main stakeholders: the donors (partners working on the WP1 case studies) and the consumers (partners conducting analytical tasks in WP3 and WP4). The next deliverable in this work package (D2.6) will be dedicated to this validation, with a deadline set for month 18, just two months after the submission of this deliverable.

Step 2. Opening the Data Space to partners. Following the validation of the data space design and incorporation of the stakeholder feedback, the initial version of the data space will be finalized and "frozen." It will then be released for immediate use by the project partners. This is anticipated to occur by the end of the first half of 2025.

Step 3. Data Space Revision. Immediately after launching the initial version of the data space, we will start working on transforming the datasets to suit the analytical tasks carried out in

WP3 and WP4, as well as preparing for integration with the European repositories such as **EmodNet** and **Copernicus**. For this purpose, in collaboration with partners involved in WP3 and WP4, we will re-organize and pre-process the raw data collected in the initial data space. In parallel, we will continue adding more raw data to the initial version of the data space as generated within the WP1 case studies. This revision process will extend beyond the project's second year and continue until its completion.

Step 4. Developing Procedures for Cataloguing and Publishing Data. As part of the development of the second version of DiverSea Data Space, two tasks within WP2 – T2.3 and T2.4 – will focus on establishing policies for publication of both raw and pre-processed data by the partners involved in supply and transformation of data which originates in WP1 case studies. The procedures will consider relevant standards and operational protocols in the corresponding data repositories.

Step 5. Completing the Final Version of the Data Space. Deliverable D2.5 is tasked to report on the completion of the final version of DiverSea Data Space. However, due to the continuing generation of raw data in the WP1 case studies and its transformation, the actual completion of the final version of the data will span a significant portion of the second and third years of the project. As a result, the deadline for submitting D2.5 is set for month 32 of the project, nearing the end of the third year.

Step 6. Validation of the Final Data Space. After completing data pre-processing and updating the data models of the initial version, we will conduct extensive validation of the final data space in collaboration with the stakeholders from broader community of potential users as identified in the stakeholder's workshops conducted within WP4. The primary focus will be to engage with national and international stakeholders interested in further analysis of the data published in the European repositories such as **EmodNet** and **Copernicus**. This validation effort forms the basis of Deliverable D2.7, which is scheduled for completion by Month 34, shortly before the end of the third year of the project.

Conclusion

This report marks the culmination of more than a year of challenging work by computer scientists, data analysts and data engineers entering the field of marine biodiversity and coastal seas ecosystems research. Throughout this period, the WP2 team collaborated closely with the teams of WP1, WP3 and WP4 and contributed to the cross-domain study in WP5. This collaborative effort resulted in the development of the initial version of DiverSea Data Space, a technical resource designed to support marine specialists in their effort to provide a comprehensive picture of the state of the marine biodiversity and the coastal environment along the shores of Europe from Black Sea in the east through the Sea of Marmara, Mediterranean Sea and the Atlantic Ocean up north all the way to the North Sea.

During the second and the third years of the project, the work within WP2 will become more technical, focusing on expanding the scope and enhancing the data processing potential of the project. In this phase, WP2 will collaborate more closely with partners from the two other technical work packages, WP3 and WP4. while maintaining its strong working relationships with data providers from the WP1 case studies. By the end of this period, WP2 will also engage with the community of marine researchers and decision makers to support the project's goals.

Based on the work reported here and aligned with the projected roadmap for implementation and deployment, the DiverSea Data Space is expected to be fully completed by the end of the third year. The fourth year of the project will then focus entirely on disseminating the outcomes and exploring the benefits for the local, national and international communities of researchers, decision makers and business operatives.

Annexes

Annex 1. Vocabulary of Terms

The vocabulary of terms is divided into several groups depending on their use for description, management, processing, and use of data in the project

A1.1 General Terms

This section defines the topmost terms used in working with data.

Data: Artefact in digital form, containing information relevant to the marine biodiversity, its ecosystem, and its dynamics; ***Example:*** Posidonia description

Meta-data: Data, providing information about the data from different perspectives – identification, localization, qualification, quantification, etc. ***Example:*** Data source for the data

Data Management Operation: Operations upon the data which do not change its meta-data and information content in any possible way, but its storage location, internal computer representation, or external presentation for the purpose of using it. ***Example:*** mixing the eDNA sequences from two locations.

Data Pre-processing Operation: Operation upon the data which does not change its information content, only its meta-data. ***Example:*** identifying zooplankton within water data samples.

Data Analysis Operation: Operation which produces new meta-data from existing data. ***Example:*** Classification of the plankton data.

Data Post-processing Operation: Operation which creates new data from existing data and changes its meta-data. ***Example:*** Filtering the noise in acoustic data

Data Pipeline: A sequence of operations along the entire life span of particular data, combining data management, data pre-processing, data analysis, and data post-processing operations. ***Example:*** Collecting water samples, filtering the water, sequencing the eDNA, showing the location on the map.

Operational Workflow: Directed graph of operations that may be applied to different data depending on conditions along the paths within the graph. ***Example:*** Collection, Aggregation and Synchronization of data from multiple sensors.

Data Ontology: High-level integral model of the data and its metadata together with the data processing operations, pipelines and workflows. Typically formulated using a logical language, serialized in one of the popular data markup languages (XML, JSON, RDF).

Data Management Policies: Rules for accessing and manipulating the data by different stakeholders in accordance with their permissions and restrictions as specified in their profile (i.e., read-only operations, read/write and destroy data, register datasets and allocate memory for it, persist data and open access points, start/stop data services and share data, register users and grant permissions for working with data, etc.). ***Examples:*** Data Administrator, Data Engineer, Data Analyst, Service Consumer, etc.

Knowledge Graph: External representation of the data ontology and the related policies in a comprehensive format, usually as a directed graph which can be easily visualized.

A1.2 Metadata

This section defines the most important qualitative descriptors of data without considering its internal structure, its meta-data.

Content: Qualitative and quantitative information related to the marine biodiversity and its ecosystem; ***Example:*** Sound of the dolphins and other large sea mammals

Type: Human understandable format of the information according to its content. ***Examples:*** sensor readings, text descriptions, symbolic encodings, sound recordings, diagrams, scans and images, map layers, spatial renderings, etc.

Format: Computer readable format of the information according to its availability in digital form; ***Example:*** plain text, CSV, SQL, HTML, JSON, XML, graphics, audio, video, etc.

Origin: Geographic location where the data comes from. ***Examples:*** Dardanelles Straight.

Source: Technical Infrastructure for obtaining the data regardless its owner. ***Examples:*** Sensor, Scanner, Water sonde, Air balloon, Dron camera, etc.

Date/Period: Calendar time of making the data available; in the case of project data this is the exact time of generating the data, in the case of external data this is the time extracted from its meta-data

(Persistence) Data Model: Data description using a formal language for the purpose of data management. ***Examples:*** free text, list, table, ER model, SQL DDL, JSON, XML, UML.

Creation Mode: Technical way of obtaining the data. ***Examples:*** Observed, Measured, Simulated, External

Transfer Mode: Technical way of transferring the data. ***Examples:*** file upload, repository import, batch transfers, online streaming, etc.

Provider: Organization owning the internal data generated or provided by a project case study

Supplier: Owner of publicly or privately donated external data; ***Examples:*** university, research institute, NGO, government institution, business organization, industrial company, online site, private person

Privacy: Restrictions imposed on data use and data sharing. ***Examples:*** Public, Limited, Licensed, Designated, Private

Standard: Regulatory information.

A1.3 Data

This section defines the structural, qualitative and quantitative descriptors of the data in digital format.

Data Attribute: The smallest amount of information about the data, represented using a single computer datatype. Different terminplogy has been used in different context - *Essential Variable* (domain), *Data Property* (logical modeling), *Field* (physical modeling), *Feature* (analysis), *Parameter* (design), *Value* (implementation). ***Examples:*** Size of the data in integer, Amount of the memory in bytes, Depth of the source in floating number, etc.

Data Record: The smallest structure that groups several data attributes together, characterizing complex self-contained information. ***Examples:*** Fish Net, Fishing Spot, Fish Catch, Fishing Boat, Fisherman, Fishmonger, Etc.

Data Collection: Ordered Set of Data Records. ***Example:*** Table of the different species of fish, List of the fishing boats, etc.

Data File: Ordered set of one or more data collections in a single computer file, ordered sequentially by time, by identifier, or by some other data attribute and available for immediate use. ***Example:*** File with data about the monthly catch in the port of Odesa area over 2024.

Data Stream: Similar to data file, but not available for immediate use; can be obtained only in real-time mode of transfer. ***Example:*** Flowmeter measurements obtained by a flowing water sonde.

Dataset: Computer file, set of computer files or data stream containing data collections in a specific computer-recognized format. ***Examples:*** CSV file, JSON stream, SQL database export, etc.

Identifier: Unique data attribute that allows individual treatment of the data on each level of granularity (record, collection, file, stream, dataset).

Relation: Data attribute which relates the data to another data. Different terminology has been used in different contexts: *Impact* (domain), *Object Property* (logical modeling), *Connection* (physical modeling), *Dependence* (analysis), *Link* (design), *Key* (implementation). ***Example:*** The dataset describing the readings of the physical conditions in the water is related to the dataset describing the area where the measurements are taken from.

A1.4 Data Management Operations

Data Management Operations do not change the information content of the data but may affect its metadata. They are necessary for preparing the data to persist in computer memory and further analyze it. Typically, they are organized in data pipelines which continuously control the flow of data between the source and the destination.

Generation: This operation obtains data values from the data source and creates separate data records depending on the content, type, and format. ***Example:*** Reading the temperature of the water from the water sondes in degrees Celsius.

Aggregation: This operation combines multiple data records into a new data record, based on common data attributes. ***Example:*** Aggregating several biochemical measurements obtained in the same place at the same time from different sensors into a single record.

Collection: Linking several data records in a collection based on a common ordered data attribute. ***Example:*** Creating a file of temperature records over a specific period of time.

Encoding: Changing the value of a data attribute according to a known encoding scheme. ***Example:*** Encoding all fish species using the prefix Zoo and all Posidonia fields with Phyto,

Encrypting: Changing the data according to a known encryption algorithm and scheme on different levels (attribute, record, collection, file, or stream).

Transportation: a separate pipeline for sending, relaying, and receiving data using a communication protocol of cyberspace. ***Example:*** Transporting data about the level of microfibre pollution of the seawater near Izmir from EmodNet over HTTP.

Persisting: Storing data into files, databases or distributed file systems for subsequent use during analysis or in additional data services. ***Example:*** directing the stream of real-time sensor readings of the temperature of the waters off the shore of the Aegean Sea near Athens into a MongoDB database.

Ingestion: a separate pipeline for downloading, buffering and storing data into designated repository. ***Example:*** Ingesting data about the level of jellyfish presence off the shores of Black Sea near Odesa and storing it into a PostgreSQL database.

A1.5 Data Pre-processing Operations

Data Pre-processing Operations change the information content by extending the data with additional data, or by contracting it without losing essential information. In both cases, the meta-data is updated with additional descriptors to accommodate the changes.

Filtering: Removing data from collections, files, or streams based on commonly applied constraints over one or more data attributes. ***Example:*** Removing all records from a file, which are not from the same place.

Completion: Adding data to records, collections, files, or streams in the case of missing values for selected attributes. ***Example:*** If the records taken on an hourly basis do not always contain some data, the missing data is substituted with the average of the existing data from the collection (interpolation).

Normalization: A data transformation technique for scaling the numerical values of the data attributes within a given common range in such a way, that it preserves the relative differences within this range. ***Example:*** replacing the fluctuations of absolute values of the amounts of sulphur dioxide in a cubic meter of water with a percentage of the difference between the minimal and maximal levels.

Smoothening: Replacing the values of selected data attributes in records, collections, files, or streams using the value of a function over the entire collection, file, or stream. There are different methods for data smoothening. ***Example:*** The attribute value of every second record in the collection is replaced by the average value of this attribute for the first and the third record in the same collection (linearization).

Enrichment: Adding new attributes to the records as a function over the entire collection, file, or stream which accounts dome logical dependencies between the attributes. ***Example:*** Identifying the route number by comparing the location and the timing of the record to the locations and the times of the previous and following records in the collection.

Mapping: Replacing attribute values with values of functions of the same type of attribute. ***Example:*** Replacing the absolute time of the day with a relative time (i.e., one hour later than the current record).

Accumulation: Replacing data records in a collection with a single record with attributes, which are the result of cumulative operation over the corresponding attributes within the same collection. ***Example:*** Summing up the time for observation to establish the total duration.

Buffering: Splitting a single data collection, file, or stream into multiple sequentially ordered fragments (collections, files, or streams). The fragments can be of the same or different size, depending on the needs and the constraints. ***Example:*** Splitting the data collections into multiple fragments, containing information for the same day.

Synchronization: Combining filtering, completion, and smoothening of the data records in a collection into a structurally homogenous and descriptively identical collection of data records. ***Example:*** Constructing data collection with daily measurements of the same physical properties of the water.

Indexing: Creating a new order of the data records in a collection based on a natural linear order of some of their attributes. ***Example:*** Order the records of the movements of a ship numerically in an increasing order of the timing and, alphabetically, by the name of the location.

Exploratory Data Analysis (EDA): Statistical data processing technique for estimating minimal, maximal, and average values, variance, correlation and other statistical characteristics of the datasets.

Outlier Detection: Identifying data records that are significantly different from the majority of data and removing them from the datasets.

Principal Component Analysis (PCA): A statistical technique used to reduce the dimensionality of a dataset while retaining most of the variance.

A1.6 Data Analysis and Interpretation

This section contains a description of the most common terminology used in data analysis and machine learning (ML). An in-depth discussion of this vocabulary will be found in WP3, while the terminology introduced here is to provide the context of data use.

(Analytical) Data Model: Formal description of the data in a suitable for data analysis formal language. Note that this data model is different from the data model used in data management

Feature Importance: A score that indicates how useful or valuable each feature was in the construction of the analytical data model.

Feature Selection: The process of identifying and selecting the most relevant data attributes that contribute significantly to the classification, pattern recognition, prediction, trend analysis, and other analytical tasks.

Feature Engineering: Creating new data attributes from existing data to improve the performance of data analysis

Training Dataset: Datasets used to prepare ("train") the analytical algorithms in the case of using supervised machine learning technique for the analysis.

Test Dataset: Datasets used to test the algorithms for analysis of the data. The results of the analysis of the test dataset must be known in advance in order to compare the results obtained by the use of data analytics with the actual known in advance results.

Fitness: A measure of the adequacy of the analytical model based on its performance during data analysis

Overfitting: When a model learns the training data with great sensitivity, including the noise, but performs poorly on new data with different statistic characteristics.

Regression: A family of statistical methods for prediction of temporary data in time series

Correlation: Statistical methods for estimating the degree of mutual dependence between synchronized data

Mean Absolute Error (MAE): The average of the absolute differences between predicted and actual values of the data attributes.

Cross-Validation: Methiod for model evaluation that assesses how the results of a statistical analysis will generalize to an independent dataset.

Root Mean Square Error (RMSE): A standard way to measure the error of a model in predicting quantitative data, computed as the square root of the average squared differences between predicted and observed values.

R-squared (R^2): A statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

A1.7 System Modeling and Factor Analysis

This section contains the most important terminology relevant to the simulation of the ecosystem and analysis of its state and internal dynamics using the methodology of system analysis. A more detailed discussion can be found in WP4 but here it has been introduced to provide context for use of data in system dynamics simulation and factor data analysis.

(System Dynamics) Model: A formal model describing the behavior of a system or a class of systems using equations. The dynamic models represent the time-varying state changes, such as *acceleration*, *velocity*, and *fluctuations* over time.

Model Parametrization: The parametric equation defines a group of parameters as functions of one or more independent variables. They usually depend on time. The independent parameters are called *free parameters* or *drivers*. Their role is to provide the input data which determines the potential dependency between parameters

Parameters: Data attributes that have a *qualitative impact* on the system dynamics. Their role is to provide the input data which determines the potential dependencies. The parameters usually depend on time, and their calculation is a separate dynamic process, the independent parameters of which are called *free parameters* or *drivers*.

Factors: Free parameters that have a *qualitative impact* on the system. Depending on their impact, they are called *influencers, stressors, stabilizers, mitigators*, etc.

- ✓ *Influencers* - factors which cause change in the parameters of the system when changing over time
- ✓ *Stressors* - factors which cause stress on the dependent parameter (s) in some time points (i.e., they cause oscillations of the values of dependent parameters, increase the fluctuations, increase the acceleration of changes, increase the frequency of changes, etc.).
- ✓ *Stabilisers* - factors which cause specific changes of the values of dependent parameter (s) when increased or decreased over time
- ✓ *Regularizer* - factors which cause smooth changes of the dependent parameters when increased/decreased
- ✓ *Mitigators* - factors which cause specific non-correlated changes of the frequency of changes of the dependent parameters

Parameter Variation: Change of the *value*, *relative distance* to another value, *frequency* of oscillation around a given value, *speed* of increasing the distance, and *acceleration* of the speed. Measured qualitatively

Dependent parameters: parameters that are impacted by the free parameters.

Space dimension: The space dimension of a mathematical or physical space (or object) is defined as the minimum number of coordinates needed to specify any point within it.

Parametric Equation: A parametric equation defines a group of data attributes as functions of one or more independent parameters. Parametric equations are commonly used to express the coordinates of the points that make up a geometric object, such as a curve or surface, which are called a *parametric curve* and a *parametric surface*, respectively.

Functional Dependence: A function that *explains the changes* in a group of data attributes by the changes in another group. Measured qualitatively as linear, stepwise, exponential, logarithmic, periodic, sigmoid, normal (Gaussian), etc.

A1.8 Software

This section contains brief description of the most typical software used in projects which include data processing

Software System: Software which can be executed on a hardware infrastructure and used by end users for performing specific data processing. ***Example:*** Mobile Phone Operating System.

Server: Software application extending the capability to process local data on remote infrastructure, which works continuously. ***Examples:*** Database Management Systems (DBMS), Web servers, File servers, Application servers, etc.

Client: Software application used locally by the end users for performing data processing remotely on servers. ***Example:*** Web browsers, Dashboards and Server Consolas.

Service: Software component which executes operations under the control of a server and can be used from a suitable client. ***Example:*** Directory service, Authentication service, Location service, Search engine, etc.

Registry: Server which registers different resources. ***Examples:*** hardware catalogue, software catalogue, file catalogue, user directory, service directory, etc.

Protocol: A set of commands of a computer language (programming language, data markup language, communication language, etc.), together with a description of their parameters. ***Examples:*** HTTP (for browsing catalogues), SOAP (for using Web Services), MQTT (for transporting sensor readings), SPARQL (for querying ontologies), etc.

API (Application Programming Interface): Abstraction of a layer of the software which translates protocol commands into executable operations. ***Example:*** Eclipse Mosquitto

Driver: Software component which facilitates the interaction between different software components within the software systems

Annex 2. Data Ontology

See the files biodiversityOntology.rdf and biodiversityOntology.owl of the data ontology in the file repository of the project at https://drive.gate-ai.eu/WP2/T2.2/

Annex 3. Data Repository

See the database scripts databaseScripts.sql for creating PostgreSQL databases in SQL DDL format in the file repository: https://drive.gate-ai.eu/WP2/T2.2/

Annex 4. Datasets with Historical and Public Data available to Project Partners

See the file Historic and Public Data for the Case Studies.pdf with description of the available datasets with historic and public data in in the file repository of the project at: https://drive.gate-ai.eu/WP2/T2.1/

Annex 5. Datasets produced within DiverSea Case Studies

See the files with descriptions of the DiverSea Case Study datasets in XSLSX format in the correspondiong sections of the file repository of the project at: https://drive.gate-ai.eu/WP2/T2.1/

GLOSSARY, ACRONYMS AND ABBREVIATIONS

Short Name	Legal Name	Country
NTNU	NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET	NO
UCPH	KOBENHAVNS UNIVERSITET	DK
VITO	VLAAMSE INSTELLING VOOR TECHNOLOGISCH ONDERZOEK N.V.	BE
UNIVPM	UNIVERSITA POLITECNICA DELLE MARCHE	IT
LPN	Liga para a Protecção da Natureza	PT
UPF	UNIVERSIDAD POMPEU FABRA	ES
AT	ASOCIACION TURSIOPS	ES
DEU	DOKUZ EYLUL UNIVERSITY OF IZMIR	TR
IMB	INSTITUTE OF MARINE BIOLOGY OF THE NAS OF UKRAINE	UA
UPORTO	UNIVERSIDADE DO PORTO	PT
GATE-SU	SOFIA UNIVERSITY ST KLIMENT OHRIDSKI	BG
AUEB	ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS - RESEARCH CENTER	EL
UGOT	GOETEBORGS UNIVERSITET	SE
VNNOR	VNNOR AS	NO
NKUA	ETHNIKO KAI KAPODISTRIAKO PANEPISTIMIO ATHINON	EL
AKD	AKDENIZ KORUMA DERNEGI	TR
UG	UNIVERSITY OF GHANA	GH
INEGI	INEGI - INSTITUTO DE CIENCIA E INOVACAO EM ENGENHARIA MECANICA E ENGENHARIA INDUSTRIAL	PT
UCSD	The Regents of the Univ. of Calif., U.C. San Diego	US

Acronyms - Countries			
BE	Belgium	NO	Norway
BG	Bulgaria	PT	Portugal
DK	Denmark	SE	Sweden
EL	Greece	TR	Turkey
ES	Spain	UA	Ukraine
GH	Ghana	US	United States
IT	Italy		

DiverSea Project Terminology Abbreviation List

Abbreviation	**Meaning**
AI-ML	Artificial Intelligence & Machine Learning
API	Application Programming Interface
BISE	Biodiversity Information System for Europe
BHD	Birds and Habitats Directive
BSTA	Beyond State of The Art
CAPED	Create Assess Protect Exploit and Disseminate
CF Conventions	Climate and Forecast Conventions
CDOM	Chromophoric/Colored Dissolved Organic Matter
CMEMS	Copernicus Marine Environment Monitoring Service
Copernicus	European Union's Earth observation program
CSV	Comma-Separated Values
CTD	Conductivity, Temperature, Depth
Darwin Core	Biodiversity Data Standard
DBMS	Database Management System
DDL	Data Definition Language
DNA & RNA	Deoxy Ribonucleic Acid & Ribonucleic Acid
DOP & DEP	Dissemination and Outreach Plan & Dissemination and Exploitation Plan
EC centre for biodiversity	European Commission Knowledge Centre for Biodiversity
eDNA/eRNA	Environmental DNA/RNA
EGS	Ecosystem Goods & Services
ELTER	Integrated European Long-Term Ecosystem, critical zone and socio-ecological Research
EMODnet	European Marine Observation and Data Network
EMSO	European Multidisciplinary Seafloor and Water Column Observatory
ER	Entity-Relationship
ERIC	European Research Infrastructure Consortium
EOV/EBVs	Essential Ocean and Biodiversity Variables
ESA	European Space Agency
GBIF	Global Biodiversity Information Facility
GEOBON	The Group on Earth Observations Biodiversity Observation Network
GEOSS	Group on Earth Observations
GOOS	Global Ocean Observing System
GUI	Graphical User Interface
HTTP	Hypertext Transfer Protocol
ICES	International Council for the Exploration of the Sea
INSPIRE	Infrastructure for Spatial Information in the European Community
ISO	International Organization for Standardization

IoT	Internet of Things
IUCN	International Union for Conservation of Nature
JSON	JavaScript Object Notation
KERs	Key Exploitable Results
KPI	Key Performance Indicators
LSTS	Laboratório de Sistemas e Tecnologia Subaquática
MARBIOSE	European MARine BIOdiversity and ecosystem goods & Services
MEDATLAS	Mediterranean Data Atlas
MOU	Memorandum of Understanding
MPA	Marine Protected Area
MSFD	Marine Strategy Framework Directive
MQTT	Message Queuing Telemetry Transport
NCA	New Concepts and Approaches
NetCDF	Network Common Data Form
NGO & SME	Non-Governmental Organization & Small and Medium-sized Enterprises
O/Obj.	Objectives
OBIS	Ocean Biodiversity Information System
OGC	Open Geospatial Consortium
OMoMaP	Observation, Monitoring, Mapping and Prediction
R&D	Research and Development
RDF	Resource Description Framework
S & SS	Scope & Specific Scope
SEWP	Spreading Excellence and Widening Participation
SOAP	Simple Object Access Protocol
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
SSH	Social Sciences and Humanities
STEM	Science, Technology, Engineering, and Mathematics
SWAFS	Science with and for Society
TRL	Technology Readiness Level
UML	Unified Modeling Language
VPN	Virtual Private Network
WFD	The Water Framework Directive
WFS	Web Feature Service
WMS	Web Map Service
WP	Work Package
XML	Extensible Markup Language