European Union Network for the Implementation and Enforcement of Environmental Law

Monitoring large marine vertebrates along fixed transects from ferries and cargo vessels: data validation, format and storage & database interoperability

Date of report: January 2023

Report number: 2022(VI) WG4

Funded by the European Union

IMPEL is funded by a "FRAMEWORK PARTNERSHIP AGREEMENT" with European Commission DIRECTORATE-GENERAL FOR ENVIRONMENT - LIFE PROGRAMME (ENV.E.4/FPA/2022/001 – IMPEL)

Introduction to IMPEL

The European Union Network for the Implementation and Enforcement of Environmental Law (IMPEL) is an international non-profit association of the environmental authorities of the EU Member States, acceding and candidate countries of the European Union and EEA countries. The association is registered in Belgium and its legal seat is in Brussels, Belgium.

IMPEL was set up in 1992 as an informal Network of European regulators and authorities concerned with the implementation and enforcement of environmental law. The Network's objective is to create the necessary impetus in the European Community to make progress on ensuring a more effective application of environmental legislation. The core of the IMPEL activities concerns awareness raising, capacity building and exchange of information and experiences on implementation, enforcement and international enforcement collaboration as well as promoting and supporting the practicability and enforceability of European environmental legislation.

During the previous years IMPEL has developed into a considerable, widely known organisation, being mentioned in a number of EU legislative and policy documents, e.g., the 7th Environment Action Programme and the Recommendation on Minimum Criteria for Environmental Inspections, and more recently in the General Union Environment Action Programme to 2030 and EU Action Plan: 'Towards Zero Pollution for Air, Water and Soil'.

The expertise and experience of the participants within IMPEL make the network uniquely qualified to work on both technical and regulatory aspects of EU environmental legislation.

Information on the IMPEL Network is also available through its website at: www.impel.eu

Title of the report:	**Number report:**
Monitoring large marine vertebrates along fixed transects from ferries and cargo vessels (Vol. 2): data validation, format and storage & database interoperability	2022(VI) WG4
Project Manager: Roberto Crosti (IT) Authors: Arcangeli A, Babey L, Correia AM, Crosti C, David L, Falconi M, Garcia-Garin O, Mifsud C, Pasanisi E, Tepsich P, Scuderi A, Vighi M, Campana I. Editors: Ilaria Campana, Morgana Vighi, Alessia Scuderi. Citation: Arcangeli A, Babey L, Correia AM, Crosti C, David L, Falconi M, Garcia-Garin O, Mifsud C, Pasanisi E, Tepsich P, Scuderi A, Vighi M, Campana I (2023) *Monitoring large marine vertebrates along fixed transects from ferries and cargo vessels (Vol. 2): data validation, format and storage & database interoperability* IMPEL report N.: 2022(VI) WG4-p 82.	**Report adopted at IMPEL General Assembly Meeting**: 08/06/2023 Stockholm **Total number of pages: 82**

Executive Summary

Marine data are expensive to collect and wide benefits can be gained from sharing and properly managing them, for scientific and management purposes.

The aim of creating common databases follows the FAIR Data Principles, which have been produced to make biodiversity data Findable, Accessible, Interoperable and Reusable.

Interoperable data can be integrated with other data, applications and workflows. This is achieved by using common metadata and data standards, and by harmonising data by using semantic artefacts.

Interoperability is strongly encouraged at EU level through the INSPIRE Directive, which implies that spatial data and services established and operated by the Member States of the EU can be combined in a consistent manner, and that datasets can be accessed via network service

Despite these requirements, interoperability and semantics are still unfamiliar to many data collectors. It has been highlighted the need to increase knowledge on existing standards for harmonisation and data interoperability, and the need for review, guidance, and validation protocols of the many existing metadata standards.

Within this context, the EU MTT project provides a good example of a network of research institutes and non-profit organisations performing marine monitoring, working to overcome the differences in the protocols and data formats, to harmonise their data and obtain interoperable and high-quality information. Indeed, many partners already contribute to fit their data to the legal requirements of

the EU biodiversity monitoring.

In this report, the main existing data repositories related to biodiversity/marine species are described, providing information on their purposes, functioning and outputs. The specific requirements of each repository, standards used, and information on how it is possible to contribute to them are also reported in detail.

The outcome of this report is to highlight the main characteristics of these databases, their similarities and connections, and their links with the international legislative framework, with the aim to provide a useful tool for the EU MTT partners to improve the integration of the data collected within their monitoring activities in these common repositories.

TABLE OF CONTENTS

1. The IMPEL EU MTT project (2022-2024), the framework

IMPEL, the European Union Network for the Implementation and Enforcement of Environmental Law, is an international non-profit association of the environmental authorities of the European Union Member States, acceding and candidate countries of the EU, EEA and EFTA countries. The IMPEL mission is to ensure effective implementation and enforcement of European environmental law by promoting professional collaboration, information, and best-practice exchange between environmental regulators.

Within this framework, IMPEL has been co-financing since 2020 a partnership of European bodies with the aim of publishing shared guidelines for monitoring marine species included in the Habitats Directive (HD), such as cetaceans and sea turtles, in high sea areas, using large vessels traveling along Fixed Line Transects, as platforms of observation.

Several research institutions and non-profit organisations have been monitoring for many years marine megafauna such as cetaceans, sea turtles, and sea birds, and their threats, such as marine litter, and marine traffic, using large vessels, including cargoes and ferries, as platforms for systematic surveys. The main networks involved in this activity are the FLT MED Network, led by ISPRA, the "North-East Atlantic" Network, led by ORCA (which publishes yearly the "State of European Cetaceans"), and the "Macaronesia sea" network CETUS Project, led by CIIMAR. The networks are expanding also across the southern countries of the Mediterranean Region.

The need to harmonise the data stewardship and all the data processes in order to fulfil the requirements of EU environmental legislation (e.g., Marine Strategy Framework Directive (MSFD), HD, Bird Directive (BD), New Waste Directive Package, Single Use Plastics Directive, the Barcelona Convention, and OSPAR) led to a "road map" with several yearly milestones.

In 2020, within the IMPEL FLT Europe project, the state of the art regarding the monitoring of large marine vertebrates along fixed transects from large vessels, was assessed through a SWOT analysis (IMPEL 2021, Campana and Vighi, 2020).

In 2021, within the IMPEL EU MTT project (2022-2024), a common recording tool was developed (AtSea on ODK open-source mobile data collection platform).

The first task foreseen for 2022 was to describe the data validation, format, and storage processes and the database interoperability of the collected datasets (this internal report).

For the upcoming years, the project members will deal with the issue of data analysis in order to respond to the legislation needs (2023). The final Guidelines will be published in 2024.

2. The main principles and issues in data sharing and interoperability

Marine data are expensive to collect and always unique in relation to time and geographical position. Wide benefits can be gained from working together to share and properly manage the data, defining access to users at different levels, from descriptive information, analytical procedures, raw data, and research outputs (Figure 1). The aim of creating common databases follows the **FAIR** Data Principles, which have been produced to make biodiversity data Findable, Accessible, Interoperable and Reusable (Wilkinson et al, 2016).

To be **Findable**:
- (meta)data are assigned a globally unique and eternally persistent identifier, described with rich metadata that specify the data identifier and are registered or indexed in a searchable resource.

To be **Accessible**:
- (meta)data are retrievable by their identifier using a standardised communication protocol, that is open, free, and universally implementable and allows for an authentication and authorisation procedure, where necessary. Metadata are accessible, even when the data are no longer available.

To be **Interoperable**:
- (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation, use vocabularies that follow FAIR principles and include qualified references to other (meta)data.

To be **Reusable**:
- (meta)data have a plurality of accurate and relevant attributes, are released with a clear and accessible data usage licence, are associated with their provenance and meet domain-relevant community standards.

FAIR principles can, and should, be applied to all research output(s). Interoperable data can be integrated with other data, applications and workflows. This is achieved by using common metadata and data standards, and by harmonising data by using semantic artefacts to describe their variables unambiguously. Likewise, technical interoperability is achieved by creating automated workflows using standards and Application Programming Interfaces (APIs) for data transfer. It is however important to always keep raw data and harmonise only for interpretations and reporting.

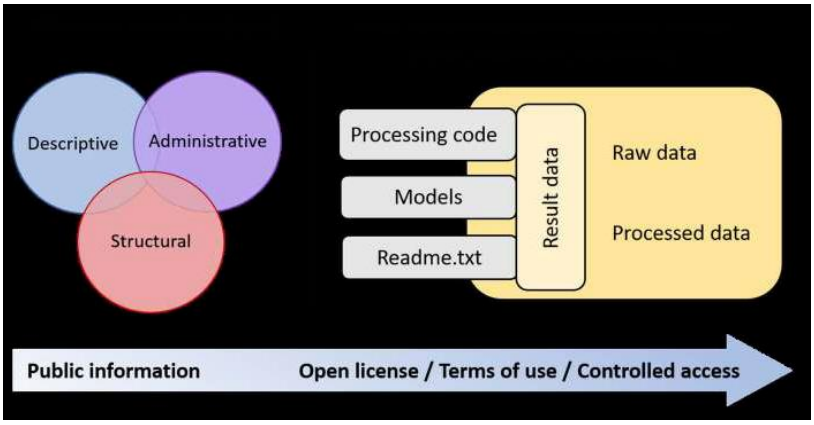


Figure 1 Diagram of the data sharing process (Source: www.biodiversa.eu/2022/11/05/workshop-on-data-interoperability-and-harmonisation/).

Metadata is used for describing the fitness-for-purpose, but not so much for data discovery. Community standards and their extensions are needed for making data available in an understandable (FAIR) format and for data Quality Assurance, depending on the purpose of data use. Vocabularies are used with standards to express the data quality as richly as possible.

Interoperability and harmonisation of biodiversity monitoring databases are among the topics considered within the Biodiversa+ work programme. Biodiversa+ is the Partnership jointly developed by the European Biodiversity Partnership and the European Commission, as part of the EU Biodiversity Strategy 2030 (https://www.biodiversa.eu/) aimed to connect science, policy and practise for transformative change, ensuring efficient science-based support for policy-making and implementation in Europe. Within this context, two workshops were organised in 2022, targeting European and global databases and initiatives, as opportunities to introduce their data architecture and data workflows (https://www.biodiversa.eu/2022/11/05/workshop-on-data-interoperability-and-harmonisation/). Three main challenges for data interoperability and harmonisation emerged during the workshops, which are briefly described below.

Lack of (meta)standards

Data collected by different agencies within governments apply different standards, and there is a continuous proliferation of metadata standards. The review of such standards, guidance, and validation protocols are needed to ensure the adherence, quality, and consistency between information.

Data and methods should be harmonised across the heterogeneous research landscape and linked with national, European and global institutions, fostering data interoperability and semantics to allow the interdisciplinary use of data to exchange information. This also means identifying the minimum metadata sets of information that can be translated in all the major

existing metadata schemes and defining strategies for data management at different stages of the cycle (collect, analyse, publish).
Finally, the integration of machine learning with human interpreted results has no standards and machine learning processes are not always transparently documented. Building machine-readable repositories requires initial human investigation and sense-making. Additionally, these types of repositories have large data storage needs and the current IT infrastructures are not supporting them entirely.

Lack of capacity building and knowledge sharing opportunities
Concepts such as interoperability and semantics are still unfamiliar to many data collectors. There is no clear mandate from funding agencies to adhere to specific standards that facilitate data interoperability, and standardising and sharing data have non-negligible costs. The understanding of these concepts needs to improve, and knowledge on existing tools/standards for harmonisation and data interoperability need to be provided to key actors and stakeholders.
The development of semantic artefact requires people who know multiple institution data and metadata architecture. Best practices and the establishment of common guidelines are needed to facilitate data interoperability between monitoring data from management and research infrastructures.

The governance for biodiversity monitoring data interoperability
Much existing data is collected in templates that lose raw data. Under a management perspective, it is important to assess which dataflows are and/or can be channelled into the official reporting mechanisms.
There is a general lack of harmonisation on what needs to be monitored and no clear roles and goals for the different organisations. There is no one-stop shop for collecting the relevant information needed in terms of monitoring initiatives and FAIR data (requiring to apply standards, providing a central register). It is important to facilitate the integration across sectors (forest, agriculture, water) and the interactions between existing organisations to create co-benefits and improve cost effectiveness, also in line with global and EU policy targets.

3. The EU requirements about data sharing and interoperability

3.1 The INSPIRE Directive

inspire-info@jrc.ec.europa.eu

The Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) was published in the official Journal on April 25th, 2007 and entered into force on May 15th, 2007.
INSPIRE is based on the infrastructures for spatial information established and operated by the Member States of the EU. This makes INSPIRE a unique example of a legislative "regional" approach.
The Directive addresses 34 **spatial data themes**, from protected sites to industrial facilities, needed for environmental applications, with key components specified through technical Implementing Rules (IR). To ensure that the spatial data infrastructures of the Member States are compatible and usable in a Community and transboundary context, the Directive requires that common IR are adopted in a number of specific areas (i.e., Metadata, Data Specifications, Network Services, Data and Service Sharing, Spatial Data Services, Monitoring, and Reporting). These IRs are adopted as Commission Decisions or Regulations, and are binding in their entirety. The Commission is assisted in the process of adopting such rules by a regulatory committee composed of representatives of the Member States and chaired by a representative of the Commission (this is known as the "Comitology procedure"). This open and participatory approach was successfully used during the development of the data specifications as well as during the preparation of the Implementing Rule on Interoperability of Spatial Data Sets. Article 21 defines the basic principles for Monitoring and Reporting.

Interoperability in INSPIRE means that spatial data and services from different sources can be combined in a consistent manner across the EU. The process does not require the involvement of humans or machines. Interoperability implies that spatial data sets can be accessed via network services, usually over the Internet. It is possible to achieve interoperability by changing, i.e., harmonising, and storing existing datasets, or transforming them via services for publication on the INSPIRE infrastructure.
The best practices concerning spatial data are compiled based on evidence of real-world application in production environments, where spatial data are delivered on the Web for end-users, and with an expected high-quality level.
The key aspects of the Best Practices are: data available on the Web, harmonisation of data discovery, data access based on current Web practices, and data connected with other data on the Web.

The Best Practices to be implemented on the INSPIRE infrastructure (Figure 2) are:
1: Use globally unique persistent HTTP URIs for Spatial Things
2: Make your spatial data indexable by search engines
3: Link resources together to create the Web of data
4: Use spatial data encodings that match your target audience
5: Provide geometries on the Web in a usable way
6: Provide geometries at the right level of accuracy, precision, and size
7: Choose coordinate reference systems to suit your user's applications
8: State how coordinate values are encoded
9: Describe relative positioning
10: Use appropriate relation types to link Spatial Things
11: Provide information on the changing nature of spatial things
12: Expose spatial data through 'convenience APIs'
13: Include spatial metadata in dataset metadata
14: Describe the positional accuracy of spatial data

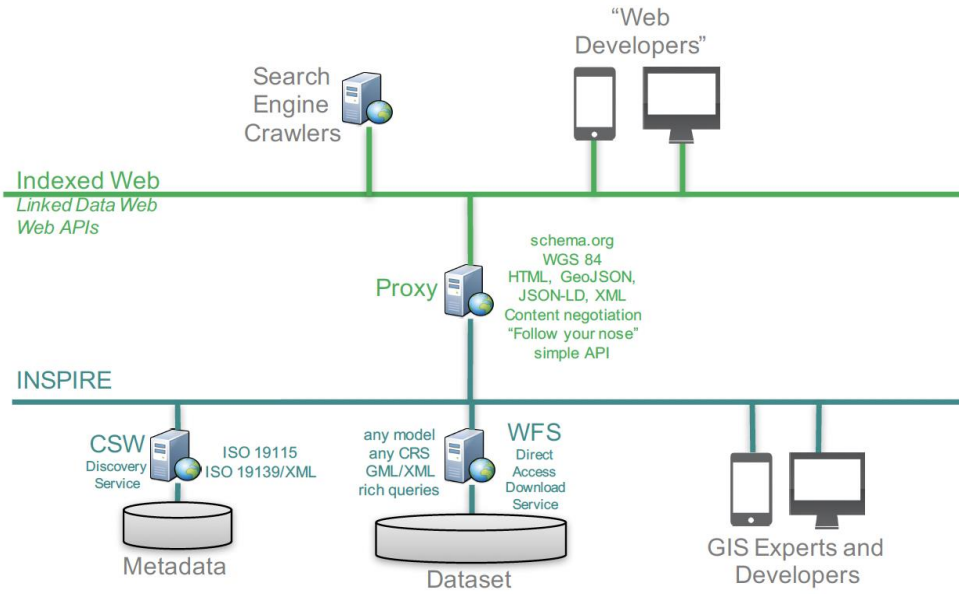


Figure 2 Diagram of the INSPIRE infrastructure
(Source: https://inspire.ec.europa.eu/events/webinar-spatial-data-web-and-inspire).

3.1.1 Theme Species Distribution – Data specification

Species Distribution is a biodiversity theme focused on the geographical distribution of the occurrence of biological organisms. The INSPIRE Directive defines Species Distribution as the geographical distribution of occurrence of animal and plant species (considered as a synonym of the correct scientific term of "taxon") aggregated by the grid, region, administrative unit, or

another analytical unit. There is an aggregation of data, converting raw observations into a distribution of occurrence, where occurrence is the spatial representation of a species at a specific location and a specific time period. The distributions may be represented in a wide range of formats, such as points, grid cells at different scales or polygons of specifically defined areas (analytical units).

The theme *Species Distribution* includes three sections: the Data Set description, the Distribution Information description and the Source Information description (in the extended part).
Each unit specifies a referenceSpeciesScheme, which refers to a choice of three widely known reference lists, and a referenceSpeciesID refers to an ID from that reference list for the given species of interest. EU-Nomen is the preferred reference list to be used. If a taxon is listed in EU-Nomen, this reference must be used as the first choice. If it is not listed in EU-Nomen, the second choice is EUNIS, if not in EUNIS, Natura2000 can be used.

An extended schema allows associating metadata to each unit via the featureType SourceInformation. A multitude of approaches and methodologies exist both for collecting data on species observations and for actually deriving the species distribution from these. In order to ascertain whether a distribution for a given species from a given country is directly comparable with a distribution for the same species from a different country, it is necessary to know the details of the methodologies used. It is important, therefore, that this information is adequately described in the associated metadata. SourceInformation is the feature-level metadata allowing the description of methodology information about each specific instance of distribution information. These metadata can be shared among several species distributions, but when downloaded by a user, they appear as part of the GML dataset rather than with the dataset-level metadata in the associated XML. The extended schema also gives the possibility to link to the observation data specified within the Environmental monitoring facilities specification (Annex III: EF). In addition, it includes a DarwinCoreTriple attribute that allows a connection to the original observational data that can be accessed from GBIF data providers.

4. The most relevant data repositories at global, European and national level

In the following section, the main existing data repositories related to biodiversity, or specifically to marine species, are listed, sorted according to their geographical scope (i.e., global, EU, regional or national data). The website and contact person(s) are indicated for each repository, along with details on their creation and development, the main topics covered, the data contributors, and the outputs/products. The specific requirements of each repository, standards used, and information on how it is possible to contribute to them are instead detailed in the final part of the report (Section 6).

4.1 Global

4.1.1 Ocean Biodiversity Information System

OBIS https://obis.org/
Contact person(s): Project Manager Ward APPELTANS (UNESCO-IODE) w.appeltans@unesco.org
OBIS nodes & node managers:
OBIS SEAMAP node manager Patrick Neal HALPIN (Duke University)
OBIS UK node manager Dan LEAR (MBA UK)
EUROBIS node manager Leen VANDEPITTE (VLIZ)
MEDOBIS node manager Christos ARVANITIDIS, Dimitra MAVRAKI (HCMR)

OBIS emanates from the Census of Marine Life (2000-2010) and was adopted as a project under the programme "International Oceanographic Data and Information Exchange" (IODE) of the "Intergovernmental Oceanographic Commission" (IOC) of UNESCO (established in 1961). OBIS purpose is to enhance marine research, exploitation and development, by facilitating the exchange of oceanographic data and information between participating Member States, and by meeting the needs of users for data and information products.

OBIS accepts data from any organisation, consortium, project, or individual, who wants to contribute. At the date of the preparation of this report, more than 20 OBIS nodes around the world connect 500 institutions from 56 countries. Collectively, they have provided over 45 million observations of nearly 120,000 marine species, from bacteria to whales, from the surface to ocean depths, and from the Tropics to the Poles. Along with the spatial database, OBIS provides useful mapping and graphical tools to explore species occurrences across the globe.

Among its objectives, OBIS provides:

- a global platform for international collaboration between national and regional marine biodiversity and ecosystem monitoring programmes, contributing to a concerted global approach to marine biodiversity and ecosystem monitoring, through guidelines on standards and best practices;
- the world's largest scientific knowledge base on the diversity, distribution and abundance of all marine organisms in an integrated and standardised format (as a contribution to Aichi biodiversity target 19), as well as the integration of biogeographic information with physical and chemical environmental data, to facilitate climate change studies;
- data, information, and tools to support the identification of biologically important marine and coastal habitats for the development of marine spatial plans and other area-based management plans (e.g., Ecologically or Biologically Significant Marine Areas, EBSAs, under the Convention of Biological Diversity), and the assessment of the state of marine biological diversity to better inform policy makers, and respond to the needs of regional and global processes (e.g., UN World Ocean Assessment, Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES)).

4.1.2 BioEco Metadata Portal - Global Ocean Observing System

GOOS https://bioeco.goosocean.org/#/about
Contact person: Serita VAN DER WAL (OBIS) s.van-der-wal@unesco.org

The GOOS BioEco portal is an open access, online platform that provides information on the sustained ocean observing programs that deliver information about the health of marine life. It provides metadata and information from all types of global ocean observations and monitoring programs of biological and ecosystem (BioEco) Essential Ocean Variables (EOV's), namely: biomass and diversity of Microbes, Phytoplankton, Zooplankton, and Benthic invertebrates; abundance and distribution of Fishes, Marine birds, Marine mammals, and Marine turtles; cover and composition of Macroalgae, Mangroves, Hard corals, and Seagrass. The portal is developed in collaboration with the OBIS, IOC Project Office for International Oceanographic Data and Information Exchange, the GOOS BioEco Panel, and the data centre of the Vlaams Instituut voor de Zee (VLIZ). It serves as an online tool that the ocean observing community can use to obtain information on marine biological and ecosystem observations and the programmes that collect them. The portal provides an interactive map that delivers a global picture of the biological and ecosystem observations collected by the contributing programmes, which can be selected or filtered by the programme name. The information about each programme includes the variables observed, the state of development of the programme, the standardisations and specifications used to collect observations, and the programme's observing capability (or readiness level). This information is known as the programme 'metadata'. The actual data collected by each

programme can be found in data systems such as the OBIS, and links to them can be added in the portal.

One of the goals of the BioEco Portal is to create an automated flow of data and metadata from ocean observing programmes to portal and data management systems, such as OBIS. By 2025, it aims for 90% of active BioEco monitoring programmes having up-to-date entries in the BioEco portal, and 80% having established continuous data flow to OBIS. By 2025, the BioEco Portal will also have a live connection with the GOOS monitoring facility, hosted by OceanOPS in Brest.

4.1.3 Group on Earth Observations System of Systems

GEOSS https://www.earthobservations.org/geoss.php
Contact person: Paola DE SALVO pdesalvo@geosec.org; secretariat@geosec.org

GEO is a partnership of more than 100 national governments and 100 Participating Organisations joining government institutions, academic and research institutions, data providers, businesses, engineers, scientists and experts that collaborated to create a Global Earth Observation System of Systems (GEOSS), aimed to better integrate observing systems and share data by connecting existing infrastructures through common standards. EOSS is a set of coordinated, independent Earth observations, information and processing systems that interact and provide access to diverse information for a broad range of users in both public and private sectors. At the date of the preparation of this report, GEOSS Platform brokered more than 150 autonomous data catalogues and information systems, including, among others, those from EEA, GBIF, NASA, NOAA, UNEP. GEOSS links these systems to strengthen the monitoring of the Earth state, and facilitates the sharing of environmental data and information collected by the observing systems of GEO parties. GEOSS ensures that these data are accessible, of identified quality and provenance, and interoperable, to support the development of tools and the delivery of information services. Through the GEOSS platform, this 'system of systems', proactively links together existing and planned observing systems around the world, supports the development of new systems where gaps exist, and promotes common technical standards to allow the combination of data into coherent data sets.

The GEOSS Portal offers a single Internet access point for users seeking data, imagery and analytical software packages, connects users to existing databases and portals, and provides reliable, up-to-date and user-friendly information. For users with limited or no access to the Internet, similar information is available via the 'GEONETCast' network of telecommunication satellites.

4.1.4 Global Biodiversity Observation Network

GEO BON https://portal.geobon.org/home
Contacts: ebvportal@idiv.de, info@geobon.org, info@europabon.org

EuropaBON https://europabon.org/ Henrique PEREIRA (iDiv)
MBON https://geobon.org/bons/thematic-bon/mbon/ Isabel SOUSA-PINTO (UPorto)

The Global Biodiversity Observation Network (GEO BON) contributes to improve the acquisition, coordination and delivery of biodiversity observations and related services to users, including decision makers and the scientific community, with the aim to support effective management policies for the world's biodiversity and ecosystem services. GEO BON is the part of GEOSS representing biodiversity, and is building up for the pathway to link biodiversity data and metadata to GEOSS. The GEO BON network is structured as Working Groups as well as national, regional and thematic Biodiversity Observation Networks (BONs).
GEO BON's most important users are the national governments who are responsible for reporting on the status and trends in ecosystems and biodiversity, in order to meet their national mandates (e.g., national biodiversity plans, recovering species at risk) and international obligations (Conventions, which are GEO BON partners).
GEO BON is focusing its efforts on the implementation and adoption of the Essential Biodiversity Variables (EBVs) and related monitoring guidelines and interoperable data management systems. There are four classes of EBVs: Community composition, Ecosystem functioning, Ecosystem structure, Species populations. Each one includes different types of variables.
The EBV Data Portal is an eShape initiative developed by the German Centre for Integrative Biodiversity Research (iDiv). It includes 27 EBV raster datasets to import and share.

At the date of the preparation of this report, the European BON (**EuropaBON**) counts 1184 Members of 64 countries, and aims at integrating data streams to support EU environmental policy. The EuropaBON Monitoring database (https://monitoring.europabon.org/) explicitly describes current workflows of monitoring networks efforts delivering biodiversity information in Europe. This database maps how biodiversity data collected in monitoring schemes across Europe flows through different institutions and programs and gets processed to produce EBVs and other EU policy-relevant indicators. The portal serves the dual purpose of being the platform for data entry, as well as to allow the visualisation and quick consultation of the collected data. It gathers three types of information for each monitoring network:

- Integration nodes or institutions/projects/initiatives who integrate/process biodiversity data to generate EBVs or other potentially relevant indicators for the EU environmental policy;

- details of the Biodiversity monitoring data or initiatives/schemes responsible for the collection of biodiversity-related information;
- Data streams representing data flows and connections of integration nodes across different scales (e.g., national and European). These streams contain two types of information: information related to the data flowing between nodes (dataset) and whether this has been integrated ("products": EBVs, indicators) or not (raw and aggregated data) (data process). Data streams allow the evaluation of the data flows between monitoring programs and institutions and what kind of integrated data are generated.

The Atlantic International Research (AIR) Centre hosts the Secretariat of the **Marine BON** (MBON) and coordinates the *Marine Biodiversity Monitoring in Europe* study (MarBioME, https://www.aircentre.org/projects/marbiome/, contact person - joana.soares@aircentre.org). MarBioME aims to provide a review and appraisal of the current status of the Marine Biodiversity Monitoring in the EU and adjacent waters (including actors, roles, infrastructures and methodologies) implemented in the framework of the MSFD and other relevant EU legislation. All the information collected in the review will be integrated in the EuropaBON biodiversity monitoring system. Among its aims, the MBON promotes best practice in data management, including development of standards that aid interoperability and data integration, and publication of data through OBIS and GOOS.

4.1.5 Global Biodiversity Information Facility

GBIF https://www.gbif.org
Contact person: Kyle COPAS kcopas@gbif.org

The Global Biodiversity Information Facility (GBIF) is an international network and data infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth. Strictly speaking, GBIF is not a data repository. Rather, GBIF indexes thousands of datasets through a distributed infrastructure involving more than 100 formal participants that supports hundreds of data-publishing institutions worldwide.
GBIF.org makes the FAIR and open data discoverable and citable, assigning each download a Digital Object Identifier (DOI) and storing it for an extended period of time. At the date of the writing of this report over three peer-reviewed research articles were making use, daily, of data from the GBIF network, in studies spanning the impacts of climate change, the spread of pests and diseases, priority areas for conservation and food security, and the GBIF network was only publishing datasets directly from organisations and not from individuals.

4.1.6 Global Ship Strikes Database - International Whaling Commission

IWC https://iwc.int/management-and-conservation/ship-strikes

Contact persons: Alex ZERBINI (Scientific committee) sc.chair@iwc.int
Lindsay PORTER (Scientific committee) vicechair.sc@iwc.int
Iain STANILAND (Secretariat Head of Science, Conservation and Management) iain.staniland@iwc.int
David MATTILA (Human Impact Reduction) david.mattila@iwc.int

In 2007, the IWC launched a long-term initiative to collect and analyse both historic and current information on reported ship strikes, at a global scale. The aim of this work is to identify 'hot spots' where the occurrence of large numbers of whales overlaps with busy shipping lanes, and to share this information, once it has been checked and analysed, with the widest possible range of stakeholders to help develop targeted and effective mitigation actions. To understand the risk for whales and relate it to the densities of ships and whales, the IWC is collaborating with other relevant organisations at both regional and inter-governmental levels to share information and expertise. In 2010, the IWC held a workshop with the Agreement on the Conservation of Cetaceans of the Black Sea, Mediterranean Sea and Contiguous Atlantic Area (**ACCOBAMS**, https://accobams.org/conservations-action/ship-strikes/), which started a continuous collaborative work, to define relevant recommendations to address collisions between marine mammals and ships. Successively, in 2019, a joint workshop with the IWC, the IUCN, and the ACCOBAMS, was held to define how the IUCN's programme to systematically identify Important Marine Mammal Areas (IMMAs) might be overlaid with shipping information and used to help pinpoint ship strike hotspots. The Marine Environmental Protection Committee (MEPC) of the International Maritime Organisation (IMO) actively collaborates providing information about these hotspots and feedback on the collision reduction plans that have been implemented in several of these areas. The Article VIII of the IWC Convention invites Member States to fill a National Progress Report, with the aim to inform about the collaboration and conservation activities implemented during the year, within the IWC general objectives. Data are then uploaded into the IWC Portal, where a specific section dedicated to ship strikes events is available.

4.2 European

4.2.1 European Marine Observation and Data Network

EMODNET https://www.emodnet-biology.eu/

Contact person: Leen VANDEPITTE leen.vandepitte@vliz.be (VLIZ)

Established in 2007, EMODnet is a long-term marine data initiative spanning seven broad disciplinary themes: bathymetry, geology, physics, chemistry, biology, seafloor habitats and human activities. It works in partnership with all the European OBIS nodes.
The Biology portal, at the date of the preparation of this report, hosts more than 670 databases and covered marine species of phytoplankton, zooplankton, angiosperms, macroalgae, invertebrate bottom fauna, birds, mammals and fishes. The portal is OGC compliant and enables access to metadata descriptions of more than 1200 thematic biological datasets; being INSPIRE compliant these metadata records can also be found through the EU Open Data Portal. The portal identifies and focuses on biological data types, species, species attributes, sampling methods, and biological indicators, to support several legislations, such as the MSFD, by creating biological data products.

4.2.2 Reportnet – European Environmental Agency

REPORTNET https://cdr.eionet.europa.eu/
Contact: helpdesk@eionet.europa.eu

The European Environmental Agency (EEA) works actively to provide policy makers and European citizens with the latest available information on European biodiversity and ecosystems. The EEA's overall work is to support and inform policy development and implementation by means of data, information/indicators, and assessments, which integrate species and habitat analysis with wider assessments of ecosystems and their services. The EEA supports the above-mentioned nature directives through reporting via Reportnet (https://www.eionet.europa.eu/reportnet/about-reportnet-1) and the Biodiversity Data Centre (https://www.eea.europa.eu/themes/biodiversity/dc/dm#c0=10&c1=Data&b_start=0). The agency also works closely with the European Environment Information and Observation Network and the European Topic Centre on Biological Diversity (ETC/BD, https://www.eionet.europa.eu/), an international consortium led by the National Natural History Museum in Paris that works with the EEA under a framework partnership agreement. The main tasks of the ETC are: to support EEA in reporting on the EU environment; to provide the relevant information to support the implementation of environmental and sustainable development policies (EU Nature Directives); and to build capacity for reporting on biodiversity in Europe, mainly through the **European Information and Observation Network** (Eionet).

Reportnet is Eionet's infrastructure for supporting and improving the flow of data and information. Reportnet is based on a set of interrelated tools and processes which all build on the active use of the Web. Reportnet was developed in 2000 and has been operational since 2002. The system integrates different web services and allows for distributed responsibilities. Reportnet

was initially used for reporting environmental data to EEA, but is now also hosting some of DG Environment's reporting tasks. The open system permits deliveries to other national and international organisations in a very transparent way.

The **Central Data Repository** (https://cdr.eionet.europa.eu/) is part of the Reportnet architecture. The Central Data Repository provides a web interface for guiding the reporter through the reporting workflow, with key steps such as uploading files and presenting quality control feedback. It contains data reports on the environment as submitted to international clients. Each country either has a collection for its deliveries or a referral to a different preferred repository. The data reports within each country collection are arranged under the relevant reporting obligations or agreements.

Based on the information included in the Data Repository, EEA provides different products for consultation, including:

- Biodiversity Information System for Europe (BISE): a web portal that centralises information about European biodiversity (i.e., policies, data, assessments) in a single location. A dedicated section presents selected data catalogues and infrastructures, and offers reference data related to biodiversity in Europe, as developed and managed by BISE key partners, and other relevant organisations and projects (https://biodiversity.europa.eu/data-maps-and-tools/biodiversity-data).
- European Nature Information System (EUNIS): provides access to the publicly available data, compiled within the framework of Natura 2000 and EEA reporting activities (https://eunis.eea.europa.eu/).
- Geospatial data catalogue (SDI): a comprehensive catalogue of datasets, tables and maps organised in different topics, following the INSPIRE themes (https://sdi.eea.europa.eu/catalogue/biodiversity/eng/catalog.search#/home).

4.2.3 Data.europa.eu – Joint Research Centre

DATA.EUROPA https://data.jrc.ec.europa.eu; https://data.europa.eu/en
Contact: info@data.europa.eu

The data.europa.eu portal is a central point of access to European open data from international, European Union, national, regional, local, and geodata portals, intended to: give access and foster the reuse of European open data among citizens, businesses and organisations; promote and support the release of more and better-quality metadata and data by the EU's institutions and countries, or agencies and other bodies; educate citizens and organisations about the opportunities that arise from the availability of open data. The portal invites all the EU Member

States to publish their public data resources, such as open datasets also deriving from legislative requirements, and to make them accessible to the public whenever possible. It consolidates and merges the activities of the former EU Open Data Portal (which served data from the EU institutions, agencies, and bodies) and the European Data Portal (which focused exclusively on EU Member States and other European countries), launched respectively in 2012 and 2015, on the basis of Directive 2003/98/EC on the reuse of public sector information. These were funded by the EU, and managed by the Publications Office of the European Union (OP), of which it follows the mission, (i.e., to support EU policies and ensure that accessible and reusable data are accessible to the public, to facilitate transparency, economic activity, and the diffusion of knowledge), the vision, and the values of Transparency; Trustworthiness; Accessibility; and Service orientation. The portal aims at improving the accessibility to open data and promoting its use by public administrations, research centres, citizens, businesses, and any other interested organisation, while enhancing the transparency of European administrations. To foster the comparability of data published across borders, it presents metadata references in a familiar format (Data Catalogue Vocabulary - DCAT application profile for data portals in Europe), using the Resource Description Framework (RDF) technology. It provides translations of metadata descriptions in all 24 official EU languages using machine-translation technologies (eTranslation). Data providers are autonomous in publishing their metadata in the portal, which also publishes datasets of other European countries and organisations beyond the EU, and is updated when new datasets and content are available.

At the date of the preparation of this report, the data portal comprises 1.561.912 Datasets from 176 Catalogues and 36 Countries, which can be explored through a search engine (data tab), a map for geospatial data, or a SPARQL endpoint and API endpoint.

Among the available information, the **Marine Geoportal EMIS** collection (http://emis.jrc.ec.europa.eu/satellite/), developed by the Joint Research Centre (JRC), relies on biological and physical variables generated from both hydrodynamic models and satellite remote sensing. A number of these variables and advanced products are available here as raster datasets to the scientific and environmental managerial community through various tools (GIS Viewer, EMIS-R, Marine Analyst, Maps), which enable the user to conduct regional assessments. The geographical extent is 70N - 10S and 30W - 42E, at the available spatial resolutions of 4 and 2km. For example, the dataset EMIS - “Favourable feeding habitat of fin whale Monthly 1998-2018” (frequency of occurrence, %) identifies daily favourable feeding habitats of fin whales by linking their ecological traits with environmental variables from satellite remote sensing (Druon et al, 2012).

4.2.4 Marine Data Archive

MDA http://www.marinedataarchive.eu
Contact person: Katrina EXTER (VLIZ) katrina.exter@vliz.be, mda@vliz.be

To prevent scientific data from being lost, the Vlaams Instituut voor de Zee (VLIZ) in Belgium developed the Marine Data Archive: a secure, online repository where researchers can archive their data files in a well-documented manner. The MDA holds data from a wide range of international, European, and Flemish projects. It is free for anyone to use: to archive and manage their data, as a personal, project, or institutional archive, and as an open repository for data publication. Data of any type can be archived in the MDA: raw, processed, structured, data products, documents, images, etc. Data in the MDA are made public by linking them to the Integrated Marine Information System (Flanders), or to any other international catalogue (*e.g.*, OBIS, GBIF).

4.3 Regional

4.3.1 Network on the Conservation of Cetaceans of the Black Sea, Mediterranean Sea and contiguous Atlantic area

NETCCOBAMS https://accobams.org/

NETCCOBAMS is an internet tool developed by ACCOBAMS that transposes in GIS layers the main and most important information received by the ACCOBAMS Secretariat. The first NETCCOBAMS platform was devoted to all experts working on cetacean conservation (*e.g.*, the scientific community, managers, members of non-profit and intergovernmental organisations, relevant national and regional administrations, students), and is still being implemented. The main objectives of NETCCOBAMS are to facilitate the visualisation of important areas for cetacean conservation; to reinforce exchanges and collaboration between all actors in cetacean conservation; to assist ACCOBAMS Parties in taking appropriate management and conservation measures.

This platform is built upon the technological basis provided by SINAY, which is formed by a Big Data architecture, High Performance Computing Facilities and Dedicated Apps, and provides data, information and indicators that cover the whole ACCOBAMS area:

- Model-based maps of shipping noise for summer 2018;
- Model-based presence maps for 8 species;

- Acoustic Risk Maps: areas where fin whales, sperm whales and Cuvier's beaked whales are under risk of loss of auditory capabilities;
- Important Marine Mammal Areas (IMMAS);
- Data collected by ACCOBAMS during the Noise Hotspot project (2005 – 2015 data on the spatial and temporal distribution of impulsive noise generating activities);
- Input data used to produce model-based noise maps and habitat maps (*e.g.*, ship traffic map from AIS data in 2018, Sea Surface Temperature, salinity, Chlorophyll-A, depth, fin whale sightings, plastic debris presence along the coasts, a list of data on human activities).

4.3.2 Intercet

INTERCET https://www.intercet.it/

Contact persons: Guido GNONE ggnone@costaedutainment.it;
Michela BELLINGERI mbellingeri@costaedutainment.it

Intercet is an operational tool created within the GIONHA Project (2007-2013, http://www.gionha.it) to facilitate cooperation between the parties engaged in research on cetaceans and sea turtles in the northern Tyrrhenian Sea. It was developed by the Genoa Aquarium for the Liguria Region and was planned as a working tool, open to the entire scientific community, in the form of a Web-GIS platform that allows the aggregation, visualisation and integrated analysis of geo-referenced data and photographs. Since 2014, within the TursioMed project (2014-2017), this tool has been implemented in the Mediterranean Bottlenose dolphin Conservation Plan of ACCOBAMS as a tool to support the sharing, network management and analysis of data (including photo-ID) at Mediterranean level (Gnone et al, 2022), and was used to support networking on bottlenose dolphins at the Mediterranean level by involving 29 partners that shared data collected on cetacean species between 2004 and 2016. Thanks to external funding, it further developed into the InterMed project (2020-2022), which gathered data from 32 partners to improve knowledge about cetaceans and sea turtle populations in the whole Mediterranean Sea, and supported the finalisation and implementation of the Conservation and Management Plans for *T. truncatus*, *D. delphis*, and *G. griseus* at the Mediterranean level.

4.3.3 Sextant - Infrastructure de données géographiques marines et littorales

SEXTANT https://sextant.ifremer.fr/eng

Contact: sextant@ifremer.fr

Sextant is a marine and coastal Geographic Data Infrastructure (GDI) that aims to document, disseminate and promote a catalogue of data relating to the marine environment. Sextant

provides tools to promote and facilitate the archiving, consultation and availability of these data to the laboratories and partners of Ifremer, as well as for national and European actors that work in the marine and coastal field.

Sextant infrastructure and the technologies used are in line with the implementation of the INSPIRE Directive and make it possible to follow the Open Data approach. Data published by Sextant are either available free or restricted and can be used in accordance with the terms of the Creative Commons Licence selected by the data owner. Some datasets published by Sextant have a DOI, which enables them to be cited in a publication in a reliable and sustainable way. The long-term preservation of the data filled in Sextant is ensured by Ifremer infrastructure, which is operated by the Ifremer Scientific Information Systems for the Sea. At the date of the preparation of this report, 160 catalogues of laboratories, Ifremer units and partner organisations were present in this repository.

4.3.4 Joint Cetacean Data Programme

JCDP https://jncc.gov.uk/our-work/joint-cetacean-data-programme/
Contact: marinemammals@jncc.gov.uk https://jncc.gov.uk/contact/?ref=jcdp

Across the north-east Atlantic, there are multiple sources of data for cetacean species, from wide scale surveys, such as the Small Cetaceans in European Atlantic waters and the North Sea (SCANS) to other local sources from non-profit organisations, marine industry, academic projects, gathered at different temporal and spatial scales. Combining cetacean datasets from the existing evidence base has proven to be highly effective in enabling the assessment of cetacean populations across wide areas, as demonstrated by the Joint Cetacean Protocol (Joint Nature Conservation Committee, JNCC) that, between 2004 and 2016, collated data to successfully identify trends in distribution and relative abundance of cetacean species in the North Sea and adjacent sea regions. Between 2019 and 2022, the Department for Environment Food & Rural Affairs (DEFRA) of the UK funded the JNCC for the development of the Joint Cetacean Data Programme (JCDP), with the aim to provide a platform for collation, standardisation, storage, and access of cetacean data, collected at sea, via ship-based or aerial observations and digital surveys. It is a growing resource that will enable the best use of all the available and comparable data from which to carry out analyses at relevant spatial and temporal scales to inform cetacean management, policy and conservation. The development of the JCDP was a collaborative project with involvement of many stakeholders and contributors, from a range of organisations across the north-east Atlantic area. In 2022, an international Working Group on the JCDP of the International Council for the Exploration of the Sea (**ICES**) took over the project management to ensure future resilience.

The JCDP aims to streamline the process of accessing and utilising cetacean survey data by collating the existing evidence-base across the north-east Atlantic into a single accessible resource. The JCDP aims to achieve this vision through:

- the development of an international platform to host cetacean survey datasets from the north-east Atlantic;
- the development of a data standard to guide data collection and storage, enabling a high-quality collation of data (e.g., INSPIRE compliant);
- the provision of regularly updated open access data products for use in strengthening cetacean science and subsequent decision-making;
- a facilitated access to the collated dataset for use in bespoke analyses.

The JCDP platform comprises the JNCC Hub and the JCDP Data Portal:

- The JNCC Hub is a series of web pages hosted on the JNCC website (https://jncc.gov.uk/our-work/joint-cetacean-data-programme/), that provide information about the JCDP, and guidance to support data contributors and users. Documents and guidances are available to help inform cetacean survey protocols and data provision.
- The JCDP Data Portal (https://www.ices.dk/data/data-portals/Pages/Cetaceans.aspx) is a website which enables data providers to upload cetacean survey data into the Portal, and data users to explore data products and download datasets. At the moment of the preparation of this report, 89 datasets deriving from cetacean monitoring activities were available. The ICES has been contracted by JNCC to build the Portal, including the database, and the associated infrastructure to explore and access the data, which is hosted in the ICES Data Centre.

4.4 National

4.4.1 UK - Marine Environmental Data and Information Network Data Portal

MEDIN https://portal.medin.org.uk/portal/start.php
Contact person: Clare POSTLETHWAITE cfpo@noc.ac.uk; enquiries@medin.org.uk

The Marine Environmental Data and Information Network (MEDIN) Data Portal is a service that allows users, through a single point of access, to find information on the marine datasets that are held at the Data Archive Centres and at other public and private sector bodies. Marine data are held by many organisations in the UK and are collected for many different purposes, such as to assess the timing of tides; to determine the position of submerged obstacles; to monitor and forecast weather and ocean states; to locate marine structures; and for scientific research and marine conservation. MEDIN has produced a standard for marine metadata to harmonise

datasets at UK level, that also complies with other international conventions. MEDIN Sponsors include a range of UK marine organisations who support MEDIN's principles and lead the UK in marine data management. The partners represent government departments and agencies, research organisations, and private companies. They are committed to practising good data management to help future-proof and secure the UK's valuable marine data. At the date of the preparation of this report, the MEDIN Data Discovery Portal contains information on about 17.375 marine datasets, from over 600 UK organisations.

4.4.2 UK - The Archive for Marine Species and Habitats Data

DASSH https://www.dassh.ac.uk/
Contact: dassh.enquiries@mba.ac.uk

The Archive for Marine Species and Habitats Data (DASSH) is the UK Data Archive Centre for marine biodiversity data for both species and habitats, an initiative that was built by the Marine Biological Association (MBA) on its historic role in the promotion of marine science. DASSH was co-funded by the DEFRA and the Scottish Government, and is the MEDIN Data Archive Centre responsible for archiving marine species and habitat data from ad hoc sightings and non-effort-based surveys of marine life. It specialises in marine biodiversity data and/or information, including species or habitat survey data, species lists, habitat or biotope lists, species or habitat/biotope distribution maps, figures, images and video clips, supporting a wide range of data types and file formats, and it provides tools and services for the long-term curation, management, preservation and publication of marine species and habitats data, within the UK and internationally. Through partnerships with other UK and European data centres, DASSH interconnects with other international data portals (e.g., EMODnet, EurOBIS, GBIF).

4.4.3 IT - National Environmental Information System - ISPRA

SINA https://www.isprambiente.gov.it/it/attivita/reti-e-sistemi-informativi-ambientali/sistema-informativo-nazionale-ambientale-sina

The data and geographic information collected by Italian Institute for Environmental Protection and Research (ISPRA) and National System for Environmental Protection (SNPA) are stored and made public and accessible also in real time, as part of the National Environmental Information System (SINA). SINA was created in 1988 as a strategic informative system to be used for the collection, elaboration, and reporting of geographic and environmental information. It guarantees the effective link between the initiatives implemented by the various parties in the collection and organisation of data, the consistent maintenance of information flows, and the divulgation of data to public administrations, researchers, experts, and citizens. Data collected

are a reference for the institutional activities of public administration bodies. SINA workflow is based on monitoring, data collection and storing, data elaboration and environmental analysis, reporting and data publishing.
Outcomes are reported in areas of interest. Some examples within the sections **Nature**, **Biodiversity**, and **Sea** are reported below:

- Habitats Directive (92/43/CEE) Reporting: this site contains the data collected and processed by Italy for the HD Report to the European Commission. The Report collects updated data on distribution, conservation status, pressures, threats, and trends relating to all animal and plant species and habitats of Community interest that are present in Italy.

- Atlas of marine protected species: created as part of the project to establish the Regional Biodiversity Observatory of the Sicily Region, the Atlas constitutes a collection of files of marine species that are either protected or threatened with extinction, considered as such pursuant to the various Community Directives and international agreements for the nature conservation (i.e., HD, BD, SPA / BIO, Bern and Bonn Directives, the Convention on International Trade in Endangered Species of Wild Fauna and Flora, ACCOBAMS) and which appear to have at least one indication in Marine Protected Areas and Natura 2000 Areas of Sicily.

- **Biodiversity and Nature – Centralised Information System - SIC http://www.db-strategiamarina.isprambiente.it/app/#/**

Contact person: Giordano GIORGI (ISPRA) giordano.giorgi@isprambiente.it

The SIC was created to respond to the requirement of collecting, managing and sharing at Community level the data from the Monitoring Programs within the MSFD (2008/56/CE). The SIC collects data from the Regional Agencies for Environmental Protection (ARPAs) of the 15 Italian coastal regions relating to the physical components of ecosystems and biodiversity with a focus on species already studied under the HD and BD (e.g., *T. truncatus*, *Calonectris diomedea, Puffinus yelkouan*). The SIC is also accessible through a Web-GIS allowing the visualisation of spatial information deriving from the monitoring programmes implemented for the MSFD.

- **National Network of Biodiversity – NNB https://www.nnb.isprambiente.it/en**

Contact: nnb@isprambiente.it

The NNB is a shared data management system that consists of a central node, which allows performing search and management operations on the data, and peripheral nodes (databases that have primary biodiversity data). It aims at guaranteeing consultation and efficient integration

of information on biodiversity, without the need to physically transfer the data, which always reside within the cooperating entities that hold the legal rights. Any local or national institution, research body, or other organisation in possession of data on biodiversity or information related to samples preserved in scientific collections, can join the NNB. Through the aggregation of the current state of knowledge on biodiversity in Italy, the NNB aims to improve the dissemination and sharing of data on biodiversity, including those collected within EU environmental legislation, making them accessible for pure and applied research, education and training, and to represent a strategic national tool to inform political decisions, and guarantee a sustainable use of Italian natural resources.

The databases owned by the individual nodes differ in structure (different fields) and architecture (different DBs, type Access, Oracle, MySQL, etc.), but they are able to communicate through the BioCASe Protocol. The latter guarantees, through a set of rules, an intrinsic communication between the nodes themselves and the international community that participates in the BioCASe network. The Network is able to ensure interoperability with similar international infrastructures (LifeWatch, GBIF, etc.) and with the National GeoPortal, in accordance with the provisions of the INSPIRE Directive (Legislative Decree 32/2010).

4.4.4 IT - Joint Cetacean Database and Mapping

JCDM https://jcdm.dss.uniroma1.it/

Contact persons: Giovanna JONA LASINIO giovanna.jonalasinio@uniroma1.it
Daniela Silvia PACE danielasilvia.pace@uniroma1.it

While cetacean data are growing exponentially, the efforts to combine them into model outputs and make them readily available and understandable have lagged behind. The Joint Cetacean Database and Mapping (JCDM) is an initiative that resulted from a joint effort of the Departments of Statistical Science and of Environmental Biology of the University of Rome “La Sapienza”, and the support of different partners that act in the field of marine conservation at the Mediterranean level, with the goal to transform complex data into accessible information by collecting and synthesising the existing knowledge, using novel modelling approaches and providing easy access to results.

Different Tools have been developed within JCDM:

- Georeferenced Database composed by marine mammals’ presence data from existing geoportals (i.e., OBIS SEAMAP, GBIF); data from dedicated monitoring surveys collected from different research platforms; presence only data extracted from social media; presence data

resulting from the photo identification of the individuals collected from dedicated monitoring surveys.

- Analytical tools including references to statistical models employed to deal with specific data features (models for presence-only data, mark-recapture models for residency patterns, e.g., Martino et al, 2021); a repository of R codes implementing the proposed statistical models freely available to the scientific community.
- A Web Tool developed to easily interact with the comprehensive georeferenced database and with the results of the analytical tools. JCDM engine allows to elaborate the presence data at different levels by aggregating at species and spatial level; visualise the case studies; represent the model results in the form of descriptive maps and tables; download the data and the model results.

5. Format, storing and validation of data collected on marine macrofauna from large vessels (ferries, cargoes) within the Europe-MTT project

As anticipated at the beginning of this report, the Europe Marine Transborder Transect for the monitoring of macrofauna and anthropogenic pressures (Europe MTT) is a follow up of the IMPEL FLT EUROPE 2020 project, aimed at harmonising monitoring methods for large marine vertebrates, namely cetaceans and sea turtles, and threats to them, such as marine litter and maritime traffic, along transborder transects, using large vessels such as ferries and cargoes as platforms of observation.
The concept behind both projects was raised from the experience gathered by a number of research institutions and non-profit organisations from several years of research and monitoring marine macrofauna onboard large vessels (i.e., cargoes and ferries) as platforms for systematic surveys.

Three main networks are involved in the projects:

- the **Fixed Line Transect Mediterranean Monitoring Network** (FLT MED NET). Led by ISPRA, the network started with a first route in the Tyrrhenian Sea in 2007 and grew since then with the introduction of new transects in the Central-Western Mediterranean Sea. It currently comprises over 12 routes that cross high seas and national waters among Italy, France, Spain, Greece, Tunisia, and Morocco. FLT MED NET includes 13 scientific partners from Italy, France, Spain, Tunisia and Greece. All partners share the same protocol to systematically survey the vertebrate marine species listed in the Habitats Directive (cetaceans, marine turtles and seabirds) and main threats to them (maritime traffic and marine litter).

- the **"North-East Atlantic" Network**. Initiated in 2001 within the Atlantic Research Coalition, and currently led by the UK charity ORCA (https://www.orca.org.uk/), this fixed line monitoring network regularly surveys nine regions: Arctic Waters, North Sea, English Channel, Celtic Sea, Irish Sea, Minches and West Scotland, Bay of Biscay and Iberian Coast, Wider Atlantic and the Mediterranean Sea, and, since 2018, also the south-west of Greenland, the west coast of the United States of America, and the South Atlantic Ocean (Argentina, Chile, the Falkland Islands, South Georgia and the South Sandwich Islands, and the Antarctic Peninsula), with the support of 14 ferry and cruise companies. ORCA's Marine Mammal Surveyor Teams are composed of 3-4 fully trained volunteers that conduct monthly surveys from the vessel bridge (or another forward-facing platform), using a standardised survey protocol, based on the distance sampling methodology, to ensure that data collection is rigorous and comparable.

- **CETUS** – the cetacean monitoring programme in Macaronesia. Led by the Interdisciplinary Centre for Marine and Environmental Research (CIIMAR - University of Porto), and in partnership with the TRANSINSULAR (ETE Group) cargo ship company, this monitoring programme has been undertaking surveys of cetacean occurrence onboard cargo ships in the Atlantic region between Continental Portugal, the Macaronesian archipelago, and the north-western coast of Africa, since 2012. Trained volunteers collect data on the presence of cetacean species and other pelagic megafauna, survey effort, weather conditions and marine traffic, among other variables, using an adapted protocol based on the one used by the FLT MED NET.

Although the networks pursue a common aim, they adopted methodologies for data collection and analysis that were not designed in the first place to respond to the European legislative drivers and some gaps still exist. The main aim of the EU MTT project is to fill such gaps in data collection, storage and analysis, in order to allow the gathering and sharing of harmonised data that could respond to the requirements of the MSFD and other EU environmental directives, and efficiently support decision-making across Europe. To do so, a report was already produced within the first phase of IMPEL FLT EUROPE project (see Campana and Vighi, 2021, for details) focusing on the data collection protocol and data needed to feed the main environment Directives for cetaceans, sea turtles, sea birds, and marine litter. Here, we highlight the main similarities and differences regarding the data gathered and the data repositories used by the network partners involved in the project

Data gathered within a large vessel survey include three groups of parameters, which are related to: operational factors (observation distance, speed, and height of the platform); environmental (i.e., meteorological) factors (wind force and direction, cloud coverage); target-related factors (i.e., identification of species, count of individuals, behaviour), and are collected either at regular intervals or whenever any sighting/modification of status occurs.

5.1 Operational parameters

The three networks perform observations from large vessels with varying speed and height, which are recorded at the beginning of each survey along with the names of the observers, of the ship, the date, and the transect code. Observations are generally performed from the command deck or the highest accessible site of the ship, focussing on a survey area located 180º ahead of the vessel, or two separate areas from the bow to 90º or 130º on port and starboard sides. Networks adopt either the distance sampling method for the reporting of cetacean sightings, or fixed width strip methods, generally for turtles and marine litter; with some, generally negligible, differences on the methodology used to measure the distance and angle to the target. During distance sampling surveys, distance and angle are recorded at every sighting of a target, using binoculars

or range sticks to measure distance, and goniometers to measure the angle either from the ship's bow, or from north. On the other hand, during fixed strip width surveys, the width of the strip is measured with a clinometer and noted at the beginning of each transect, and is kept constant along the survey. The main characteristics of the surveys adopted by the three networks are summarised in Table 1.

Table 1. Operational parameters applied during the monitoring surveys of the FLT MED NET, ORCA and CETUS networks.

Network/programme	FLT MED NET	ORCA	CETUS
Platform type	Ferry	Ferry	Cargo
Platform height (range or mean)	12-29 m	10-29 m	13.5-16 m
Platform speed (range or mean)	14-29 knots	10-25 knots	11-16 knots
Number of observers and rotation shifts	2-4, rotation every 30-60 minutes.	3-4, rotation every 30 minutes.	2, rotation every 60 minutes.
Area observed (total and by each observer)	270º ahead (130º each side)	180º ahead (100º each side)	180º ahead (90º each side)
Sampling technique	Cetaceans: Distance sampling Sea turtles and litter: 50 m fixed strip width.	Distance sampling	Distance sampling based protocol for data collection; the ship position alone is used for analyses to compute large-scale habitats.
Method of angle and distance measurement	Distance: rangefinder stick or reticle binocular; angle: compass or goniometer.	Distance: reticle binoculars or estimate by eye (if target closer than 250 m); angle: angle board (goniometer).	Distance: reticle binocular; angle: binocular compass (along with ship bearing, which is used for calculating the angle from the bow).

5.2 Meteorological parameters

Observations are undertaken during daylight hours, from sunrise to sunset, and data regarding the environmental conditions are generally recorded by the three networks at the beginning of

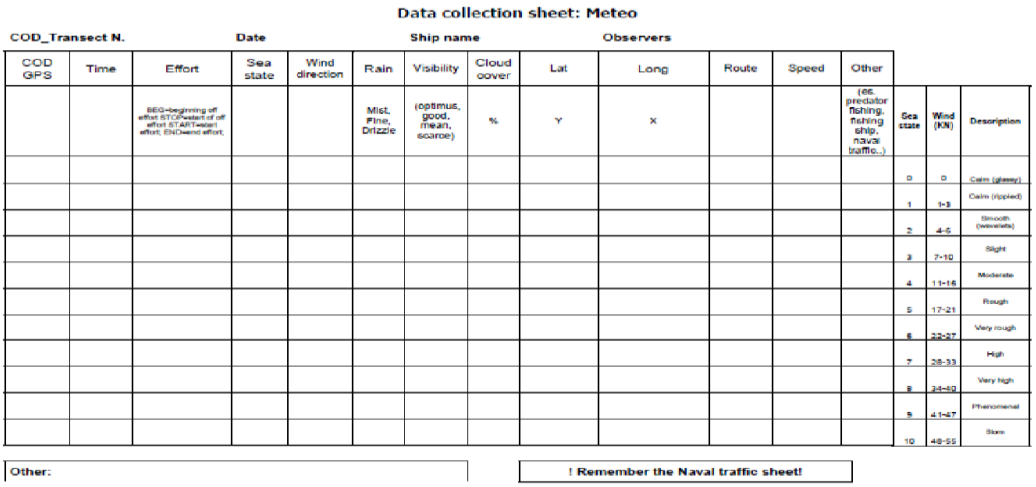
each transect, and whenever any significant change in the conditions occurs. The FLT MED NET protocol considers positive effort conditions when the wind force is less or equal to three on the Beaufort scale. On the other hand, favourable conditions for ORCA surveys are considered with a swell height ≤2 m, and Beaufort Sea state ≤6 (Matear et al, 2019). Finally, within the CETUS program, monitoring is mostly conducted during summer, to allow more stable weather conditions, and stops whenever sea state or wind state are higher than four (on the Douglas or Beaufort scales, respectively), visibility is lower than 1 km, or during heavy rain (Correia et al, 2015, 2020). Details of survey conditions and meteorological data collected by the networks are shown in Table 2, evidencing some differences in their recording.

Data on meteorological variables are recorded within the FLT MED NET in the 'meteo data collection sheet' (Morgado et al, 2017; Figure 3) at the beginning and end of each effort period; and each time a change occurs; as the ship position is automatically recorded by a handheld GPS or applications each minute, meteorological information are available for the complete track of effort. Meteorological parameters include: Beaufort Sea State (still used by most partners to collect sea conditions, although it would be recommended to use Douglas scale to assess sea state and Beaufort scale to assess the wind), wind direction, precipitation, visibility, and cloud cover (in %). Data are collected under all weather conditions even if only data collected when Beaufort ≤3 are used for the analysis. The definition of the Beaufort scale is done through descriptive observation of the sea state and range of wind speed when available (both described in the data collection sheet).

Table 2. Meteorological parameters collected during the monitoring surveys of the FLT MED NET, ORCA and CETUS networks.

Network/programme	FLT MED NET	ORCA	CETUS
When			
	start and end of each survey period; whenever weather conditions change	start and end of each survey period; every 30 minutes or if weather conditions change	start and end of each survey period; whenever weather conditions change
How			
Data collection form	Dedicated sheet-App	Dedicated sheet	Marker on GPS
Conditions for positive effort	Beaufort Sea state ≤3	swell height ≤2 m, Beaufort Sea state ≤6	Douglas and Beaufort ≤4; visibility higher than 1 km, no heavy rain
Meteo conditions			

Sea State	Sea state and wind together on the Beaufort scale (some partners use Douglas scale for sea)	Beaufort Sea state 0-6; Swell (1 Absent, 2 Light (>0-1m), 3 Moderate (1-2m), 4 Heavy (>2m))	Douglas scale
Wind State	-	Beaufort scale	Beaufort scale
Wind Direction	Yes	-	Binoculars compass
Cloud Coverage	Expressed in %	-	Expressed in %
Glare	-	NA None, P Port, S Starboard, A Ahead, PA Port-ahead, SA Starboard-ahead, All the entire 180º survey area	-
Visibility	Optimus, good, mean, scarce	Expressed in km	Categorical scale 1 to 10 (from 0 to 50 km)
Precipitation	Mist, fine, drizzle	1 None, 2 Rain, 3 Hail, 4 Fog/Mist, 5 Sleet, 6 Snow	Occurrence of rain



Data collection sheet: Meteo

COD_Transect N. Date Ship name Observers

COD GPS	Time	Effort	Sea state	Wind direction	Rain	Visibility	Cloud cover	Lat	Long	Route	Speed	Other
		BEG=beginning off effort STOP=start of off effort START=start effort, END=end effort,			Mist, Fine, Drizzle	(optimus, good, mean, scarce)	%	Y	X			(es. predator fishing, fishing ship, naval traffic..)

Sea state	Wind (KN)	Description
0	0	Calm (glassy)
1	1-3	Calm (rippled)
2	4-6	Smooth (wavelets)
3	7-10	Slight
4	11-16	Moderate
5	17-21	Rough
6	22-27	Very rough
7	28-33	High
8	34-40	Very high
9	41-47	Phenomenal
10	48-55	Storm

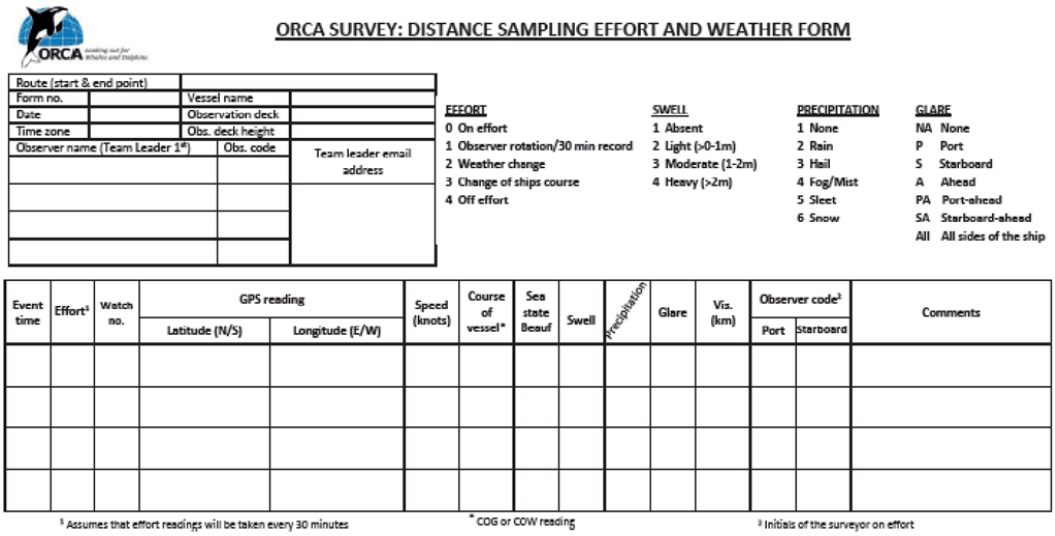
Other:

! Remember the Naval traffic sheet!

Figure 3 Data collection sheet used within the FLT MED NET to record meteorological conditions.

Within ORCA, environmental information is recorded within the dedicated “distance sampling effort and weather form” (Robbins et al, 2020; Figure 4) at the start and end of each survey period and is updated every 30 minutes or if sighting conditions change. The survey period ends when all observers are off effort; whether for lunch break, poor weather, dusk etc. Recorded environmental conditions include: Beaufort

Sea State; Swell (1 Absent, 2 Light (>0-1m), 3 Moderate (1-2m), 4 Heavy (>2m)); precipitation (1 None, 2 Rain, 3 Hail, 4 Fog/Mist, 5 Sleet, 6 Snow); glare (NA None, P Port, S Starboard, A Ahead, PA Port-ahead, SA Starboard-ahead, All All sides of the ship); Visibility (how far you can clearly see from the observation platform, expressed in km).



ORCA SURVEY: DISTANCE SAMPLING EFFORT AND WEATHER FORM

Route (start & end point)			
Form no.		Vessel name	
Date		Observation deck	
Time zone		Obs. deck height	
Observer name (Team Leader 1st)		Obs. code	Team leader email address

EFFORT
0 On effort
1 Observer rotation/30 min record
2 Weather change
3 Change of ships course
4 Off effort

SWELL
1 Absent
2 Light (>0-1m)
3 Moderate (1-2m)
4 Heavy (>2m)

PRECIPITATION
1 None
2 Rain
3 Hail
4 Fog/Mist
5 Sleet
6 Snow

GLARE
NA None
P Port
S Starboard
A Ahead
PA Port-ahead
SA Starboard-ahead
All All sides of the ship

Event time	Effort[1]	Watch no.	GPS reading		Speed (knots)	Course of vessel*	Sea state Beauf	Swell	Precipitation	Glare	Vis. (km)	Observer code[2]		Comments
			Latitude (N/S)	Longitude (E/W)								Port	Starboard	

[1] Assumes that effort readings will be taken every 30 minutes
* COG or COW reading
[2] Initials of the surveyor on effort

Figure 4 Effort and weather data collection form used within the ORCA network.

Within CETUS, all the information is recorded on a tablet, using a dedicated app to record the entire transect, and specific points where/when further information is collected. The observers assess sea state using the Douglas scale, wind speed using the Beaufort scale, visibility on a categorical scale of values from 1–10 (covering visibility ranges from 0 m to over 50,000 m, estimated based on the definition of the horizon line and reference points at a known range, e.g., ships with an AIS system) and the occurrence of rain (Correia et al, 2019). A "meteo" point with weather data is recorded at the beginning, ending and every time the weather changes, following a specific coding system as shown below and in Figure 5:

Marker name: me (=meteo) **Marker type**: me (=meteo) **Description**: s (=sea state), w (=wind state), wd (=wind direction), c (=cloud cover), v (=visibility), r (=rain).

Marker name: ef **Marker type:** MMOs **Description:** on/off ef – effort MMO – observers (first and last) **Marker name:** me **Marker type:** me **Description:** s:w:wd:v:c:r me - meteo s – sea state w – wind state wd – wind direction v – visibility c – cloud cover (%) r – rain	**Marker name:** v **Marker type:** v **Description:** s:m:b v – vessels s – small(<5m) m – medium(5-20m) b – big (>20m) **Marker name:** ds **Marker type:** l/r;MMO **Description:** d:a:v ds – distance sampling l/r – left/right d – scale distance a – angle v – vessel angle	**Marker name:** ce **Marker type:** sp **Description:** :::t/f:h:ap/in/av:s:m:b:comments ce – sighting sp – sp code ::: – min.max.best h – heading s:m:b – vessels **Marker name:** as **Marker type:** sp **Description:** :.. as – associated sp – sp code/name	**Sea state:** 0. 0 1. 0-0.1m 2. 0.1-0.5m 3. 0.5-1.25m 4. 1.25-2.5m -- 5. 2.5-4m 6. 4-6m 7. 6-9m 8. 9-14m 9. >14m	**Wind state:** 0. Calm 1. Light air 2. Light breeze 3. Gentle breeze 4. Moderate breeze -- 5. Fresh breeze 6. Strong breeze 7. Near gale 8. Gale 9. Strong gale 10. Storm 11. Violent storm 12. Hurricane	**Visibility:** 1. <50m 2. 50 to 200m 3. 200 to 500m 4. 500 to 1000m -- 5. 1000 to 2000m 6. 2000 to 4000m 7. 4000 to 10000m 8. 10000 to 20000m 9. 20000 to 50000m 10. >50000m	**Rain:** 0. No rain 1. Little rain -- 2. Medium rain -- 3. Lots of rain
Delphinus delphis – dd; *Lagenodelphis hosei* – lh; *Phocoena phocoena* – pp; *Stenella atenuata* – sa; *Steno bredanensis* – sb; *Stenella coeruleoalba* – sc; *Stenella clymene* – scl; *Stenella longirostris* – sl; *Stenella frontalis* – sf; *Tursiops truncatus* – tt; *Lagenorhynchus albirostris* – la; NI *Delphinidae* – d; *Balaenoptera acutorostrata* – ba; *Balaenoptera borealis* – bb; *Balaenoptera edeni* – be; *Balaenoptera physalus* – bp; *Balaenoptera musculus* – bm; *Megaptera novaeangliae* – mn; NI Baleen whale– b; *Grampus griseus* – gg; *Globicephala sp.* – gm; *Hyperoodon ampullatus* – ha; *Kogia breviceps* – kb; *Mesoplodon densirostris* – md; *Orcinus orca* – oo; *Peponocephala electra* – pe; *Physeter catodon* – pc; *Pseudorca crassidens* – poc; *Ziphius cavirostris* – zc; NI *Ziphiidae* – z; NI *Cetacea* – c							

Figure 5 Coding instructions for data collection used by CETUS.

5.3 Species and other target-related parameters

During a sighting, the minimum data collected by the three networks include the position of the ship expressed in latitude and longitude, the date and time, the identification of the species and approximate group size and composition. If distance sampling methods are used, also the angle and approximate distance to the sighting are noted. If sightings are not identified to the species level, it is generally agreed to allow the identification to the "lowest taxonomic level possible at sea" (Alves et al, 2018), or to associate the identification of the species with the certainty of their identification (Kiszka et al, 2007). Similarly, as the definition of group is not straightforward for highly mobile cetaceans, and is often interconnected with the definition of behaviour, groups are generally defined as animals seen in close proximity at the same time, showing similar behaviour, and heading in the same direction.

Within the FLT MED NET, during each cetacean sighting, the ship location is marked on the handheld GPS, and data are annotated on the "sighting data collection sheet" (Figure 6), specifying the time, observer, side of sighting, the angle between the detected group and the ship track line, the linear distance from the ship, species, number of individuals, direction of swimming, and surface behaviour.
Data on radial distance and angle between the detected group and the track line are recorded by graduated binoculars, or using a personal measuring stick and a goniometer. The maximum range considered for cetacean sightings is approximately 4 km from the ship, as only large whales can be detected at a further distance (Morgado et al, 2017).

Data collection sheet: Sightings

COD Transect N. | Date | Ship name | Observers

N/COD GPS	Time	Ship position		Side	Obs	Species	N° Tot			N° Juv	Distance	Angle	Direction of swim	Responce to ship			Behaviour	Ph	Collision or Near collision	Ships (see sheet)
		Lat (Y)	Long (X)				Min	Max	Best			[0-180°]	(0-360°)	Apr	Esc	Indif				

Behaviour

Behaviour	Superficial	Speed	Progress	Direction	Group association
Travel	Half Leap	Slow	Straigth	Same	(i.e. 2+2; 2+2+1)
Rest	Full Leap	Normal	Irregular	Different	
Play	Dorsal Fin	Fast	None	Circle	**Type**
Feeding Wild	Surfing	Porpoising	Zig-zag		Males
Feeding Net	Blow	Floating			Fem/juv
Mating	Breach				Mother/calves
Unknown	Spyhopp				

Unidentified Species (US):

- L LARGE "whale" species
- DD Distinct Dorsa
- ID Indistint Dorsal
- M MEDIUM "cetacean" species
- LF Large Fin
- SF Small Fin
- S SMALL patterned "dolphin" species

Figure 6 Data collection sheet used within the FLT MED NET to record cetacean sightings.

Where species identification cannot be confirmed, sightings are either downgraded to unidentified dolphin/whale, unidentified “large whale species” (L), “medium cetacean species” (M), “small dolphin species” (S), or left in the unidentified category “US”. Similarly, when it is not possible to determine the exact number of individuals, a minimum and a maximum number of animals are registered, as well as the most probable number of individuals according to the observer´s perception (best estimate). Data on the general behaviour of the species is collected, classified in categories, such as travelling, resting, socialising, and feeding/foraging. The group composition (i.e., presence of young individuals) and the behaviour of the animals towards the ship (i.e., indifferent, avoiding or approaching) are also registered, as well as, for travelling animals, information on the heading of the group and swim speed.
Cases of collision or near-collision are also reported with details of the dynamic of the event, considering “Near collision” any event when animals are sighted in front of the ship at a minimum distance of 50 m ahead of the bow and 25 m on the side, without showing evident approaching behaviour.

Within ORCA, for each cetacean sighting, the ship location is collected via the ships onboard GPS console, the time of the sighting, species identity, certainty of the species identity, group size, number of calves, distance and angles from the ships’ bow to animals are recorded along with information on the animals’ behaviour, direction of travel, cue that led to the sighting and if it was a mixed species sighting, using the *distance sampling sightings form* (Robbins et al, 2020; Figure 7).

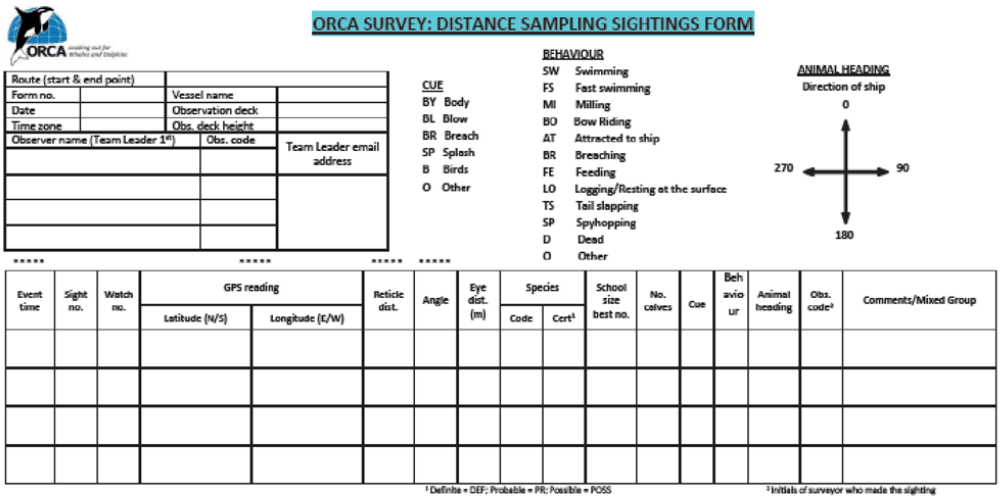


Figure 7 Sighting data collection form used within ORCA.

The fields required in the form include: the time and sighting number, watch number (which links to the effort and weather form), GPS position of the ship, reticle distance (radial distance to the sighted animal(s)), angle from the bow of the ship to the animal or centre of a group of animals,

eye distance in metres to the sighted animal(s) (only collected if reticle distance is not possible), species, along with its identification certainty ('DEF' (definite), 'PR' (probable) or 'POSS' (Possible)), school size, number of calves, cue for the animal detection (BY Body, BL Blow, BR Breach, SP Splash, B Birds, O Other), behaviour at the time of sighting (SW Swimming, FS Fast swimming, MI Milling, BO Bow Riding, AT Attracted to ship, BR Breaching, FE Feeding, LO Logging/Resting at the surface, TS Tail slapping, SP Spyhopping, D Dead, O Other), including primary, secondary and tertiary behaviours, animal heading in relation to the ship, if the sighting was a mixed species group.

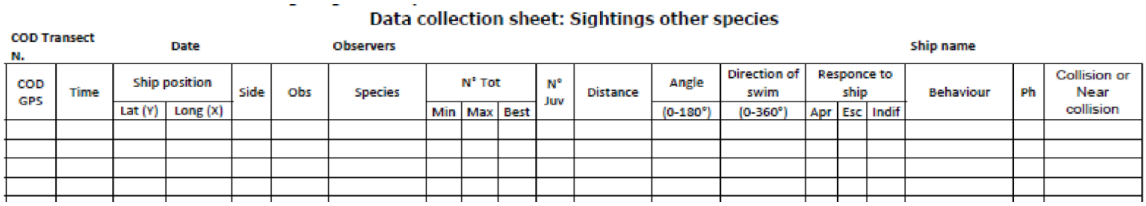
Within the CETUS network, during cetacean sightings, observers record the GPS position, species identity, the distance and angle of the position of the animal(s) in relation to the ship, the number of animals within the group, their behaviour towards the ship (if any), and information on the heading of the group (Correia et al, 2020). Species identification is attempted to the species level, although the identity assigned is always at the taxonomic level at which the observers are confident of their identification, being either the genus or unidentified dolphin, baleen, beaked whale, or cetacean. Groups are defined as animals swimming in the same direction, other than having the same behaviour, and being in close proximity. When the exact number of individuals cannot be assessed, a minimum and a maximum number of animals are registered, as well as the most probable number of individuals according to the observer´s perception (best estimate).

Data on the general behaviour of the species is also collected, classified in categories, such as travelling, resting, socialising, and feeding/foraging. The group composition (i.e., presence of young individuals) and behaviour of the animals towards the ship are also registered (i.e., indifferent, avoiding or approaching), as well as, for travelling animals, information on the heading of the group and swim speed.

Data are recorded first in the form of a distance sampling point with data on distance and angle of the sighting, following specific codes (Figure 5), as:
Marker name: ds (=distance sampling) **Marker type**: l/r (=left/right); MMO **Description**: d (=distance), a (=angle), v (=vessel angle),
and then adding another marker with the sighting information, in the form of:
Marker name: ce (=sighting) **Marker type**: sp (=sp code) **Description**: (min;max;best), t/f (=true/false (not sure of the sighting)), h (=heading), ap/in/av (=approach/indifferent/avoiding (behaviour towards the ship), s;m;b (=vessels).

5.4 Other data collected within the networks

Even though, within the networks, priority is given to cetacean detection, data on other marine megafauna are collected. The FLT MED NET collects information on two species of sea birds (*Puffinus yelkouan* and *Calonectris diomedea*), selected considering their priority for conservation purposes, five large fish groups (*Mola mola*, *Mobula mobular*, *Xiphias glaudius*, *Thunnus ssp*., Istiophoridae, and sharks), and jellyfish. Sea turtles are considered of high priority and systematic data on their presence is also collected with the fixed-strip width methodology by the observer dedicated to marine litter monitoring. Data regarding these species are collected using a dedicated form with similar fields as the data collection form for cetaceans (Figure 8). For species usually observed in large groups, groups are categorised into small (<10 individuals), medium (10-100), and large (>100).



Data collection sheet: Sightings other species

COD Transect N. | Date | Observers | Ship name

COD GPS	Time	Ship position		Side	Obs	Species	N° Tot			N° Juv	Distance	Angle	Direction of swim	Responce to ship			Behaviour	Ph	Collision or Near collision
		Lat (Y)	Long (X)				Min	Max	Best			(0-180°)	(0-360°)	Apr	Esc	Indif			

Figure 8 Heading of the FLT MED NET data collection sheet for sightings of species other than cetaceans.

Sightings of other large vertebrates (e.g., turtles, sharks, sunfishes) are also registered by CETUS. However, for pelagic megafauna other than cetaceans, data are always collected opportunistically. In these cases, only taxonomic information and the number of individuals (as well as optional comments about the sighting, e.g., animal behaviours, presence of calves or others) are registered (Correia et al, 2019).

Data on maritime traffic are collected both by the FLT MED NET and CETUS networks, using a similar sampling protocol that was specifically designed and is applied to all transects to provide real-time information on maritime traffic in the presence and absence of cetaceans. Within the FLT MED NET surveys, all vessels longer than 5 m and visible by eyesight are recorded at the beginning and end of the survey effort, each time a cetacean sighting occurs (record in presence of cetaceans); and at approximately every hour throughout the transects when animals are not sighted, on a dedicated data collection sheet (Figure 9). Vessels are classified as Small (<5m), Medium (5 - 20m, distinguished in: Motor, Sailing, Fishing) and big (>20m, such as cargoes, tankers, passenger ships).

The CETUS network uses a protocol based on that applied by FLT MED NET to monitor maritime traffic: the number of vessels, categorised by type and size, detected with or without binoculars

all around the ship position, are registered at the beginning and end of each survey, at every sighting of cetacean species, and at every hour throughout the survey effort.

Data collection sheet: Naval Traffic

COD Transect N.		Date		Ship name				Observers				
				< 2 NM				> 2 NM				
Cetacean Presence/Absence	Time	Ship position		Small	Medium		Big	Small	Medium		Big	Other
		Lat (Y)	Long (X)	< 5m	5m < X < 20m		> 20m	< 5m	5m < X < 20m		> 20m	
					Motor	Sail	Fishing		Motor	Sailing	Fishing	

Figure 9 Heading of the FLT MED NET data collection sheet for maritime traffic data.

Within the FLT MED NET, floating marine litter presence is also recorded by a dedicated observer within a fixed width strip, using a specific data collection form based on the master list of marine litter items produced within the MSFD.

5.5 Data quality control and data storage by the three networks

As environmental monitoring through commercial vessels travelling along fixed routes is mostly based on operationally defined parameters, the precise application of agreed protocols, supported by training, photographic documentation, and ground-truthing approaches are the means to assure the quality of measured data. To guarantee that consistent data are collected according to the parameters described above and reported using standard formats, after being collected in the field, data undergo a quality control process.

As reading a protocol and implementing it can lead to variable interpretations, the FLT MED NET foresees the development of dedicated training sessions, collecting data on the same ship, among partners with a person from each partner, to assure they operate the same way. As one team can ask for more precision than another (e.g., rounding measurements, use different tools), all partners need to be sure the procedure they follow and tools they use are appropriate and consistent.

As well, both FLT MED NET and CETUS researchers make verifications and validations of data at several stages during their processing and the development of the final dataset, including the digitalisation of the data to excel files, within the MySQL database, in ArcGIS and in R, while and after structuring the final dataset. CETUS has added data quality variables to the dataset, based on the observers' experience and on photographic validation (when possible) (Oliveira-Rodrigues et al, 2022).

Similarly, all data recorded within ORCA undergo checks by trained scientists to ensure a high level of quality control and are systematically checked using set protocols to ensure that records are accurate, and that spatial data are correctly formatted. Additional checks are completed annually for any data collected within that year, to reduce the potential for erroneous data being archived (Matear et al, 2019).

Collected data are usually first stored in individual project-based repositories or national databases. Separate datasets are kept by each institution involved in the FLT MED NET with data regarding each of the transects surveyed within the network. However, the standardisation of data collection forms, including the use of the same fields and codes for data categorisation, and the development of a common data recording tool (AtSea on ODK open-source mobile data collection platform), allows the data from the network to be joined together for the purpose of large scale or regional analyses. In addition, in the last year, a dedicated SQL-based database with a Web-GIS user-friendly interface was specifically designed to gather all data from the FLT MED NET partners. The infrastructure is hosted in the NNB (ISPRA) and currently all data from 2007-2021 period are stored there (Di Stefano et al, 2023).

Data collected within ORCA are also stored in both physical and digital formats through archives, log sheets, ArcGIS Geodatabases and relational databases, and in the ORCA Data Portal (Matear et al, 2019).

Data collected within CETUS are logged into an Excel database, then structured using MySQL internal database, and processed in R Software to structure the final dataset to be published open access (Correia et al, 2019, 2022).

6. Potential use and interoperability of data collected in the framework of the Europe MTT project within international databases/repositories

Several partners of the EU MTT project already share the data collected from large vessels in some of these international repositories, providing both data on the sightings of species and on the effort, except for the IWC ship strike reporting.

The following section describes the main characteristics of the databases listed in the first part of the document in terms of access requirements, use of standard formats and vocabularies, data licensing and accessibility, etc. (summarised in Table 3), to assess the level of data interoperability among the most important international databases and the compliance with the main environmental legislative tools.

In some cases, the same dataset can be found in several repositories due to the different purpose and coverage of each one (i.e., biodiversity/cetacean, European/national), but also because of the existing connections among them (e.g., OBIS-EMODnet, see Correia et al, 2019, 2022).

Few repositories have been created to respond to legislative obligations, so the contribution of data to them is mandatory and provided by institutional bodies (e.g., REPORTNET, SIC ISPRA). However, most repositories are based on the autonomous contribution by individual researchers, associations, or agencies that own scientific data (e.g., GOOS, GEOBON, MDA, DASSH).

Similarly, data providers for some repositories (e.g., GEOSS, data.europa.eu, SIC ISPRA) are national and international institutional bodies, and scientific agencies, but in most cases the contribution can come from any type of research entity. For this reason, some databases have been improving their catalogues through funded projects aimed to support data standardisation and upload (e.g., Intercet, EMODnet Biology).

Indeed, specific procedures often need to be followed in order to make data compliant to each database, particularly for those at global/European level. Some data repositories have implemented the standards of the INSPIRE Directive (EMODnet, SEXTANT, JCDP, MEDIN, DASSH, NNB ISPRA), while different data formats are required among the other repositories. For this reason, almost all databases have integrated quality control tools or validation processes before the final dataset upload.

Table 3. Summary table of the characteristics of the main repositories

	Repository	*Topic*	*Link to policies*	*Link to other DBs*	*Providers*	*Standards*	*Validation/ quality control*	*Upload*	*Data access/ Licence*	*DOI*
Global	**OBIS**	Biodiversity	UN, CBD	GBIF, EMODNET, DASSH	Anyone	DwC – EML - vocabularies	QC tools	IPT (GBIF)	Restricted /CCBY	Yes
	GOOS	Marine	-	OBIS	Anyone	Specific	GeoNode map	Direct/Geo Node	Restricted /CCBY	Yes
	GEOSS	Biodiversity	UN SDG	GBIF, OBIS	Institution	Discovery Access Broker	Discovery Access Broker	Discovery Access Broker	Restricted /Data CORE	No
	GEOBON	Biodiversity	UN, MSFD, WFD, Nature Directives	OBIS, GBIF, GEOSS, GOOS	Anyone	Specific, GBIF taxonomy	Yes	Direct	CCBY	Yes
	GBIF	Biodiversity	-	OBIS, EMODNET	Institution	DwC – EML - vocabularies	Yes	IPT	Open access	Yes
	IWC	Cetaceans	ACCOBAMS IUCN	-	Institution	Specific templates	No	Direct	Only registered	No
European	**EMODNET**	Marine	MSFD, INSPIRE	GBIF, OBIS, DATA.EU	Anyone	DwC – EML- vocabularies	Biockeck, OBIS QC tools	IPT (GBIF)	Restricted /CCBY	Yes
	REPORTNET	Biodiversity	UNEP, Nature Directives, INSPIRE	-	Institution	Data Dictionary	Data Dictionary QC	Data Dictionary	Only registered	No
	DATA.EUROPA.EU	Biodiversity	Nature Directives, MSFD, INSPIRE	EMODNET	Institution	Data Catalogue vocabulary	Data Catalogue vocabulary	Direct	Open access/ Reuse Directive, CCBY	Yes

	MDA	Marine	-	GBIF, OBIS, EMODNET	Anyone	Specific templates	Templates	Direct	Restricted /CCBY	Yes
Regional	**NETCCOBAMS**	Cetaceans	ACCOBAMS	-	Anyone	Not available	Not available	Direct	Only registered	No
	INTERCET	Cetaceans	ACCOBAMS	-	Anyone	Specific templates	Web-GIS	Direct	Only registered	No
	SEXTANT	Marine	INSPIRE	-	Anyone	Specific templates	No	Direct	Restricted /CCBY	Yes
	JCDP	Cetaceans	JNCC, OSPAR, ICES, INSPIRE	MEDIN	Anyone	MEDIN data standards - vocabularies	INSPIRE validator, Data Portal	GeoNetwork	Restricted /CCBY	No
National	**UK - MEDIN**	Marine	INSPIRE	GBIF, OBIS, EMODNET	Anyone	Discovery metadata standard - vocabularies	Metadata Editor, Maestro, Schematron	Direct	Restricted /CCBY	No
	UK - DASSH	Marine	INSPIRE	GBIF, OBIS, EMODNET, MEDIN	Anyone	Discovery metadata standard	No	Direct	Restricted /CCBY	No
	IT- SIC ISPRA	Marine	MSFD	REPORTNET	Institution	Specific templates	Validation tool	Direct	Only registered	No
	IT - NNB ISPRA	Biodiversity	Nature directives, INSPIRE	GBIF	Anyone	Specific templates	No	BioCASe	CCBY	No
	IT - JCDM	Cetaceans	-	-	Anyone	Specific templates	No	Direct	CCBYNC	No

UN: United Nations
SDG: Sustainable Development Goals
CBD: Convention of Biological Diversity
MSFD: Marine Strategy Framework Directive
WFD: Water Framework Directive
IUCN: International Union for Conservation of Nature
UNEP: United Nations Environment Programme
JNCC: Joint Nature Conservation Committee
ICES: International Council for the Exploration of the Sea
DwC: Darwin Core
EML: Ecological Metadata Language
QC: Quality Control
IPT: Integrated Publishing Toolkit
CC: Creative Common Licences

In most of the cases, the data providers who share their datasets in these repositories keep the ownership of the data, and traceability of the source can also be assured by assigning a DOI to the datasets.

The access to the datasets can be decided on the basis of different licence types. The most used licences for datasets are the Creative Commons Licences, which allow freedom: to share (to copy, distribute and use the database); to create (to produce works from the database); to adapt (to modify, transform and build upon the database).

- CC-0 (public domain): The owner waives any copyright they might have over the data(set) and dedicate it to the public domain, and cannot be held liable for any (mis)use of the data. Although CC-0 doesn't legally require users of the data to cite the source, it does not take away the moral responsibility to give attribution, as it is common in scientific research.
- CC-BY (attribution): any public use of the database, or works produced from the database, must be attributed in the manner specified in the licence. For any use or redistribution of the database, or works produced from it, the licence of the database must be made clear to others and any notices on the original database must be kept intact.
- CC-BY-NC (non-commercial) is equivalent to CC-BY but does not allow commercial use. It can be problematic when the data is re-used in scientific journals.

In the following part of this section, the procedures to standardise datasets in the formats required by each repository are summarised, based on the instructions retrieved on the relative websites, guidelines, template files and documentation from specific training courses. In case of data required by EU policies obligations, such as for the MSFD, each Member State has implemented a national database system; here, only the Italian repository is reported as an example.

6.1 Global

6.1.1 Ocean Biodiversity Information System

OBIS https://manual.obis.org/data_standards.html

OBIS Data Sources are the authors, editors, and/or organisations that have published one or more datasets through OBIS. OBIS only harvests data from their recognised nodes.

OBIS has championed the use of international standards for biogeographic data. Without this agreement, OBIS would not have been able to build a large central database. The following standards are required:

- **Darwin Core (DwC)**: is a body of standards for biodiversity informatics. It provides stable terms and vocabularies intended to facilitate the sharing of information about biological diversity. DwC terms correspond to the column names of a dataset, related to the Classes Taxon, Identification, Occurrence, Record level, Location, Event, Material Sample.

OBIS currently requires eight DwC terms: occurrenceID, eventDate, decimalLongitude, decimalLatitude, scientificName, scientificNameID, occurrenceStatus, basisOfRecord.
Event core sampling dates are required in the ISO 8601 format. All species names need to be matched against an authoritative taxonomic register, such as the World Register of Marine Species.

- **Ecological Metadata Language**: OBIS uses the Ecological Metadata Language (EML) as its metadata standard, which is specifically developed for the earth, environmental and ecological sciences. EML is implemented as XML and for OBIS the following 4 terms are the bare minimum: Title, Citation, Contact, Abstract.

More EML terms can be used to describe datasets, such as a *Shortname* that serves as an identifier for the resource, or *Coverage* (for indication of temporal, spatial and species coverage), and *Licence*. The recommended licences for datasets published in OBIS are the Creative Commons Licences, with preference for CC-0.

- **Darwin Core Archive** and dataset structure: Darwin Core Archive (DwC-A) is the standard for packaging and publishing biodiversity data using Darwin Core terms. It is the preferred format for publishing data in OBIS. A DwC-A contains a number of text files, including data tables formatted as CSV. Its conceptual data model is a star schema with a single core table, for example containing occurrence records or event records, at the centre of the star (Figure 10). Extension tables can optionally be associated with the core table. A one-to-many relationship connects the core and extension records, so each core record can have zero or more extension records linked to it, and each extension record must be linked to exactly one core record. Besides data tables, a DwC-A also contains two XML files: one describing the archive and data file structure (meta.xml), and the other containing the dataset metadata (eml.xml).

With the project “Expanding OBIS with environmental data OBIS-ENV-DATA” (Figure 10), OBIS introduced an extension, ExtendedMeasurementOrFact (eMoF) which extends the previous one with 4 new terms: *occurrenceID*, which links measurement records in the eMoF extension to occurrence records in the Occurrence extension, *measurementTypeID*, *measurementValueID*, *measurementUnitID* to standardise the measurement types, values and units. Thanks to these terms it is possible to store a variety of measurements and facts linked to either events or

occurrences: organism quantifications, facts documenting a specimen, abiotic measurements, facts documenting the sampling activity. The three “measurement” terms should be populated using controlled vocabularies referenced using Unique Resource Identifiers (URIs). OBIS recommends using the internationally recognized NERC Vocabulary Server, developed by the British Oceanographic Data Centre (BODC) (https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/).

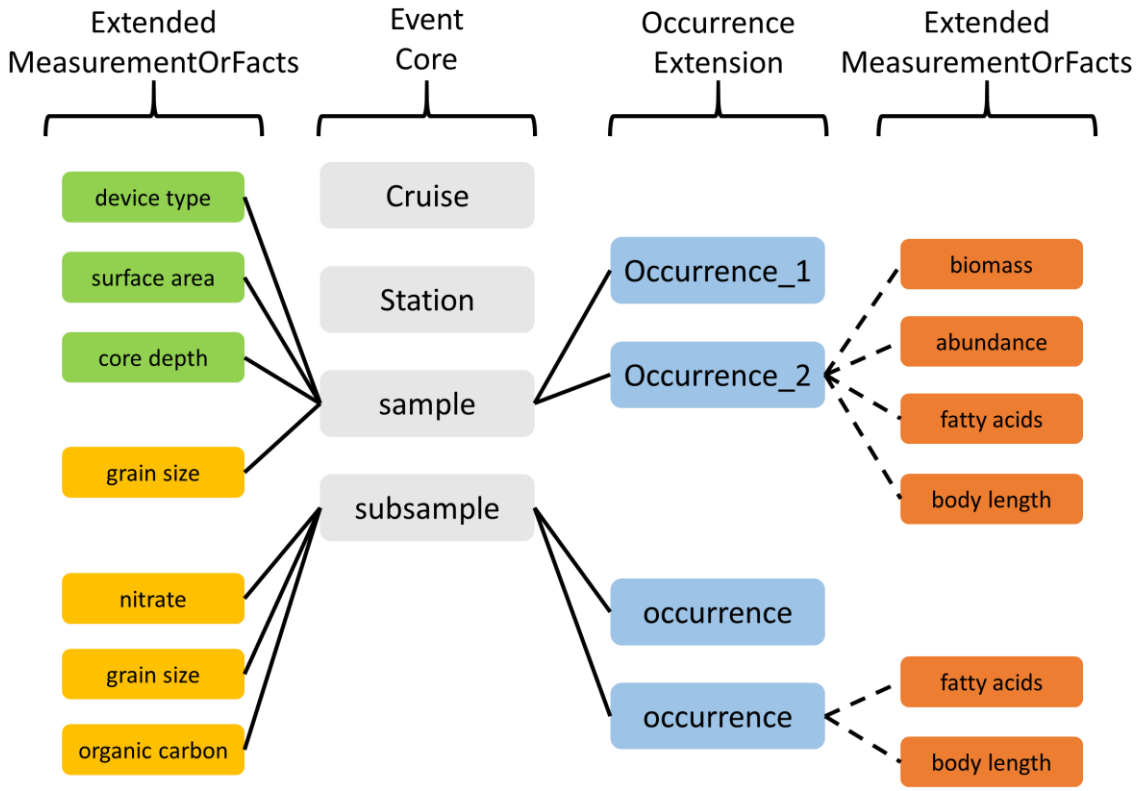


Figure 10 Overview of an OBIS-ENV-DATA format. Sampling parameters, abiotic measurements, and occurrences are linked to events using the eventID. Biotic measurements are linked to occurrences in the ExtendedMeasurementOrFact Extension (Source: OBIS).

OBIS ignores records that do not meet a number of standards. OBIS checks, rejects and reports the data quality back to the OBIS nodes. A number of Quality Control tools are developed to help data providers and OBIS nodes, for species names and geography and data format.

OBIS nodes publish these data on their specific Integrated Publishing Toolkit (IPT), which are harvested by central OBIS (Figure 11). The IPT is a standard software for exchanging data, used in biological data management and is developed and maintained by the Global Biodiversity Information Facility (GBIF). The IPT software assists the user in mapping data to valid DwC terms

and in archiving and compressing the DwC content with: (i) a descriptor file (meta.xml) that maps the core and extensions files to DwC terms, and describes how the core and extensions files are linked, and (ii) the metadata file (eml.xml), which contains the dataset metadata. All these components, compressed together as a .zip file, comprise the DwC-A.

Specific indications for submitting data directly to EUROBIS are available at https://www.eurobis.org/submit where a detailed form can be used to describe the dataset and also to archive and share the dataset by creating a formal data publication and assigning a DOI.

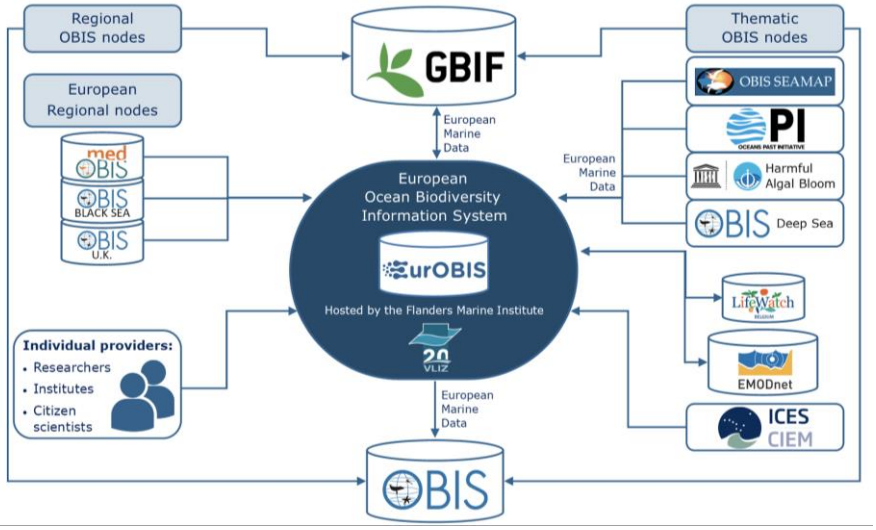


Figure 11 Graphical representation of the network of repositories providing data to the European Ocean Biodiversity Information System (Source: online training course by EMODnet Biology).

6.1.2 BioEco Metadata Portal - Global Ocean Observing System

GOOS https://iobis.github.io/bioeco-docs/

GOOS encourages visitors of the portal and registered users to contribute ocean observing programmes, regardless of their scale or developmental stage. All metadata and information available in the portal are uploaded through the backend platform, the GOOS BioEco GeoNode, to ensure a user/data provider driven approach. In GeoNode, registered users can upload, edit and manage contributed monitoring programme details and the associated metadata themselves. It can be done independently, using the Portal and GeoNode documentation available to all users and visitors, which contains step-by-step instructions and guidelines on how to upload, edit or manage the metadata, supported by videos and examples.

To contribute a monitoring programme to the portal, the user should first register as a new user or metadata data provider. From the Profile page, personal information and options can be added, such as a user picture, organisation, position, location. Interaction with other users is possible by Message User or Invite Users.

GeoNode offers two ways in which new monitoring programmes can be added, which depend on the format of the spatial information of the programme: the preferred one is by uploading a shapefile, setting permissions for other GeoNode users to view, download and edit the uploaded data. The other option is by using the GeoNode map tool, which allows creating a new layer with its own geometry type and name. The empty layer can then be edited by adding an attribute number, features, coordinates.

Registered users have exclusive access to edit and update the metadata of their assigned programme. Metadata providers are encouraged to regularly update their programme metadata to keep the BioEco portal current and representing the true current state and status of ocean observation. Through the Editing Tools, the spatial and metadata layer can be edited, removed or updated. If the editing tools are not available to the owner, it is likely due to the default setting of admin as owner. In this case, users should contact admin at s.van-der-wal@unesco.org / helpdesk@obis.org to have the ownership changed and gain access to the editing tools.

Once the programme is uploaded, a title and abstract with a short introduction on the monitoring programme should be provided. The Owner will be indicated as the metadata provider that uploaded the metadata indicated either by their email address or username. The field for DOI will be completed by the admin and should be left open. A space is provided for free-text keywords that are useful for user filtering options. The licence applicable to the data/metadata of the programme should be provided along with the temporal extent of the programme, a Uniform Resource Locator (URL) or link to the main project website, and indication if the data is published on the OBIS database. The applicable regions where the monitoring is done should be selected.

The GOOS EOV observed/monitored should be selected from a drop-down menu, along with their level of readiness (see Figure 12 for details), which is used for evaluating new components for possible inclusion in the global ocean observing systems.

FRAMEWORK PROCESSES BY READINESS LEVELS

Readiness Levels	Requirements Processes	Coordination of Observational Elements	Data Management & Information Products
Mature			
Level 9 "Sustained"	Essential Ocean Variable: • Adequate sampling specifications • Quality specifications	System in Place: • Globally • Sustained indefinitely • Periodic review	Information Products Routinely Available: • Product generation standardized • User groups routinely consulted
Level 8 "Mission qualified"	Requirements "Mission Qualified:" • Longevity/stability • Fully scalable	System "Mission Qualified:" • Regional implementation • Fully scalable • Available specifications and documentation	Data Availability: • Globally available • Evaluation of utility
Level 7 "Fitness for purpose"	Validation of Requirements: • Consensus on observation impact • Satisfaction of multiple user needs • Ongoing international community support	Fitness-for-Purpose of Observation: • Full-range of operational environments • Meet quality specifications • Peer review certified	Validation of Data Policy • Management • Distribution
Pilot			
Level 6 "Proven capacity"	Requirement Refined: • Operational environment • Platform and sensor constraints	Implementation Plans Developed: • Maintenance schedule • Servicing logistics	Demonstrate: • System-wide availability • System-wide use • Interoperability
Level 5 "Verification"	Sampling Strategy Verified: • Spatial • Temporal	Establish: • International commitments and governance • Define standardized components	Verify and Validate Management Practices: • Draft data policy • Archival plan
Level 4 "Trial"	Measurement Strategy Verified at Sea	Pilot project in an operational environment	Agree to Management Practices: • Quality control • Quality assurance • Calibration • Provenance
Concept			
Level 3 "Proof of concept"	Proof of Concept via Feasibility Study: • Measurement strategy • Technology	Proof of Concept Validated: • Technical review • Concept of operations • Scalability (ocean basin)	Verification of Data Model with Actual Observational Unit
Level 2 "Documentation"	Measurement Strategy Described • Sensors • Sensitivity • Dependencies	Proof of Concept: • Technical capability • Feasibility testing • Documentation • Preliminary design	Socialization of Data Model • Interoperability strategy • Expert review
Level 1 "Idea"	Environment Information Need and Characteristics Identified: • Physical • Chemical • Biological	System Formulation: • Sensors • Platforms • Candidate technologies • Innovative approaches	Specify Data Model • Entities, Standards • Delivery latency • Processing flow

Figure 12 Framework processes for varying levels of readiness of the GOOS EOVs (Source: A Framework for ocean observing, https://unesdoc.unesco.org/ark:/48223/pf0000211260).

Additional measurements taken / observations made of the BioEco EOV(s) selected under GOOS EOVs, should be provided under EOV subvariables and EBVs. Non-BioEco EOVs, such as Physics,

Biochemistry, and Cross-disciplinary EOVs, should be provided under GOOS Physics, Biochemistry, and Cross-disciplinary EOVs.

Newly contributed programmes and updated programme metadata are automatically incorporated into the GOOS BioEco portal, where users can then observe, compare and connect with programmes, researchers, and institutions of interest. Interactive maps can also be created using the GeoNode interactive map tool (©OpenStreetMap contributors).

6.1.3 Group on Earth Observations System of Systems

GEOSS https://www.earthobservations.org/geoss.php#gci

The GEOSS Portal is implemented and operated by the European Space Agency and offers a single Internet access point to global Earth observation data, information and knowledge for users with different backgrounds and from different disciplines. Through the Portal, the GEOSS Platform connects the users to the collections of databases and other portals and provides reliable, up-to-date and user-friendly information (Figure 13).

The GEO Discovery and Access Broker (GEO DAB) is the primary mechanism by which all data and information is discovered and accessed. The GEO DAB implements the necessary mediation and harmonisation services through Application Program Interfaces (APIs), which allow data providers to share resources without having to make major changes to their technology or standards. GEO DAB publishes three types of APIs:

- Standard Web Services: standard interfaces which comply with geospatial standards such as OGC, CSW, WCS, WMS, WFS, OAI-PMH, etc.
- Client-side APIs: high level client-side library, designed and developed to facilitate the development of web and mobile applications, presently available as a Javascript library.
- Server-side APIs: a set of server-side APIs, including OpenSearch and RESTful APIs.

These APIs are used by the GEOSS Portal to discover, visualise and access GEOSS resources. Besides, the APIs are used to define and utilise specific views of the GEOSS resources.

Data providers are constantly being added and brokered, according to user needs, thematic, and geographic balance of the data and relevance of resources shared. New data providers should follow the GCI Registration and Brokering Process. After having discussed with GEO Secretariat, the data provider should fill a set of administrative information in the yellow pages at:

"https://docs.google.com/forms/d/e/1FAIpQLSf5rNY1k2y1Dm1aRl4l7ShimYKrvucTc3DFp-uaIl8toWuD8A/viewform?c=0&w=1" and provide the logo of their organisation; the GEO Secretariat team validates the provided information and stores it, and then connects the data provider to the GEO DAB team to start the brokering process.

The data provider sends the Web service endpoint URL of the data catalogue/information to the GCI/GEO DAB team, which runs interoperability tests and sends results to the provider and the GEO Sec. Within the process, the GCI/GEO DAB team and the provider work together to address possible issues. GCI/GEO DAB team adds the provider system to GEOSS brokered sources and informs the GEO Secretariat, with the provider in cc. After this final step, the data catalogue is accessible from the GEOSS portal.

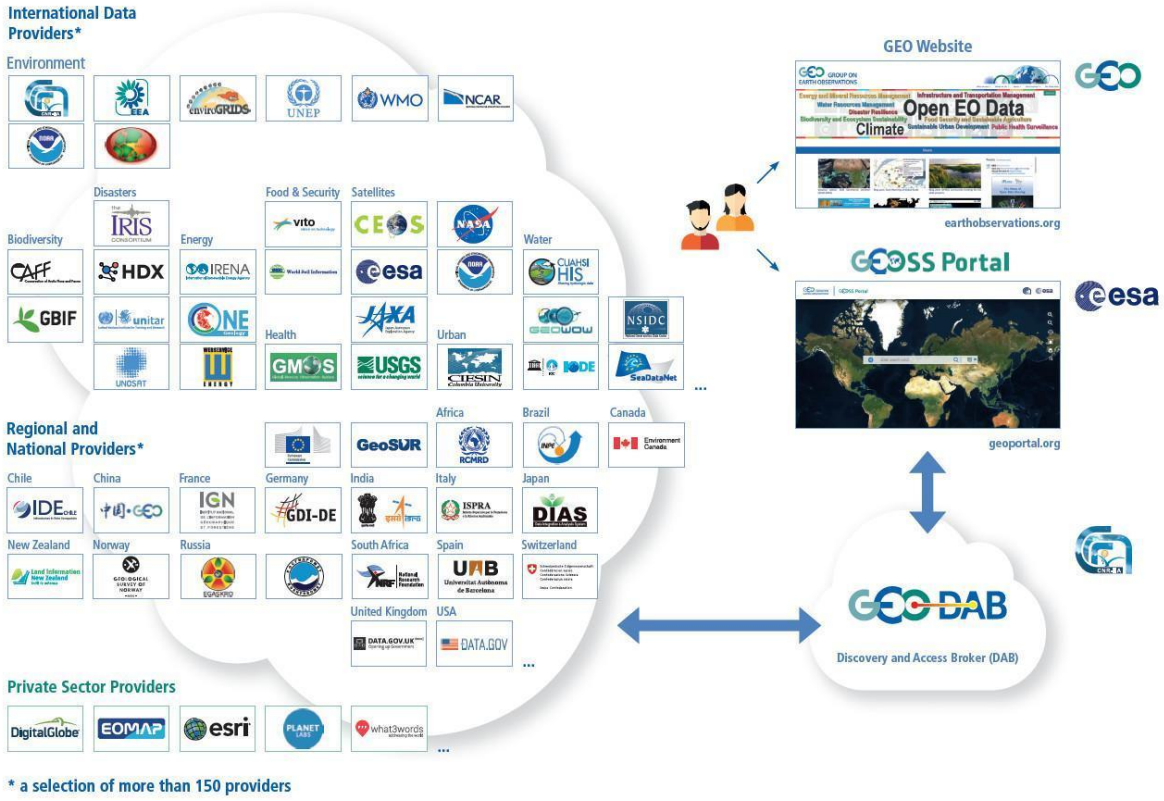


Figure 13 Graphical representation of the functioning of the GEOSS system (Source: GEOSS).

6.1.4 European Biodiversity Observation Network

EUROPABON https://monitoring.europabon.org/monitoring/

Anyone joining a BON as a member can also upload EBV datasets for sharing with others in the EBV Data Portal. To request a registration, an email should be sent to ebvportal@idiv.de.

The EuropaBON Monitoring database is open for consultation purposes only. The interface API is based on REST and JSON and is used by the portal to discover, visualise and access resources. Anyone part of a European-wide monitoring network can join the database to contribute data by contacting info@europabon.org. Each contributor to the database will receive a personalised username and password to login to input data of specific integration initiatives. For each integration initiative, it is necessary to fill in all the fields within three different tables (Figure 14):

- The **IntegrationNodes** table collects relevant information for each of the nodes within an integration activity. Each node is an institution integrating data and/or generating 'products' (EBVs or other indicators). These data and products can flow into another upper-level integration node (e.g., European-level node) or represent 'final products'.
- The **BiodiversityData** table includes information about the field biodiversity monitoring initiatives underneath integration nodes (field data collection). Users can create an entry for each monitoring program underneath an integration node, but grouping those that follow the same sampling protocol so their data are comparable and uniform.
- The **DataSteams** table collects information about how the data is processed, how it flows between lower-level and upper-level integration nodes, how data flows between European-level Integration Nodes and Policy and the final "products" (EBVs, indicators) that have the potential to inform policy.

Once the contributor has identified how many nodes the integration initiative has, and how many BiodiversityData programs feed into those nodes, these can be uploaded into the database. Within each table, the contributor will find multiple information fields to be filled; each field is accompanied by an explanation of what type of information it seeks to collect. The common field called "Target species" requires uploading a .csv file with the list of taxa, which will be normalised against the GBIF taxonomy. The contributor has to start by filling the "BiodiversityData" form ("Add new BiodiversityData"), then fill in the form of "Integration nodes", starting with the Sub-European-level nodes, as these will be those directly linked to BiodiversityData. Once all Sub-European nodes are entered into the database, the contributor can enter the European-level node. Finally, the information about how the data flows between nodes can be entered by using the "DataStreams" form ("Add new DataStream"). Within each of these DataStreams the contributor has to provide information about the data flowing between nodes and about how the

data is processed or which products are generated (Species distribution maps, species change index, Figure 14).

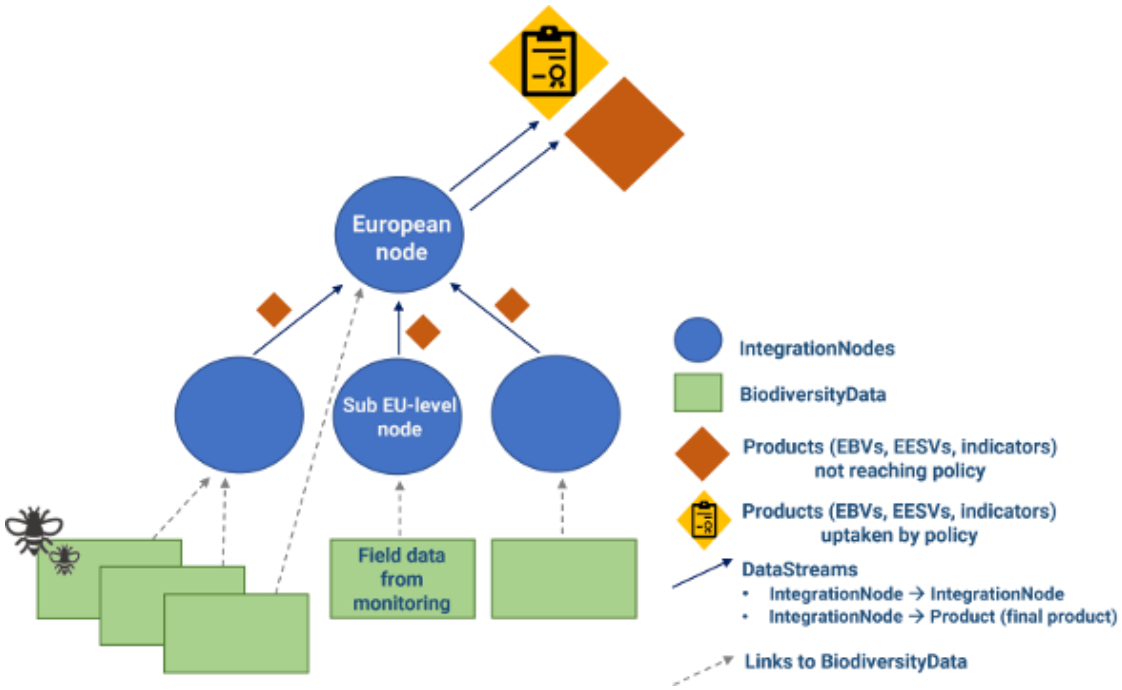


Figure 14 Graphical representation of the EuropaBON monitoring database (Source: europabon.org).

6.1.5 Global Biodiversity Information Facility

GBIF https://www.gbif.org/publishing-data

GBIF is not a data repository, but it indexes thousands of datasets through a network of data-publishing institutions worldwide. For example, the OBIS nodes need to become a data publisher in GBIF in order to make data available through this facility. Organisations that want to share their data through GBIF have to register to request endorsement as a data publisher.

Publishing process (Source: https://ipt.gbif.org/manual/en/ipt/latest/how-to-publish)

GBIF.org supports the publication of four classes of datasets using widely accepted biodiversity data standards (DwC-A, EML). GBIF ensures that each dataset is associated with a unique DOI to

streamline data users' citation of these resources. The classes start simply and become progressively richer, more structured and more complex.

Resources metadata: at its simplest level, GBIF.org allows institutions to create datasets describing undigitized resources like those in natural history and other collections. All three other dataset classes include this basic information, but this 'metadata-only' class offers researchers a valuable tool for discovering and learning about evidence not yet available online. GBIF uses the EML as its metadata standard. One can simply use the metadata editor in the IPT to populate the metadata or transform it into an XML metadata document, which is based on the EML, and upload it to the IPT.
Required metadata fields are: title, description, publishing organisation, type, licence, contact(s), creator(s), metadata provider(s).
Recommended metadata fields are: sampling methodology, citation.

Checklist data: Datasets can also provide a catalogue or list of named organisms, or taxa. While they may include additional details like local species names or specimen citations, these 'checklists' typically categorise information along taxonomic, geographic, and thematic lines, or some combination of the three. Checklists function as a rapid summary or baseline inventory of taxa in a given context. The data needs to be transformed into a table structure using DwC term names as column names. Template files for transformation are provided.
Required DwC fields are: taxonID, scientificName, taxonRank.
Recommended DwC fields are: kingdom, parentNameUsageID, acceptedNameUsageID.
Alternatively, if data is stored in a supported database, an SQL table (view) can be written using DwC column names. The basic descriptive information of the metadata is also required (see Resource Metadata).

Occurrence data: Other datasets published through GBIF.org have sufficiently consistent detail to contribute information about the location of individual organisms in time and space—that is, they offer evidence of the occurrence of a species (or other taxon) at a particular place on a specified date. Occurrence datasets make up the core of data published through GBIF.org, and examples can range from fossils in natural history collections, observations by citizen scientists, or data gathered from remote-sensing satellites. Occurrence records in these datasets sometimes provide only general geographic information, sometimes simply identifying the country, but in many cases more precise locations and geographic coordinates support fine-scale analysis and mapping of species distributions. The data need to be transformed into a structured table using DwC term names as column names. Template files for transformation are provided.

Required DwC fields are: occurrenceID, basisOfRecord, scientificName, eventDate
Recommended DwC fields are: taxonRank, kingdom, decimalLatitude, decimalLongitude, geodeticDatum, countryCode, individualCount
Alternatively, if data is stored in a supported database, an SQL table (view) can be written using DwC column names. The basic descriptive information of the metadata is also required (see Resource Metadata).

Sampling-event data: Datasets sometimes provide greater detail, not only offering evidence that a species occurred at a given location and date, but also making it possible to assess community composition for broader taxonomic groups or even the abundance of species at multiple times and places. These quantitative or sampling-event datasets typically derive from standard protocols for measuring and monitoring biodiversity, like bird censuses or marine samplings. By indicating the methods, events and relative abundance of species recorded in a sample, these datasets improve comparisons with data collected using the same protocols at different times and places—in some cases, even leading researchers to infer the absence of particular species from particular sites.
These datasets include the same basic descriptive information included under Resource metadata and the same standard elements as in Occurrence Data. Data need to be transformed into two tables using DwC term names as column names: one table of sampling events and another table of species occurrences derived from (associated to) each sampling event. Template files for transformation are provided, showing the required fields for sampling event and species occurrence record (same as the Occurrence data).
Required DwC fields are: eventID, eventDate, samplingProtocol
Recommended DwC fields are: sampleSizeValue, sampleSizeUnit, parentEventID, samplingEffort, locationID, decimalLatitude, decimalLongitude, geodeticDatum, footprintWKT, footprintSRS, countryCode, occurrenceStatus
After choosing the right class of dataset and transforming the data in the required tables, data are uploaded to the IPT by requiring an account on a data hosting centre (DHC) that uses GBIF's IPT software and is located in the data owner country (https://ipt.gbif.org/manual/en/ipt/latest/data-hosting-centres). The DHC provides data publishers with a hosted IPT account that enables them to manage and publish datasets through GBIF.org. The publisher organisation must be registered with GBIF and added to their IPT by the IPT administrator (see Figure 11).
The data are subsequently mapped (e.g., Checklist Data gets mapped to the Taxon Core, Occurrence Data gets mapped to the Occurrence Core, Sampling Event Data gets mapped to the

Event Core). The resource metadata must be filled in using the IPT's metadata editor or upload the metadata file. Finally, the dataset is published and registered with GBIF.

6.1.6 Global Ship Strikes Database - International Whaling Commission

IWC https://portal.iwc.int/login

Each *Country administrator* gives authorisation to the *Organisation administrators*, generally a representative of public or private bodies responsible for the shared data. The Organisation can add several *Users* and control the data they upload. After signing in the portal (https://portal.iwc.int/login) each contributor can add specific data on ship strikes in the section 'Report a ShipStrike' where information on the species, date and location of the event have to be provided. Other data on cetacean observations can be also added into the portal as follows.
Data collected can be added to the IWC portal with the following specifications:
Type of data: *Systematic Surveys*
Name: route or project name
Description: e.g., Fixed line transect...
contact: Organisation administrator
Once the type of data is complete, the user has to save and obtain a "reference number".
Then select ADD:
Cetacean Dataset, database:
Name: same as above
Description: e.g., Systematic surveys with "reference number"...
SHORT NAME: Project-Area-Year
YEAR; DATE: period
Type: sightings
Species: SELECT ALL SPECIES by adding in parentheses the number of sightings for each one
Cetacean database considering all sightings together
Number of Records: total sightings
Manager: User
Contact: Organisation administrator

6.2 European

6.2.1 European Marine Observation and Data Network

EMODNET https://www.emodnet-biology.eu/

EMODnet Biology requires the same data format as OBIS repository, as datasets are flowing into it. A guided online course provides examples on how to transform different types of datasets in the required formats. Here is a summary of the standardisation process required for EMODnet Biology.

A first step is to provide a technical **metadata** of the dataset, that is stored in a separate database and include different information (e.g., instruments, sampling descriptors, and protocols), as well as the contact of the person providing the metadata, the dataset title in English, the contact of the data creator, the dataset citation, the licence type, the abstract, the geographical, temporal and taxonomic coverages, and the functional groups.

Data processing in EMODnet Biology is based on three blocks: data structure, field nomenclature, and content.

The **data structure** follows the OBIS-ENV data format (https://www.eurobis.org/data_formats), with a core table (Event table) and extension tables (Occurrence table and Extended Measurement or Facts - eMoF), each one with a minimum number of required fields. Therefore, a dataset organised as a flat table has to be split into three tables. The Event core table stores sampling and observation information (e.g., time, location, depth), the Occurrence table stores the occurrence details (e.g., taxonomy, identification), and the eMoF table stores the sampling information and additional biological/abiotic measurements (i.e., organism quantifications, species biometrics, abiotic measurements, facts documenting the sampling activity).

When a flat one-table database is standardised for EMODnet Biology, the following mandatory fields have to be included in the three DwC tables:

- Event table - eventID, datasetName, institutionCode, eventDate, decimalLatitude, decimalLongitude.
- Occurrence table - eventID, occurrenceID, scientificName, scientificNameID, occurrenceStatus, basisOfRecord.
- eMoF table - eventID, measurementType, measurementTypeID, measurementValue, measurementUnit.

These three tables have to be standardised using the **field nomenclature** of the DwC terminology (https://dwc.tdwg.org/terms/). The tables are connected by the codes EventID (https://dwc.tdwg.org/terms/#dwc:eventID) and OccurrenceID (https://dwc.tdwg.org/terms/#dwc:occurrenceID), creating an event hierarchy as "one-to-many" relationships.

The **content** includes data, time, latitude and longitude, EventID and OccurrenceID, taxonomic information (https://www.itis.gov/ws_lsidApiDescription.html) and has to follow standards, such as the ISO 8601 standard for sampling dates in the Event core. The on-effort transect events include an approximate representation of the survey track in WKT format, and are indicated as line string in the *footprintWKT* field. The coordinates of the centre of the segment are added to the *decimalLongitude* and *decimalLatitude* terms and the distance between this point and the start or end of the line is included in the *coordinateUncertaintyInMeters* term. Occurrence locations are added as Points and coordinates, with additional information specified in the Occurrence Table. Information on taxonomy has to be related to WoRMS (Word Register for Marine Species, https://www.marinespecies.org) to guarantee track of the taxonomy changes, and the WoRMS Taxon Match Tools (that allows easy match for multiple taxa presence in a database) could be also used to standardise taxonomic data. Other parameters can be added to the eMoF table depending on the dataset and consistency of vocabulary is guaranteed thought the use of the NERC Vocabulary Server (NVS) developed by the British Oceanographic Data Centre (BODC), as recommended for semantic standardisation (https://vocab.nerc.ac.uk/search_nvs/).

An example of data collected within the CETUS network and formatted for EMODnet Biology is reported by Correia et al. (2019). Figure 15 describes the hierarchy of the three DwC Tables and shows for each category the fields that can be included according to DwC terms.

Once the formatting process is terminated, it is possible to verify if the dataset meets the requirements using the LifeWatch-EMODnet Biology Quality Control tool (BioCheck https://rshiny.lifewatch.be/BioCheck/) and to eventually correct mistakes. All the tools developed by OBIS (e.g., tools to check required fields and points on land) can also be used to control the quality of the formatted dataset.

Once checked, the dataset is included in the EMODnet Biology harvest and published on the IPT – Integrated Publishing Toolkit of GBIF. IPT is adopted by OBIS and EMODnet Biology for harvesting data (see Figure 11). All datasets judged fit for harvesting will then be processed by

automated procedures and become available through the EMODnet Biology Portal (both the viewer and the download toolbox). New IPT resources will be created on the EurOBIS IPT, from where the dataset can be accessed by OBIS and GBIF. Finally, it is recommended to assign a DOI to the dataset, either by the provider of the dataset or by the EMODnet Biology team.

Data collected by the FLT MED NET have been provided to this repository (Arcangeli et al, 2018, https://ipt.vliz.be/eurobis/resource?r=cet_flt_cbar_14-18; Scuderi and Martìn Moreno, 2020, https://ipt.vliz.be/eurobis/resource?r=straitgib_cetacea) and are now available through OBIS (https://obis.org/dataset/d5847ecb-6f9b-4599-888a-461cb26f8018; https://obis.org/dataset/f5005f2d-c05b-436f-9d1a-5dbc186ee4a8).

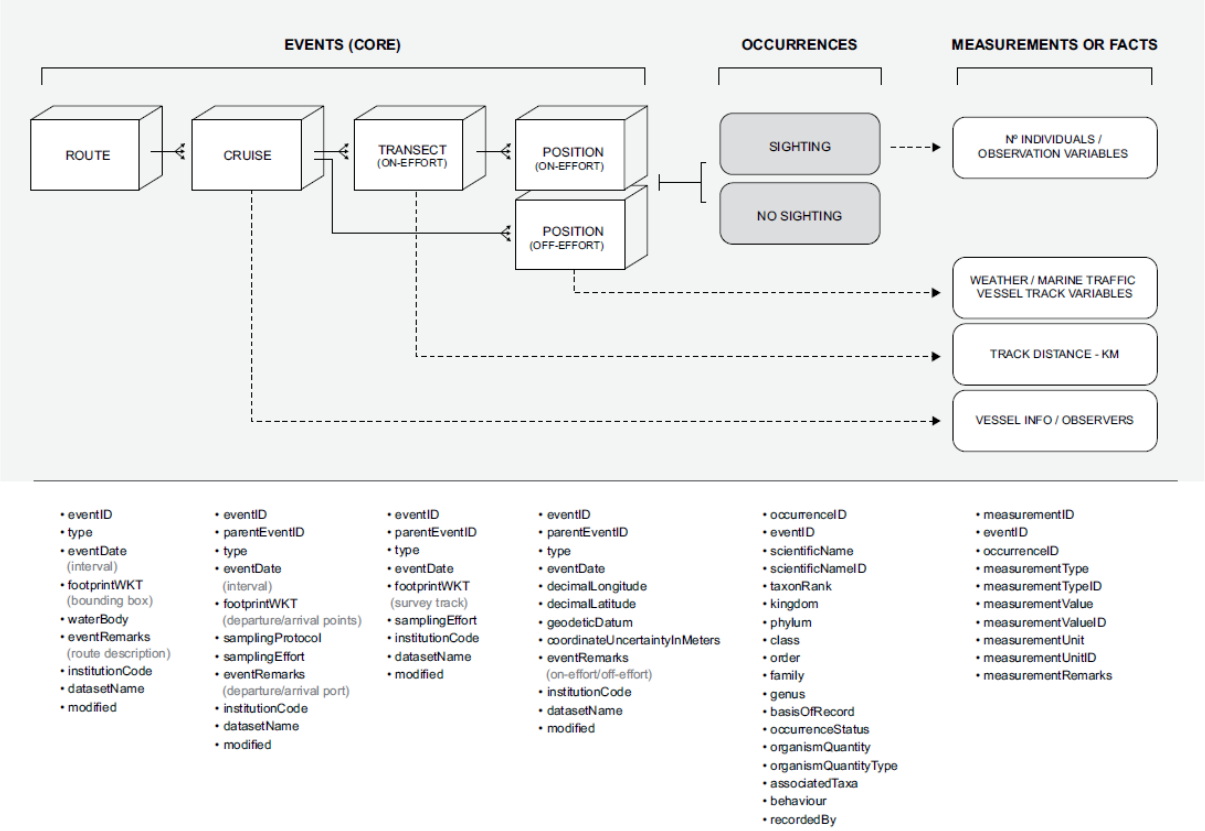


Figure 15 Graphical representation of the hierarchy of the three DwC Tables for data formatting process (Source: Correia et al, 2019).

6.2.2 Reportnet – European Environmental Agency

REPORTNET http://dd.eionet.europa.eu/documentation/doc1

Reportnet is Eionet's infrastructure for supporting and improving data and information flows, based on a set of interrelated tools and processes (Figure 16). The **Data Dictionary** (DD) is used to support the delivery of environmental data by countries to Reportnet. The contents in the Data Dictionary are technical specifications for dataflows as well as Excel templates. The DD also introduces the possibility of simple automated validation of deliveries by countries and facilitates the development of data submission interfaces. The users of the DD are data reporters and data definers as well as dataflow administrators. Reporting obligations are usually long-lasting. The definitions found in the DD, therefore, are relatively stable, but are updated prior to a new dataflow collection.

The dataset is a collection of tables containing the reported data. Conceptually, the DD datasets are stand-alone databases. In many cases a dataset corresponds to a dataflow.

A data element in DD is a column in a table. Data elements are categorised in two classes: non-common elements belong to a certain single table, where they are defined, and they cannot be used in any other table; common elements can be used in many tables (harmonised elements, e.g., CountryCode). Data elements could be for example StationCode, StationType, Longitude, and they can be of two types: with fixed values (e.g., code lists) or with quantitative values (measured values).

These resources are managed by the GeoNetwork for spatially referenced data, using metadata standards and online editing tools (ISO19115/119/110, Dublin Core).

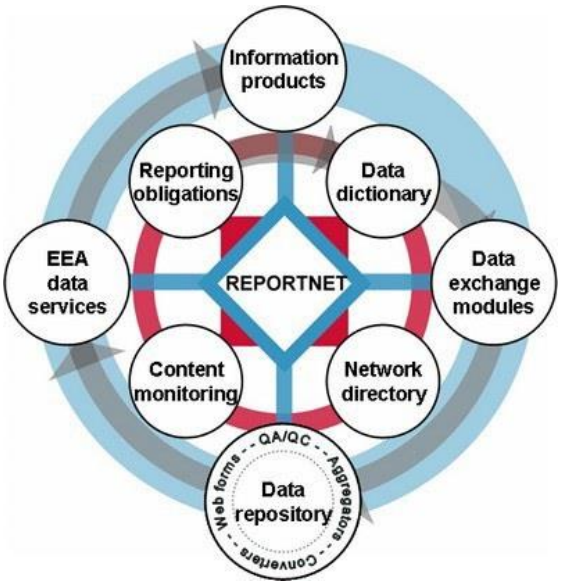


Figure 16 Graphical representation of the REPORTNET system and data flow (Source: eionet.europa.eu).

Datasets, tables and elements are defined by a set of attributes, the core set of which corresponds to ISO 11179 standard for describing data elements. The rest of the attributes are usually specific to Reportnet business rules and in any case the attribute set is flexible. The most common attribute is 'Name', which stands for the name of the defined object. Other attributes could be for example 'Definition', 'Version', etc. There are two types of attributes:
Simple attribute (name/value pair): e.g., Datatype, MaxSize/MinSize, MaxValue/MinValue
Complex attribute (a set of name/value pairs=field): e.g., SubmitOrganisation, RegistrationAuthority
Each attribute used in the DD is provided with online help for the users about its meaning and point. The whole attribute set is flexible and attributes can be added/removed from/to the system.

6.2.3 Data.europa.eu – Joint Research Centre

DATA.EUROPA.EU https://dataeuropa.gitlab.io/data-provider-manual/how-the-portal-works/

EU institutions, agencies and other bodies, and the Member States (the 'data providers') are autonomous in publishing their open data. Harvesting is the recommended method for publishers who manage their data in a data catalogue. In order to have their data harvested by data.europa.eu, publishers must fill a questionnaire available at "https://data.europa.eu/en/contact-us?type=feedback-suggestions" and provide information about the data catalogue, including: URL to interface (REST, CSW); URL to homepage; Title of the catalogue; Description of the catalogue; Publisher of the catalogue; Email address of the catalogue; Default language of the catalogues datasets; How often can/should the site be harvested (e.g., once a week); How often the site should not be harvested (e.g., scheduled maintenance). Data are processed overnight and every incoming format is transformed to DCAT-AP 2.1.1 with a hash built over every harvested dataset. This hash value is compared to the existing hash value before a dataset is potentially updated in the portal triplestore. Updates take place only when an inequality is found. The harvester is configured specifically for each harvested portal. For harvesting to take place, the source site needs to have in place one of the interfaces as described in detail in the Interface supported for harvesting section. The main supported interfaces are the following:

- DCAT-AP/Comprehensive Knowledge Archive Network (CKAN) compliant sites (for 'normal' datasets). This is the officially recommended method, based on the data catalogue vocabulary. Datasets are provided as linked data and can be represented in multiple ways,

including rdf/xml, n-triples or turtle). Environmental data are included within the ENVI data category.

- CSW/Inspire catalogue services (for geospatial datasets); with the metadata model considered as defined in the INSPIRE Technical Guidance on Discovery Services and on Metadata
- OpenSearch (GEO/EOP) (for geospatial datasets).

Most information on the portal constitutes metadata, *i.e.*, information about data (title, description, publisher, etc.), which links to the actual data, in most cases in the form of a downloadable file. The metadata is stored in the databases of data.europa.eu, where the data usually remains with the original data publisher. However, data.europa.eu is capable of storing both metadata and data. With the data provider interface (DPI) it is possible to provide the metadata and the actual data. The functioning of the data portal is shown in Figure 17.

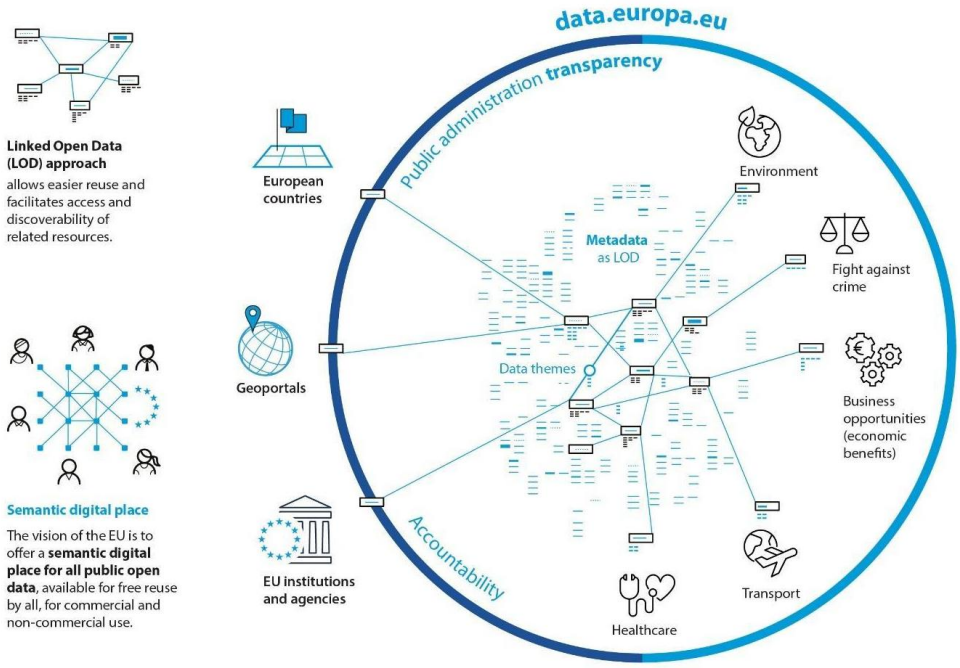


Figure 17 Graphical representation of the functioning of the data.europa.eu portal (Source: data.europa.eu).

In order to provide data, it is necessary to register by contacting the EU Publications Office, which will provide username and platform. A dataset can be created with a wizard-like form that guides the data owner through the provision of the metadata and data. The form is divided into five main steps:

- Essential Properties
- Advised Properties
- Additional Properties
- Distributions
- Dataset Overview

The form used to upload data consists of specialised input fields, supporting the various properties of DCAT-AP. In the first step, the essential metadata about the data are provided. An important property is the dataset ID, which will be used in the URL to resolve the dataset after publication (http://data.europa.eu/88u/dataset/\[dataset-id\]). In the second and third step, the remaining properties (advised and additional) of the dataset can be provided. In the fourth step, an overview of all distributions of the dataset is provided. If data is already hosted and publicly available, the URL can be provided by selecting the type 'Provide a URL'. Data can also be directly uploaded, by selecting the type 'Upload a file'. The final step provides an overview of the dataset. The publisher can edit, publish and delete datasets. Depending on whether the dataset is public or a draft, the access to the functions differs. The data provided also has to choose a licence. DOIs for datasets are minted upon request. If the publication already has a DOI, there is no need to assign it.

6.2.4 Marine Data Archive

MDA https://www.marinedataarchive.eu/archive.php?action=upload

Users need to register first to request access to the MDA (https://mda.vliz.be/account.php). After approval, the user has access to a private folder, the public folders and a general shared folder in the MDA. The root folders and subfolders are structured in a tree view: Root folders (private, public, shared); Subfolders, with a Quarantine subfolder for temporary destination for uploaded files, available for 8 months.

The user has the possibility to select a context, allowing to filter the number of folders visible depending on the context or project. There are three levels of controlled access to the folders: read only, write limited, and write full. Access is granted on the root folder and applies automatically to all subfolders and files stored under this folder. The Access rights regulate the different actions a user can perform.
To be able to use all functionalities offered in the archive, it is important to manage and describe the files in an efficient way. The more standardised and completed the metadata fields are, the more representative the search result in the Archive will be, allowing traceability and usability of

a file. Users in fact can search for other filesets based on their metadata information, then download and share: a unique link to a file of fileset can be shared with a partner or colleague with access to the folder where the file or fileset is archived.

To manage files and folders, users should use the **Upload** tool (https://www.marinedataarchive.eu/archive.php?action=upload). Each user can manage their data files as they prefer. In order to combine files in another folder without duplicating the files in the archive (project, report, product, analysis, etc.), the function 'group in a fileset' can be used. The fileset is made downloadable as a zip file.
Archiving data files is a three-step procedure: upload, describe and archive.
When uploading, the file(s) will be saved in the quarantine folder of the selected root folder.
The user needs to add a minimum of information in the metadata, that can be done for an individual file by selecting 'Edit metadata', or simultaneously for a selection of files by selecting 'apply Template'. Metadata contains information about:

- **File properties**: mandatory fields Name, Authors, Data provider (Institute), Data type.

The list of data types includes: Biotic data - In situ Instrument data - Lab instrument data - Experimental data - Geographic data - Remote sensing data - Image and movies - Model Data Software-Scripts – Documents.

- **Extra attributes**: depending on the choice of the data type, the Template will adapt automatically and display the attributes linked to a data type. Mandatory fields: Content Characteristics (e.g., biotic composition) and Temporal scope (first date – last date).
- **Description**: to store an extended description of the file content.

When all required fields are added, the user will be able to archive the data file that can be moved from the quarantine folder to the chosen destination folder. At this point, data archived can be distributed following different pathways:

- The user can choose to communicate on an archived dataset. In this case, VLIZ will create a dataset description in the information system and include a link to the file. At this moment the file is not public, but can be requested for use.
- The user can decide that his data file can be made publicly available (CC BY licence). VLIZ will create a metadata record, include a link to the archived file and request a DOI.
- The user can decide that the data in the data file should be integrated in existing online data systems. VLIZ will create a metadata record, include a link to the archived file, request a DOI, integrate the data in the system or transfer it to the data managers and add a link for download in the metadata record.

An example is the dataset of cetacean sightings in the north-east Atlantic provided by CETUS (Correia et al, 2022, https://doi.org/10.14284/547) which is open-access and linked to other data systems, such as OBIS, EMODnet, GBIF.

6.3 Regional

6.3.1 Network on the Conservation of Cetaceans of the Black Sea, Mediterranean Sea and contiguous Atlantic area

NETCCOBAMS https://hub.sinay.ai/accobams/home

This platform is built upon the technological basis provided by SINAY. A web-interface is made accessible to new users, after they request a login and password to the Secretariat to create their new accounts. NETCCOBAMS is still being implemented.

6.3.2 Intercet

INTERCET https://www.intercet.it/

In order to share data in Intercet, registration it is necessary to obtain a username and password. It is required to provide some essential data that must be standardised according to the platform to guarantee the good quality of the shared information. Contributors can provide files with the extensions .plt and .wpt, .txt or shapefiles to be uploaded to the platform, but respecting some formatting parameters. Given that most databases are in a table format, the easiest way is to create specific Excel files with required fields and transform them into shapefiles to be uploaded to the platform. The coordinate system required is decimal WGS84. Three groups of data are required: survey tracks and associated data; sighting data of the target species and associated data (species, number, coordinates); photographic images when available, to assist in the identification of the species and the individuals.

Survey tracks fields: Track code (date, organisation acronym, platform type, id); Observation Height (m); Number of observers; Speed (km/h); Track mode (acoustic, visual); Platform; Start/End track (date and time); Sea State (drop-down menu); Total number of sightings.

The table created can be joined to the Model shapefile which has the correct formatting of each field (e.g., Append function in ArcGIS).

Sighting data fields: Sighting code: (date, organisation acronym, sighting number, platform type, id); Target species (one or mixed) with number of total individuals and calves; Detection Mode (acoustic, visual); Total individuals and Young (automatic sum of 'Target species' field); Anthropic activities; Track code; Associated species; Start/End of sighting (date and time); Photos (yes if uploaded to the platform); Single waypoint (yes if the sighting has only one waypoint). In case of one waypoint per sighting, there is the need to copy each sighting twice. The table created can be joined to the Model shapefile as for the tracks.

Once the shapefiles are ready, they can be uploaded to Intercet by selecting only the .dbf, .shp and .shx); if some fields are not filled in correctly, a wizard will start to resolve the situation. It is good to upload the tracks first, and then the sightings that are related to them through the Track code field (Figure 18).

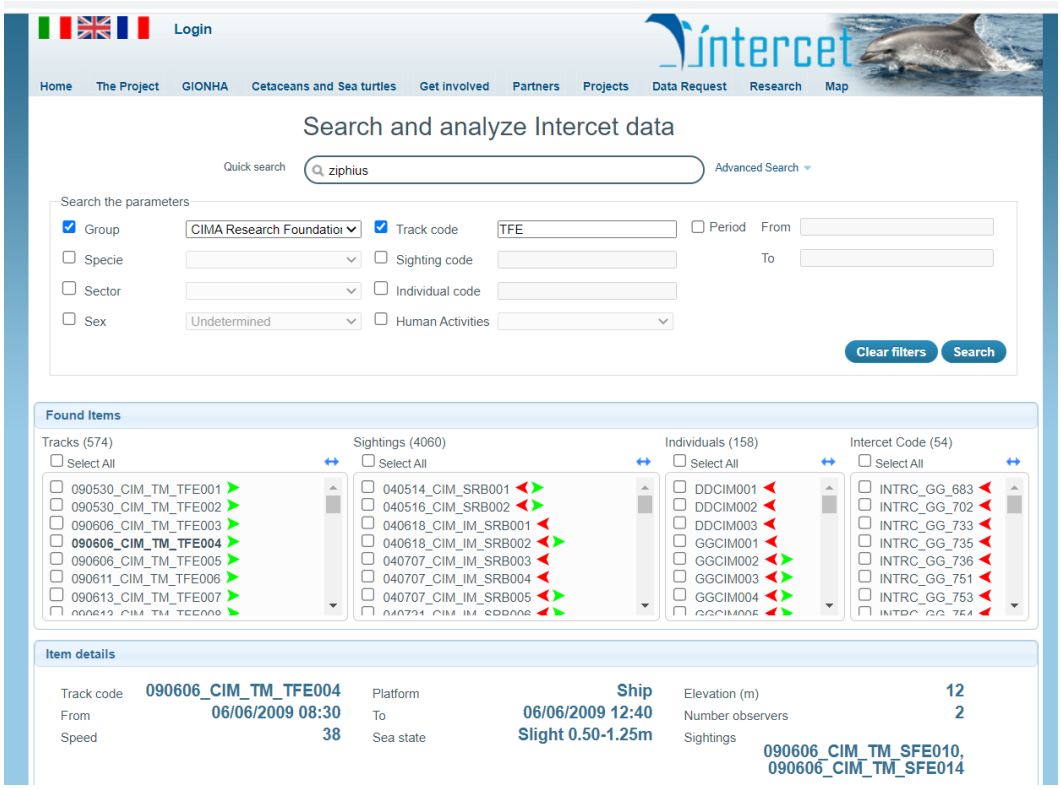


Figure 18 Examples of datasets from the FLT MED NET uploaded into the Intercet platform (Source: Intercet).

In case of sightings with **photo-identified** individuals, it is possible to upload the photos together with an individual's file in the platform, containing an individual code, information about the

species, sex or if the animal was seen dead. In the photo-identification section, a photo comparison of the photos uploaded with those in all catalogues shared by the research entities in Intercet can be performed, and if the individual's file matches, the new photos can be associated with the individual.

6.3.3 Sextant - Infrastructure de données géographiques marines et littorales

SEXTANT https://sextant.ifremer.fr/eng/Data/Submit-a-data

Sextant is organised in thematic catalogues based on data of a project, a species, a laboratory, or a geographic zone. The procedure to submit data consists in creating a thematic catalogue that will host data and allow the owners to administer it (writing metadata, managing access rights) and declare user accounts associated with their catalogue (using extranet accounts, unlimited creation). Data might be hosted by an existing catalogue already present in Sextant. To create a thematic catalogue, an email should be sent to sextant@ifremer.fr with a description of the catalogue using the form available in the site (https://sextant.ifremer.fr/eng/Ressources/List-of-thematic-catalogues) along with personal information (Name, Organisation, Email) to become catalogue administrator. A semi-automated and simplified data filling procedure is being developed.

Data published by Sextant are either available free or restricted, according to the licence selected by the data owner. Some datasets also have a DOI.

6.3.4 Joint Cetacean Data Programme

JCDP https://jncc.gov.uk/our-work/jcdp-resources/

The JCDP collates effort-related cetacean survey data collected from vessel or aerial platforms. Organisations or individuals collecting these data which follow the agreed Data Standard are encouraged to contribute via the Data Portal. The Data Portal provides the dashboard and tools necessary to search, filter and download third party data held within the JCDP. There are two levels of access with some data openly available to download, and some restricted data that require a request to the data owner/provider. The Portal also hosts a number of data products based on the dataset.

The JCDP provides a number of resources to data providers to assist in the collection and provision of cetacean survey data for inclusion in the JCDP database, as well as documents regarding data

access to set out the responsibilities that the data owners/providers and end users of the data stored within the JCDP have (https://hub.jncc.gov.uk/assets/1b35ddf6-c469-4bf8-8300-86ec21da1c2d).

A guidance on Data Collection supports the collection of data using different methods (aerial, shipboard surveys, strip transects) in a way that fits the JCDP Data Standard. The data stored within the JCDP database include effort-related cetacean survey data collected by dedicated, opportunistic, aerial digital imagery surveys.

Metadata guidance helps to create valid metadata to accompany the datasets to be submitted. Metadata is essential as part of any data submission to ensure data are understood and used appropriately. JCDP developed a standardised approach following the INSPIRE format, to coordinate across datasets to simplify use of the JCDP. The Metadata system is hosted by GeoNetwork and is publicly accessible (https://gis.ices.dk/geonetwork). Metadata are organised by topics, based on INSPIRE topics and resource types, and by levels.

Database Metadata (top level): one database-level metadata entry for the JCDP Data Portal, managed by the JCDP Governance Group and Secretariat.

Survey or Data Product Metadata (user level): a metadata record is required from the data custodian for data submission to the JCDP, and can cover a number of datasets submitted to the JCDP Data Portal where appropriate (with the same methodology, within the same project). Dataset or data product metadata entries are categorised under a JCDP group and so can be filtered and viewed together. Users can only create a metadata record within the JCDP group which will have its own page, themes and dataset metadata records.

Dataset child records Metadata (user level): records that can be created under a survey metadata record when a route is surveyed multiple times over the course of years while following the same field data collection protocol. This functionality may be useful for long-term monitoring (as the ones carried out by the EU MTT partners), where each child metadata record would cover the surveys within a calendar year.

At the moment of the preparation of this report, ORCA had uploaded 86 datasets with child records regarding data collected from ferries since 2006 (Figure 19).

The structure of metadata records can be adjusted to best suit the survey regime and data structure. To upload a new metadata, the user has to login to the GeoNetwork (account to be requested to marinemammals@jncc.gov.uk), select “Contribute”, “Add new record” (metadata), where the JCDP metadata template can be found, highlighting the mandatory fields that need to be filled out (to comply with INSPIRE more fields are needed): Title and Abstract, Date, Hierarchy level, Online Resource, Language, Spatial representation type, Encoding, Classification of Data and Services (INSPIRE categories), Keywords, Geographic coverage, Temporal reference, Reference System information, Quality, Conformity, Access and Use, Responsible Organisation. Once prepared, the metadata record has to be validated and a report will be produced with a link to the INSPIRE **validator**. Once it has been checked, corrected and validated, the user can select ‘Save and Submit’ which will send your metadata to a reviewer to check and publish. The metadata record URL will be needed when uploading the data to the Data Portal.

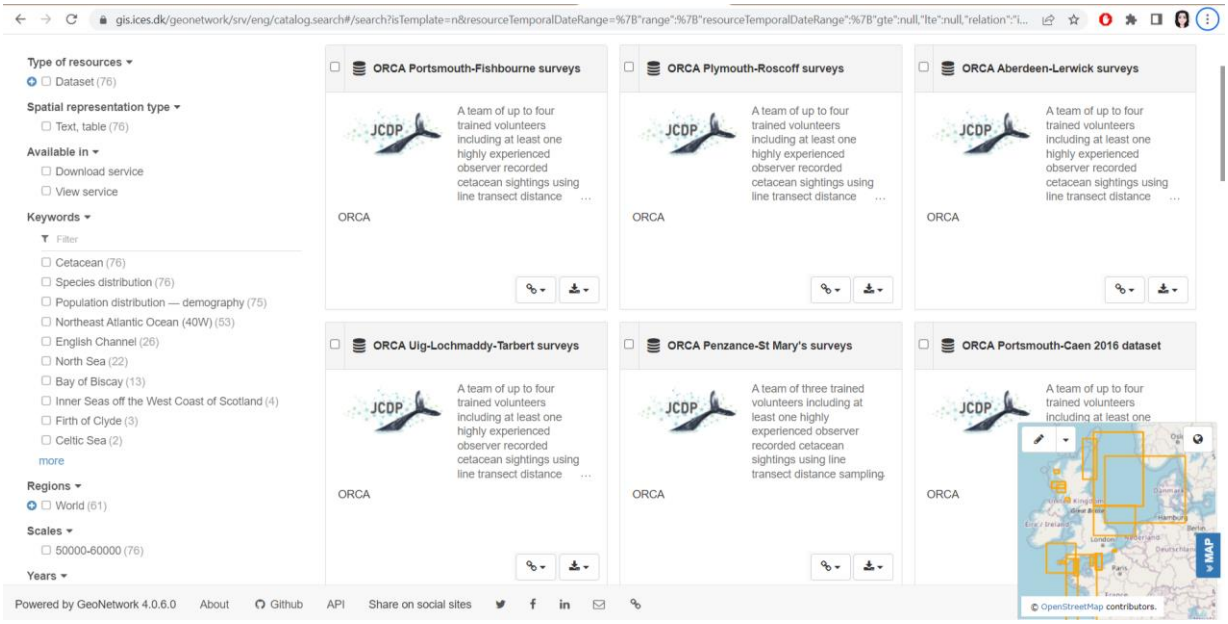


Figure 19 Examples of datasets from ORCA uploaded into the JCDP portal (Source: JCDP).

Data **standards** have been developed in collaboration with MEDIN to meet high standards of data management and formatting. The JCDP is in fact developing a UK cetacean data guideline which will be adopted by MEDIN and made available for all relevant data collectors in the north-east Atlantic, therefore becoming a recognised and appropriate standard for wide scale application. The Data Standard guidance contains information about the core and some non-essential data fields. Mandatory, conditional and optional fields are in fact present in each data table, and all fields and options are fully described in the guidelines. Drop-down menus are used for controlled vocabulary (referring to http://vocab.ices.dk).

There are three data tables: Identifiers; Effort and Environment; and Sightings Records.

The **Identifiers** table contains information relating to the survey platform, methodology and data holder.
Required fields: SurveyID (year organisation location), SurveyType, SurveyAbstract, DataAccess, DataRightsHolder, DataCustodian, DataCustodianContact, TargetTaxa, PlatformClass, NominalAngleOfSearch, Methodology, NominalNumberOfObservers.

The **Effort and Environment** table reports on the environmental and weather conditions, as well as the information relating to when the observation of animals was actively performed.
Required fields: EffortID, SurveyID, StartDate-Time, EndDate-Time (ISO format), StartLatitude-Longitude, EndLatitude-Longitude (decimal degrees), PlatformHeight, PlatoformSpeed (km/h), NumberOfObservers, AngleOfSearch, SeaState (Beaufort), SwellWaveHeight, Glare, Precipitation, PrecipitationIntensity, Visibility, Sightability, CloudCover (%), WaterTurbidity.

The **Sighting Records** table provides all details on the observation of the animals.
Required fields: Sighting ID, SurveyID, EffortID, Latitude, Longitude, AphiaID (from https://www.marinespecies.org/), IdentificationConfidence, Cue, SurveyTeam, BestGroupSize, Behaviour. Additional fields on group size and composition, radial distance and angle of observation can be provided according to the guidelines.

Feedback and comments on the data standard, vocabularies and metadata can be provided to help inform future developments of the JCDP system via a feedback form on the JCDP webpage. All data should be prepared according to the JCDP standard and saved in .xml format. The JCDP data template is an excel spreadsheet provided to help prepare data; it has separate tables to input survey data, and includes references to the vocabularies and a script to which re-formats the data into the necessary .xml format needed for upload (Export data to XML).

Guidance on data submission supports the process of uploading data through the JCDP portal hosted within the ICES data centre. The user has to login to the JCDP Data Platform (account to be requested to accessions@ices.dk); the Data Portal has inbuilt data validation steps to check for data standards. Once the screening has been completed, results are displayed by email or on the same page. If there are no errors that prevent data to be imported to the database, the user can confirm the submission using the “Import cetacean data to the database” button.

6.4 National

6.4.1 UK - Marine Environmental Data and Information Network Data Portal

UK - MEDIN https://medin.org.uk/data-standards/medin-data-guidelines

Discovery metadata is a list of information that accompanies a dataset and allows other people to find out what the dataset contains, where it was collected, and how they can get hold of it. MEDIN has produced a standard for marine metadata and a set of tools to create metadata records that comply with the MEDIN Metadata Standard (https://medin.org.uk/medin-discovery-metadata-standard). The MEDIN **Discovery Metadata Standard** is a marine profile of the UK government Standard GEMINI2 that also complies with other international conventions such as INSPIRE and ISO19115. The MEDIN Discovery Metadata Standard is hosted on GitHub (https://github.com/medin-marine/Discovery-Standard-public-content).

Mandatory elements for datasets: Resource title, Resource abstract, Resource type (controlled vocabulary), Unique resource identifier, Resource language (controlled vocabulary), Topic category (controlled vocabulary), Spatial representation type (controlled vocabulary), Keywords (controlled vocabulary), Geographic bounding box (controlled vocabulary), Spatial reference system (controlled vocabulary), Temporal reference (controlled vocabulary), Lineage, Limitations on public access, Conditions applying for access and use, Responsible party, Data format (controlled vocabulary), Frequency of update, Conformity (controlled vocabulary), File Identifier, Metadata date, Metadata standard name, Metadata standard version, Metadata language.

Optional elements: Alternative resource title, Extent, Additional information, Parent ID.

The Discovery Metadata **Editor** is an online tool to create, export and validate an XML metadata record, or to upload directly to the MEDIN Data Discovery Portal. Metadata Maestro allows users to create, edit, validate and transform discovery metadata records to MEDIN, GEMINI and ISO standards. Metadata Maestro operates in desktop standalone mode as well as using a web service to access required vocabularies when an internet connection is available. If the user wishes to **validate** XML metadata records within their own systems, MEDIN provides a dedicated Schematron and XML Schema Definition. This search facility is intended to help users to understand error messages that are generated by the Online Tool when a MEDIN discovery metadata record does not contain all the required information.

Three requirements ensure that the user is MEDIN compliant, if he supplies:

- General Metadata (common to all guidelines): project, survey
- Detailed Metadata: Methods (specific for each guideline)
- Data: in the standardised formats.

For several metadata fields, MEDIN insists that terms come from a controlled vocabulary to ensure consistency between records. Controlled vocabularies for the marine sector commonly used by MEDIN are from NERC, BODC, ICES, SeaDataNet; for species names WORMS and MSBIAS (https://www.marinespecies.org/msbias/aphia.php?p=search).

Data Guidelines provide a list of information that should be collected with data to ensure they can be re-used in the future. The guidelines are tailored to different methods and provided by theme with template files, highlighting the mandatory, conditional and optional fields. The following information was retrieved from the file "Data guideline for transect surveys":

Project metadata include name, code, date as mandatory fields. **Survey metadata** include name, type, abstract, originator, owner, startdate, timezone, access and reuse conditions as mandatory fields, but users can provide additional information as for platform type or dates. Some of these fields have controlled vocabulary.

Methods metadata for transect surveys have three mandatory fields: methodID, transectLenght (in meters) and QCScheme (description of any quality control scheme during the analysis).

Other suggested fields are: heightOfObservations, transectDescription, samplingDevice, protocolsUsed, speedOfTravel. Some of them have controlled vocabulary and units are specified.

Image and Station data are optional. **Survey Event data** are required and mandatory fields are: surveyEventID, surveyCode, SurveyEventCode, Y-Xcoordinate, spatialCRS, positionFix, accuracy, eventStartDate-Type, eventStartTime-Precision. Some of them have controlled vocabulary and units are specified. Other conditional fields are included, also linking the information with the Station data (optional), Survey metadata and Detailed Metadata forms.

Similarly, the mandatory fields for **Sample Event data** are sampleEventID, methodID, sampleStartDate-Type, sampleStartTime-Precision, and other conditional and optional fields.

Taxon Data are linked to the survey/sample event data: eventID, eventIDType, taxonID, taxonName, aphiaID, abundanceUnits are the mandatory fields. Additional information about gender, developmentStage, behaviour can also be provided. *Taxon History* is also linked to the Taxon Data form adding information about the determination date.

Similarly, **Biotope Data** provides information about biotope reference and determiner and is related to eventID, eventIDType, and *Biotope History* adds information with the determinationDate field.

Taxon Attribute Data is linked to taxonID, including several measurements and counts fields. Linked to the eventID, eventIDType there is the **Event Attribute Data** where to add measurements and attributes about the event, and the **Geological Data** form, for geological

information collected. Matrices of taxon data, taxon attributes and event attributes can be created.

6.4.2 UK - The Archive for Marine Species and Habitats Data

UK - DASSH https://www.dassh.ac.uk/submitting-data

To submit data, the DASSH team should be contacted by email (dassh.enquiries@mba.ac.uk) or by phone (+44 (0)1752 633291) to discuss data/images and archive requirements. DASSH can accommodate a wide range of electronic data types and file formats, and data can be submitted by post (on suitable media), by email or by file transfer. Electronic data (spreadsheets, databases, images etc.) are stored in a digital archive.
A Metadata file is required, describing who collected the data, where and when they were collected, what the dataset(s) describes, how the dataset was collected, and a contact person/organisation or data provider from which a copy of the data can be obtained. DASSH Metadata is collated using the UK GEMINI Discovery metadata standard. This is a defined element for describing geo-spatial, discovery level metadata within the United Kingdom. This profile is derived from the ISO 19115 Geographic Infomation - Metadata and the UK eGovernment Metadata Standard (eGMS).
Metadata on all archived material are held at DASSH and shared with MEDIN via the Data Discovery Portal.

6.4.3 IT – Centralised Information System - ISPRA

IT – SIC - ISPRA http://www.db-strategiamarina.isprambiente.it/app/#/upload

Each body in charge for the activities implemented within the MSFD have access to the Centralised Information System in order to download specific template files for each activity/descriptor. The users have to format the data derived from such monitoring activities as indicated in the specific guidelines and fill two tables, one with the sampling stations information, the other with the complete survey data.
Examples of required fields are: CountryCode, NationalStationID, which links the two tables, NationalStationName, Region, Latitude, Longitude, ClosestCoast, Time and sightings information. After verifying if the files are compliant through the specific utility in the Tools sections, the users can finally upload data files in the System.

6.4.4 IT – National Network of Biodiversity - ISPRA

IT – NNB - ISPRA https://www.nnb.isprambiente.it/it/strumenti-e-risorse/manuali-e-guide-tecniche-per-i-partner

To become a partner of the NNB, the interested party should send its expression of interest through the form (https://www.nnb.isprambiente.it/it/il-network/come-entrare-a-far-parte-della-rete-dei-partner-nnb/compila-modulo). After sending the request, the new partner will be contacted to sign a Memorandum of Understanding with the National Ministry of Ecological Transition. To be uploaded into the BioCASe system, the dataset has to be digital (access, database, .xls, .shp, etc.) and have a text document describing the content and the fields, and a metadata file. Different guidelines and template files are provided in order to publish a dataset (https://www.nnb.isprambiente.it/it/strumenti-e-risorse/manuali-e-guide-tecniche-per-i-partner), which describe the BioCASe logic, database creation, data standardisation guidelines, Web-GIS manual, geographic metadata creation and Data policy. Data standardisation of an existing dataset is required following the indications.

The minimum information required to map a database into BioCASe is:
DataSets/DataSet/ContentContacts/ContentContact/*Email*
DataSets/DataSet/ContentContacts/ContentContact/*Name*
DataSets/DataSet/Metadata/Description/Representation/*@language* (IT, EN, FR)
DataSets/DataSet/Metadata/Description/Representation/*Title*
DataSets/DataSet/Metadata/RevisionData/*DateModified* (last modified version)
DataSets/DataSet/TechnicalContacts/TechnicalContact/*Email*
DataSets/DataSet/TechnicalContacts/TechnicalContact/*Name*
DataSets/DataSet/Units/Unit/Gathering/*LocalityText* (name of locality, region, city)
DataSets/DataSet/Units/Unit/Gathering/NamedAreas/NamedArea/*AreaName* (name of SIC, MPA, etc.)
DataSets/DataSet/Units/Unit/Identifications/Identification/Result/TaxonIdentified/ScientificName/*FullScientificNameString*
DataSets/DataSet/Units/Unit/*SourceID* (Institution name)
DataSets/DataSet/Units/Unit/*SourceInstitutionID* (Institution code)
An excel file containing all fields established by INSPIRE guides the creation of the geographic metadata. Information about the data owner, data type, classification, access regulation, distribution (download), and data quality, have to be filled according to the drop-down menus

and controlled fields. Coordinates are required in decimal degrees WGS84. Finally, an XML file is created.

Databases can be published following the principles of interoperability through BioCASe Protocol, that follows the operational and logic choices of BioCASe Europe systems. A guide for BioCASe installation is provided. In other cases, users can share a database through OGC and Open API standards, available from the provider, by communicating with the access point and filling the Excel file for the geographic metadata creation. Finally, a dataset can be integrated in the NNB database when the provider could not publish them according to interoperability. All datasets are published open access with licence CC BY.

The recently developed Geodatabase of the FLT MED NET is hosted in the NNB portal.

6.4.5 IT - Joint Cetacean Database and Mapping

IT - JCDM https://jcdm.dss.uniroma1.it/partners

To contribute with data, the interested partner should write to jcdm.dss-dba@uniroma1.it and will receive a standard research agreement that can be personalised according to the partner's needs. By sharing information through an accessible knowledge base, JCDM Partners can support scientific knowledge in species distribution and abundance.
Data acceptable for JCDM can be presented in different formats (.xls; .xlsx; .csv; .txt; .mdb; .accdb; .dbf; .shp; .gdb; .json). Data should provide coordinates (latitude and longitude decimal degrees WGS84) of sightings, species (scientific name and common name), date (any format, specify which one). Effort data is also needed, by specifying which format.
By default, each dataset submitted and registered into the JCDM database falls in the CC BY-NC licence, unless the owner explicitly chooses another option. Once registered, the user can download data from the JCDM data Engine according to the security level of the registration:

- level 0: allows the visualisation and use of JCDM data features. The download of the data is not allowed as well as the access to the JCDM Database.
- level 1: allows the visualisation and use of all JCDM data features. Allows the download of the database following the agreement established with partners and to generate results (maps and tables).
- level 2: allows the download of the entire JCDM Database.

The partners will retain ownership of the data. Users are required to cite data sources and the JCDM team will provide the proper citation information for each data source.

7. Final remarks

Monitoring the marine environment is often a costly and effort demanding activity, for which huge investments in terms of economic and human resources are necessary. However, the processes that drive the functioning of marine ecosystems are often interconnected, and need to be assessed at large geographical and temporal scale. Thus, the collection of harmonised data using shared standardised protocols, and their storage in common databases, would allow the study of marine ecology and biology at a wide scale. The result is the current co-existence of many repositories that collate data obtained by different research organisations and by regional, national and basin-scale monitoring programmes (Section 4, Table 3).

While joining such databases may amplify the possibilities and benefits for researchers and managers, it can be challenging and not straightforward. In fact, there is a lack of common guidelines to facilitate data interoperability between monitoring data from different infrastructures (Section 2). Within this framework, under the European Strategy Forum on Research Infrastructures, the **European Long-Term Ecosystem Research** network (https://www.ilter.network/network/lter-europe) is developing a globally distributed Research infrastructure of long-term research (the eLTER Research Infrastructure, https://elter-ri.eu/) that focuses on harmonised parameters and methods and that will serve the scientific community and establish standardised data, services and training. In addition, it is worth noticing the launch in 2022 of the Biodiversa+ *BiodivMon Call for Research Proposals* on "Improved transnational monitoring of biodiversity and ecosystem change for science and society", dealing with the topic of this report, i.e., harmonisation of methods and tools for biodiversity monitoring, data collection, and use of available biodiversity monitoring data.

Indeed, the EU legislation strongly encourages this process, through the INSPIRE directive, which establishes an Infrastructure for Spatial Information in the European Community, based on the FAIR data principles, according to which data should be findable, accessible, interoperable, and reusable. Within this context, several common databases have been created with the aim of pooling data together and making them accessible, either at a global, regional, or national scale, but only a few are compliant with the directive's standards and share common formats (Section 6). In fact, each of these repositories has specific requirements, and data collected by research networks and monitoring programmes often need to be formatted or modified in order to fit with the required fields and be comparable with already existing data.

This is one of the main challenges for data interoperability and harmonisation highlighted during the Biodiversa+ workshop. Knowing the background, the recent FLT MED NET Geodatabase is an

example of *best practice* of data harmonisation, being INSPIRE compliant, allowing database search and data extraction among the partners, but also allowing interoperability with other GIS-based platforms and external databases (Di Stefano et al, 2023).

The EU MTT project provides a good example of a network of research institutes and non-profit organisations that perform marine monitoring (mostly of marine mammals) working to harmonise their protocols in order to obtain comparable data. The development of a common recording tool (AtSea, developed within the MTT IMPEL project in 2021 on open-source mobile data collection platform ODK) within this same project is the result of the common desire to overcome the slight differences in the protocols and data formats (e.g., sea state, visibility) and improve data standards of the involved networks. As well, the planning of training programmes among observers from the different partners will be useful to guarantee that all involved institutions operate in a consistent way. All partners, in fact, are already cooperating in order to allow interoperability of their databases and contribute to fit their data to the legal requirements of the EU biodiversity monitoring (e.g., MSFD, INSPIRE etc).

While the effort made within the EU MTT project is an example of good practice, at a larger scale, it would be desirable to work towards improving data interoperability among the existing repositories, to allow datasets to fit to different catalogues according to their specific thematic or geographical scope and needs, and avoid the duplication of effort by both data providers and data users.

8. References

Alves, A., Ferreira, R., Fernandez, M., Halicka, Z., Días, L., Dinis, A. (2018). Analysis of occurrence patterns and biological factors of cetaceans based on long-term and fine-scale data from platforms of opportunity: Madeira Island as a case study. Marine Ecology 39(2), e12499.

Arcangeli, A, Campana, I., Paraboschi, M.; ISPRA, Rome, Italy (2018). Presence of cetacean species collected through Fixed-Line-Transect monitoring across the Western Mediterranean Sea (Civitavecchia-Barcelona route) between 2014 and 2018. https://dx.doi.org/10.14284/533

Campana, I., and Vighi, M. eds (2020). Monitoring large marine vertebrates along fixed transects from ferries and cargo vessels: a state of the art. IMPEL report 2020/17. 93 pp.

Correia, A. M., Tepsich, P., Rosso, M., Caldeira, R., Sousa-Pinto, I. (2015). Cetacean occurrence and spatial distribution: Habitat modelling for offshore waters in the Portuguese EEZ (NE Atlantic). Journal of Marine Systems, 143, 73-85.

Correia, A. M., Gandra, M., Liberal, M., Valente, R., Gil, Á., Rosso, M., ... & Sousa-Pinto, I. (2019). A dataset of cetacean occurrences in the Eastern North Atlantic. Scientific data, 6(1), 1-8.

Correia, A. M., Gil, Á., Valente, R. F., Rosso, M., Sousa-Pinto, I., Pierce, G. J. (2020). Distribution of cetacean species at a large scale-Connecting continents with the Macaronesian archipelagos in the eastern North Atlantic. Diversity and Distributions, 26(10), 1234-1247.

Correia, A.M., Oliveira-Rodrigues, C., Gandra, M., Liberal, M., Valente, R., Gil, A., Rosso, M., Pierce, G.J. & Sousa-Pinto, I., CIIMAR - UP. (2022). CETUS: Cetacean monitoring surveys in the Eastern North Atlantic. Marine Data Archive. https://doi.org/10.14284/547

Di Stefano, C. Tepsich, P., Grossi, F., Pasanisi, E., Santini, E., Arcangeli, A. (2023). Database and WEB-Gis platform to manage large dataset of marine megafuna in the open seas: an essential tool to help EU environmental policies. Poster at 34th ECS conference, Galicia, Spain.

Druon, J. N., Panigada, S., David, L., Gannier, A., Mayol, P., Arcangeli, A., Canadas, A., Laran, S., Di-Méglio, N., Gauffier, P. (2012). Potential feeding habitat of fin whales in the western Mediterranean Sea: an environmental niche model. Marine Ecology Progress Series, 464, 289-306. https://doi.org/10.3354/meps09810

Gnone, G., Bellingeri, M., Molinari, Y., Dhermain, F., Labach, H., Díaz López, B., ... & Cañadas, A. M. (2022). The Seabed Makes the Dolphins: Physiographic Features Shape the Size and Structure of the Bottlenose Dolphin Geographical Units. Journal of Marine Science and Engineering, 10(8), 1036.

Kiszka, J., Macleod, K., Van Canneyt, O., Walker, D., Ridoux, V. (2007). Distribution, encounter rates, and habitat characteristics of toothed cetaceans in the Bay of Biscay and adjacent waters from platform-of-opportunity data. ICES Journal of Marine Science, 64(5), 1033-1043.

Martino, S., Pace, D.S., Moro, S., Casoli, E., Ventura, D., Frachea, A., Silvestri, M., Arcangeli, A., Giacomini, G., Ardizzone, G., Jona Lasinio, G. (2021). Integration of presence-only data from several sources: a case study on dolphins' spatial distribution. Ecography, 44, 1533-1543. https://doi.org/10.1111/ecog.05843

Matear, L., Robbins, J. R., Hale, M., Potts, J. (2019). Cetacean biodiversity in the Bay of Biscay: suggestions for environmental protection derived from citizen science data. Marine Policy, 109, 103672.

Morgado, C., Martins, A., Rosso, M., Tepsich, P. (2017). Fin Whale presence and distribution in the Pelagos sanctuary: temporal and spatial variability along 2 fixed-line transects monitored in 2009-2013. International Journal of Marine and Environmental Sciences, 1(1), 1-14.

Oliveira-Rodrigues, C., Correia, A.M., Valente, R., Gil, A., Gandra, M., Liberal, M., Rosso, M., Pierce, G., Sousa-Pinto, I. (2022). Assessing data bias in visual surveys from a cetacean monitoring programme. Scientific Data 9, 682. https://doi.org/10.1038/s41597-022-01803-7

Robbins, J. R., Babey, L., Embling, C. B. (2020). Citizen science in the marine environment: estimating common dolphin densities in the north-east Atlantic. PeerJ, 8, e8335.

Scuderi, A., Martín Moreno, E. (2020). Cetacean monitoring programme along fixed transect using ferries as platforms in the Strait of Gibraltar, 2018, FLT Med Network.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1), 1-9.