DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

Integration of biodiversity monitoring data into the Digital Twin Ocean

dto-bioflow.eu

September 2024

Acknowledgements

This report has been produced by the DTO-BioFlow project (GA No. 101112823), which received funding from the European Union's Horizon Programme call HORIZON-MISS-2022-OCEAN-01.

Disclaimer

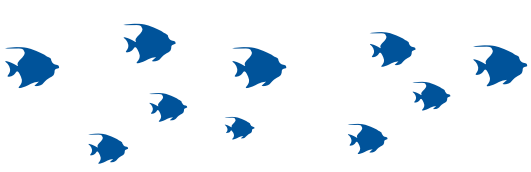


Table of Contents

©Thanos Dailianis | HCMR

Introduction

The DTO-BioFlow project, funded by the European Union's Horizon Europe research and innovation program, aims to enhance the flow of marine biodiversity data into the European Union Digital Twin Ocean (EU DTO). This initiative is critical for achieving the EU Green Deal and Biodiversity 2030 targets by creating a comprehensive digital replica of ocean processes, which helps in understanding, predicting, and simulating various ocean scenarios for better decision-making.

A myriad of actors (researchers, public bodies, blue economy operators, NGOs and citizen science groups) are collecting marine biodiversity data for various purposes and using a diverse set of collection methods. Yet, despite significant advancements in Europe to collect, harmonize, and make these marine biodiversity data available, large amounts of biodiversity data do not find their way into repositories due to a variety of reasons, including technical barriers. Many systems were not originally designed to handle the complexity of the biodiversity data being generated today.

DTO-BioFlow is establishing data flows for several new biodiversity data types produced using different techniques and instruments and which do not yet have established dataflows to make them available in long-term data repositories. The diagram shown is a simplified version of the data flows that will be captured within DTO-BioFlow. All activities start with data collection, which can be done ***in situ*** or via **remote sensors** and/or **human observation**. Data collected by DTO-BioFlow partners will be subjected to several processing steps (not covered in this diagram).

Depending on the type of data, data may be first submitted to an integrator or repository or go directly into EMODnet Biology. Data in repositories might also be shared with integrators before being made available in EMODnet Biology. In EMODnet Biology, various procedures are implemented to ensure the data are quality checked to avoid publishing data with issues. If issues are encountered, the originators are informed so these can be addressed ahead of publication, thus ensuring that all data are of high quality. All data available via EMODnet Biology **will be harvested by the EU DTO and will be findable in the EU DTO Data Lake.** This will require a few processing steps, more specifically reformatting to cloud optimised formats which will allow for faster and more efficient queries.

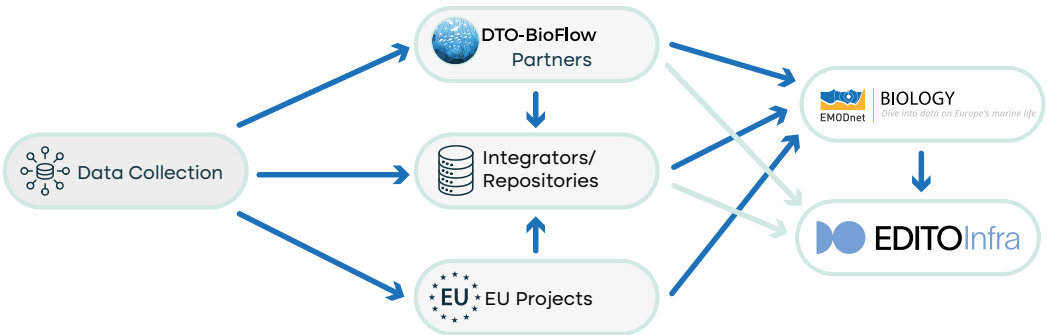
Data Integrator vs Data Repository

A data integrator refers to a facility that allows the storage of data and that provides tools that allow for the data to be analyzed. The terms integrator and aggregator are used interchangeably throughout the document. In contrast, a data repository is a facility that allows for data storage but does not provide functionalities for data analysis within the platform.

The data flows include data from genomics observations, plankton imaging observations, fish, mammal and bird biologging, cetacean passive acoustic observations, and other relevant biodiversity data sources.

Several **Demonstrator Use Cases (DUCs)** will be developed to utilize these new data flows in the DTO, integrating and assimilating them into existing models using harmonized protocols. These DUCs are scientific investigations that guide the implementation of biological data and analysis resources into the DTO. In the first 10 months of the DTO-BioFlow project, Work Package 4 (WP4) has conceptualised eight DUCs covering a wide range of ecosystem management topics. These include **invasive species management**, **adaptive offshore construction**, **pelagic biodiversity assessments**, **marine aquaculture**, **management of marine protected areas**, and **marine carbon sequestration**.

By showcasing these use cases, DTO-BioFlow aims to **demonstrate the practical applications of integrating diverse marine biodiversity data into the DTO, highlighting the project's impact on various aspects of marine ecosystem management and policymaking.**



DTO-BioFlow contribution to EMODnet Biology

About EMODnet

The European Marine Observation and Data Network (EMODnet) is a European Commission marine data service of the EC Directorate-General Maritime Affairs and Fisheries (EC DG MARE) and funded by the European Maritime Fisheries and Aquaculture Fund. Established in 2009, EMODnet plays a pivotal role as a trusted source of *in situ* marine environmental and human activities data and data products, serving a diverse user base across various sectors.

The EMODnet Portal (emodnet.ec.europa.eu) is a single access point to all EMODnet services, providing easy and free access to a wealth of marine data, metadata, and data products. Covering seven disciplinary themes – bathymetry, geology, physics, chemistry, biology, seabed habitats and human activities – EMODnet spans the entire marine environment from coast to open ocean, surface to deep seafloor, whilst also offering data and information on Blue Economy operations, from vessel density and offshore platform sitings, to hosting EU Member State Maritime Spatial Plans.

In collaboration with over 120 partner organisations and stakeholders, EMODnet aggregates data into Findable, Accessible, Interoperable, and Reusable (FAIR) pan-European data layers, with data and accompanying metadata harmonized according to EU e.g., INSPIRE and International e.g., ISO standards. In addition, EMODnet experts produce unique data products such as EUSeaMap, a broad-scale seabed habitat map, the EU Digital Terrain Model for harmonized bathymetry, and the pan-European Marine Litter database.

EMODnet caters to a broad user community, including the private sector and Blue Economy, national authorities and administrators, research and academia, non-governmental organisations, and citizen science initiatives.

This dynamic network of experts continuously evolves EMODnet into the operational EU marine data service it is today.

Source: EMODnet Annual Report 2023

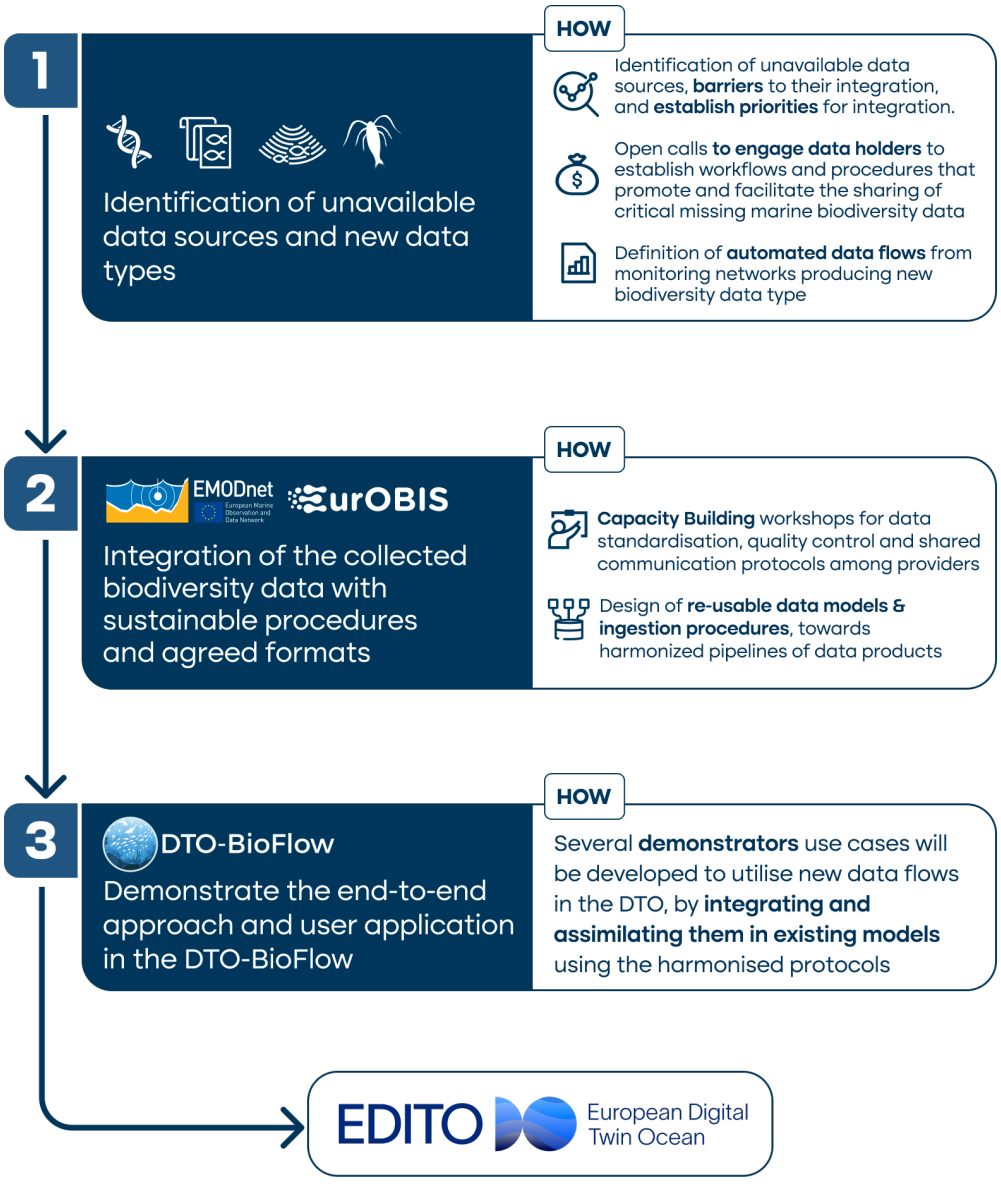
DTO-BioFlow is designed to streamline the integration of new biodiversity data types into EMODnet Biology, enhancing its repository with more comprehensive and diverse datasets. This integration ensures that the data collected can be effectively utilized for policymaking, scientific research, and environmental monitoring. **Within DTO-BioFlow, data pipelines for each of the data types addressed in the project will be established to maintain a sustained flow of biodiversity data towards EMODnet Biology, and ultimately into the EU DTO infrastructure.** The status of the data flows towards EMODnet Biology varies according to the diverse types of data addressed by DTO-BioFlow. In some cases, there is no data flow implemented, in other cases a data flow is already in place and will be subject to improvements during the project's lifetime.

From data sources to users' applications

DTO-BioFlow targets the current challenges in the collection, harmonisation, accessibility, and analysis of marine biodiversity relevant data, the integration of these data and analytical tools into DTO architectures, and the applicability of a functioning DTO biodiversity component to address policy-relevant use cases.

By fostering community action and working with marine biodiversity data providers and users, DTO-BioFlow will make biodiversity data more accessible and usable and inform policy development. This will enhance the overall understanding and management of marine biodiversity, contributing to sustainable ocean governance.

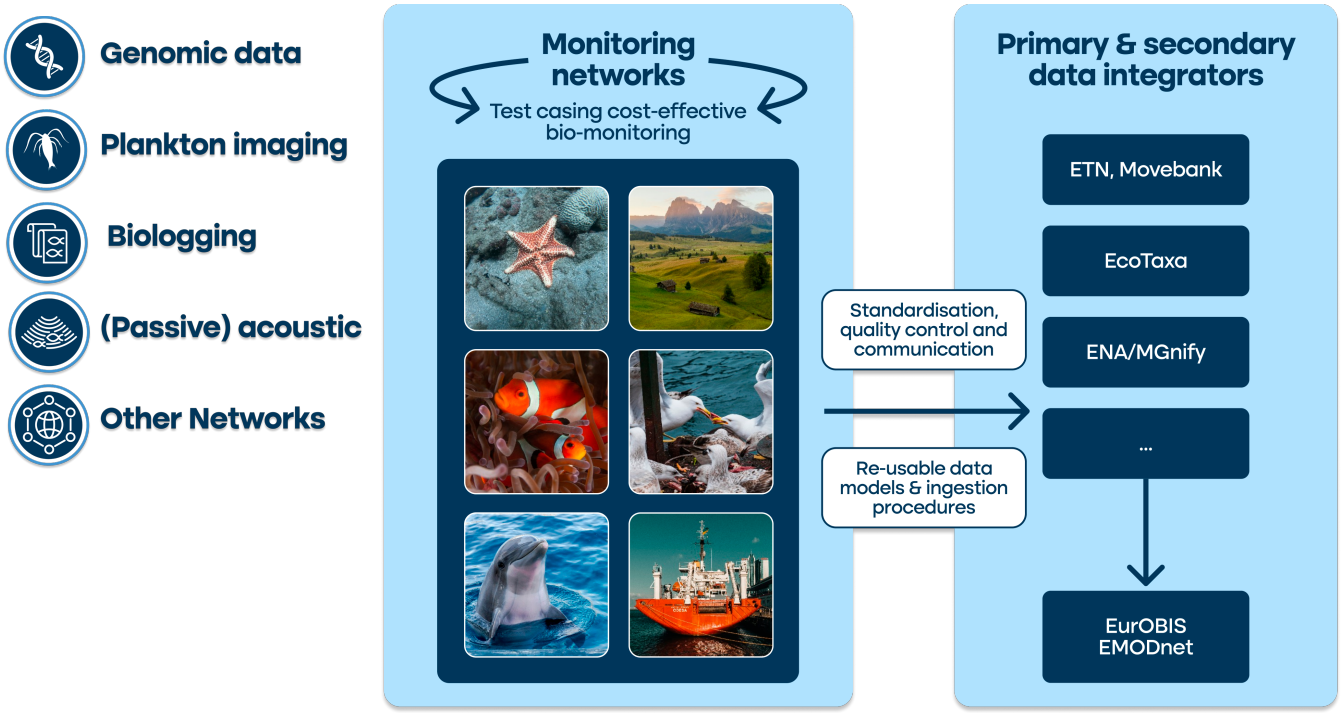
DTO-BioFlow approach




©Thanos Dailianis | HCMR

What type of data will become available?

DTO-BioFlow is establishing data flows for several new biodiversity data types produced using different techniques and instruments which do not yet have established dataflows to make them available in long-term data repositories. These include data from **genomics observations, plankton imaging observations, fish, mammal and bird biologging, cetacean passive acoustic observations, and other relevant biodiversity data sources.**



Which format for the data: EMODnet processing level

The status of the data flows towards EMODnet Biology varies according to the diverse types of data addressed by DTO-BioFlow. In some cases, there is no data flow implemented, in other cases a data flow is already in place and will be subject to improvements during the project's lifetime.

To harmonise the DTO-BioFlow outputs, in terms of data and data products, DTO-BioFlow adhere to the Processing Levels established by EMODnet that can be found in the Data and Data Product Portfolio available through the EMODnet website. The DTO-BioFlow data and data product outputs are mapped against the levels described in the following table.

Data product:
an output resulting from the analysis or modelling of *in situ* data that can include various interpolation techniques and parameters derived from the original data.

Table: EMODnet Data and Data Products Processing Level. Information included in the EMODnet Data and Data Product Portfolio document available at https://emodnet.ec.europa.eu/en/communication.

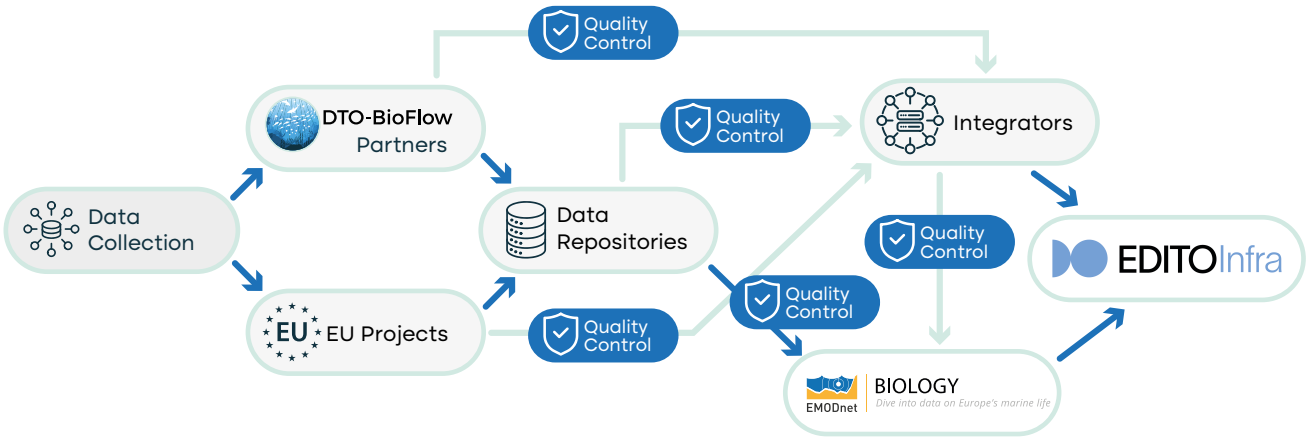
Level	Description	Applies to
L0	Raw data. Unprocessed instrument data at full resolution, including synchronisation methods (e.g. elimination of CTD up-down duplicates) and excluding communication artifacts.	Data
L1	Full resolution data reconstructed with calibration coefficients, geo-and time-referenced.	
L2	Geo- and time-referenced processed (derived) data with a minimum QC. Near-real time (NRT) with full spatial and/or temporal resolution.	
L3	Delayed mode data with further QC, usually with some completeness, consistency and space/time uniformity, Data QC checks may include comparison with historical data and/or Level 5 products such as climatologies or gridded data.	
L4	Collated data from different measurements, samples and/or sources that have been integrated in a data system by means of standardisation and/or categorisation, and subset or otherwise selected or derived to fulfil a specific requirement. Data can represent numerical values and resence/absence of a category or entity. Integration of datasets at this level enables further QC based on parameter to parameter relationships (e.g. TS diagrams).	
L5	Model or analysis output that uses data of Level 2 and/or 3 as input. Data products of this level represent the spatial distribution of a single parameter derived from multiple measurements. Data are aggregated and undergo some level of geo-processing and spatial or temporal interpolation to cover data gaps and/or solve data discrepancies.	Data Products
L5A	One-dimensional distribution of a specific parameter, without variations on the temporal or depth dimensions.	

The following sections provide an overview of the specific types of data that will become accessible through DTO-BioFlow, including their status, processing level, and location of the data.

Genomics Data

"Genomic data" is an umbrella term used to cover any kind of **nucleotide sequence data**, irrespective of whether it includes whole genomes, whole transcriptomes, or only specific loci. It may refer to the nucleotide sequence information of individual organisms (e.g. for genetic barcoding), as well as of multiple organisms simultaneously by applying the prefix "meta-". Whole genome sequence data of e.g. microbial community samples will consequently be called ***"metagenomic" data***. In cases where multiple organisms are assessed simultaneously but the focus are individual loci (most commonly marker loci used for taxonomic identification of species), the term ***"metabarcoding"*** is used. Genomic data may include taxonomic and/or functional annotations.

- **Current status:** there is no established genomics data flow to EMODnet Biology.
- **EMODnet Data and Data Products Processing Level:** L3
- **DTO-BioFlow upgrade:** occurrence data of biological taxa in delimiter-separated tabular format, which will be annotated with literature information for species of interest, and environmental metadata, which can be retrieved from ENA BioSamples or from other repositories (e.g. PANGAEA).
- **Data repositories:** MGnify (based on ENA), GBIF, OBIS

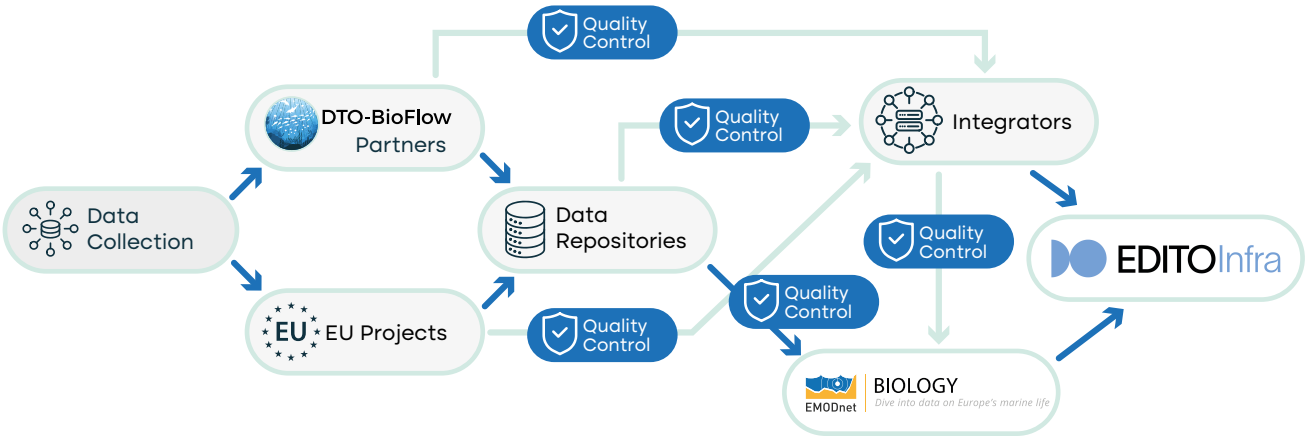


Plankton Imaging

Plankton imaging data covers all data generated by quantitative imaging instruments that provide many images, consistently, to reliably extract quantitative information, such as individual counts and/or biovolumes. These observations of planktonic organisms allow us to better quantify their role as key trophic and functional links in open ocean ecosystems. In addition, the *in situ i*nstruments or samples are processed shortly after collection and can provide information on so-called **marine snow**, which constitutes 80% to 90% of the particles in seawater. This marine snow plays a crucial role in marine ecosystems by facilitating the transfer of carbon to the deep ocean through the biological pump mechanism.

A wide variety of instruments can collect imaging data *in situ* or in the laboratory, each one with its own workflow. **It is expected that data from plankton and particle imaging will flow into the EU DTO mostly via data integrators/repositories** that are deployed across several *in situ* observatories and networks, notably including those overseen by DTO-BioFlow partners. The improved flows of data from plankton imaging expected from DTO-BioFlow will enhance biomonitoring efforts and global carbon flux estimations.

- **Current status:** After collection, images are processed, usually to extract Regions of Interest (ROIs) from full frames and to measure them in numerous ways. While there is an existing flow of plankton imaging data to EMODnet Biology, several types of data from these images are not yet included in EMODnet Biology. This will be improved for various data integrators/repositories within the scope of the DTO-BioFlow project.
- **EMODnet Data and Data Products Processing Level:** L2; L3; L5 (possibly)
- **DTO-BioFlow upgrade:** 3 data sets will be produced as part of WP4 Demonstrator Use Cases (DUC)
  - microscopic count data from traditional monitoring; data from high-throughput imaging; next generation sequencing (DUC 4.3)
  - biomass monitoring from stationary and on-demand mobile sensor platforms to assess and monitor biological processes like primary production or blooms and inform marine stakeholders (e.g., fisheries, DUC 6)
  - maps of carbon export (EMODnet level 5 product) at global scale with some time resolution (e.g. seasonal) (DUC 7)
- **Data repositories:** EcoTaxa, EcoPart, PyOPIA, LifeWatch Observatory, SUBSIM



Biologging

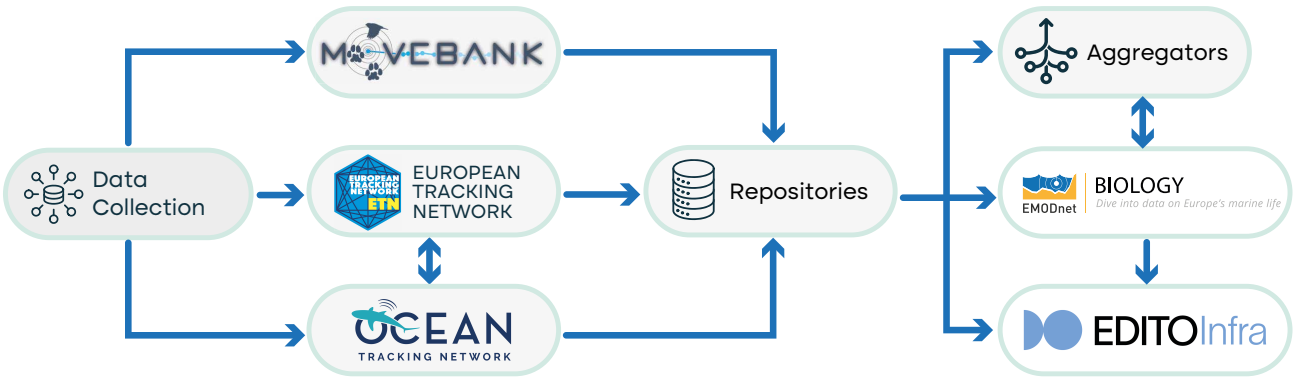
Biologging data are data of animal positions/presences obtained by animal-borne electronic devices. For this project, we consider biologging data from animals tagged in Europe, with mainly marine positions. They can be divided into three types:

1. GPS tracking involves using electronic devices, like GPS and SPOT tags, to record the GPS positions of larger marine animals that live above or near the water surface, such as large birds, marine mammals (cetaceans), fish (i.e. sharks) and marine reptiles (turtles). This data is managed in Movebank.

2. Archival tracking uses electronic tags which store environmental data like water temperature, pressure (i.e. depth), acceleration, and light. The data is either downloaded from a detached and retrieved device (data storage tags) or directly sent through the ARGOS satellite system. Animal trajectories are reconstructed through geolocation, requiring data processing. This data can be managed in ETN.

3. Acoustic telemetry uses a network of receivers to detect signals from animal-borne acoustic transmitters, mainly for tracking fish and crustaceans. The transmitters emit an acoustic signal (on a fixed frequency) with a unique ID on a predefined time interval, generating a trajectory as the animal moves through the receiver network. Detection range averages 200 meters but varies with environmental conditions. This data is managed in ETN.

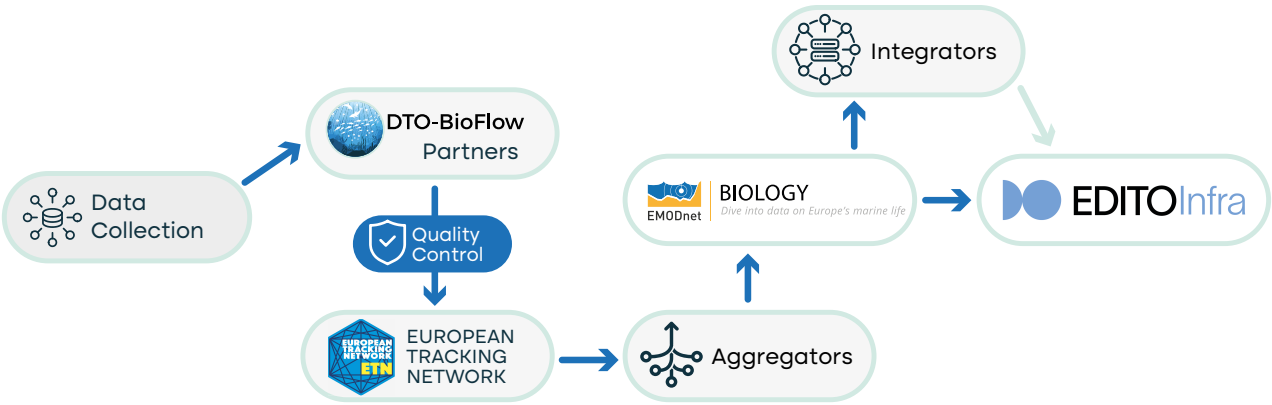
- **Current status:** Biologging data **are not operationally shared with EMODnet Biology yet**, however data from biologging projects that are part of LifeWatch Belgium have been published as open data. These data were formatted as a Data Package and published as a Marine Data Archive for ETN, and a Zenodo deposit for Movebank. The latter GPS tracking data were also transformed and published as Darwin CoreArchive that are available in GBIF and OBIS and harvested from one of these aggregators by EMODnet Biology.
- **EMODnet Data and Data Products Processing Level:** None
- **DTO-BioFlow upgrade:** automated report files providing a summary of the tagged animals within a specific time frame or project, animal tracks and device maintenance reports. In addition, dynamic maps will be developed to map the detections or presences of tagged animals. DUC 2 will use and apply these data.
- **Data repositories:** Movebank, ETN



Passive Acoustic

Passive Acoustic Monitoring (PAM) of sounds made by marine animals is an important method for estimating the distribution, density, and abundance of species that vocalise. It is particularly useful for animals that frequently produce species-specific vocalisations. Harbour porpoises (*Phocoena phocoena*) are excellent candidates as seldom does a minute go by without a porpoise producing species-specific echolocation clicks. The biological interpretation of acoustic data requires detecting signals produced by the animals, and the development and evaluation of detectors for classification is an active area of research. DTO-BioFlow is proposing the primary passive acoustic data to flow from ETN into EMODnet Biology.

- **Current status:** Currently, one dataset from passive acoustic monitoring can be found on EMODnet Biology for Phocoena phocoena from observation network in the **Belgian Part of the North Sea.** In addition, several PAM monitoring projects are collecting these data in the European seas which are currently not flowing to any data repository or aggregator. For the UK and North Sea waters a comprehensive inventory has been created, commissioned by the UK Department for Environment, Food and Rural Affairs. **Most PAM data do not flow to EMODnet Biology or any other international data portal,** except for the Baltic data that flows to the HELCOM repository. Metadata about the type of sensor and the settings used, and metadata about the deployment of the sensor are necessary but **currently not included** in the data available via EMODnet Biology. This metadata information should include the period of good quality recording, latitude, and longitude.
- **EMODnet Data and Data Products Processing Level:** L5
- **DTO-BioFlow upgrade:** DTO-BioFlow will identify, compare and compile standards already used for observation, data exchange and reporting for specific species/species groups and specific habitats/regions by using existing data ingestion into HELCOM database as a starting point. We will also aim to develop script to generate maps. DUC 2 will use and apply these data.
- **Data repositories:** HELCOM



Other Networks

DTO-BioFlow is considering other networks and data sources. These include species occurrence data from global platforms not yet integrated into EMODnet Biology, gridded species distributions from various projects, reporting data relevant to EU Directives, as well as data from industry, citizen science, and literature.

New data sources becoming available: data projects selected in DTO-BioFlow first open call

DTO-BioFlow is also onboarding data holders of new or unavailable biodiversity data sources to contribute those data to the EU DTO and thereby enhance the flow of relevant biodiversity data. Nine data providers have been selected via the first open call in 2023, for their ability to facilitate the ongoing and long-term inclusion of previously inaccessible biodiversity data.

The chosen projects span a wide range of topics and institutions, ensuring a comprehensive approach to data collection. Using diverse methods such as citizen science observations, AI-assisted imaging, net trawls, sediment traps and benthic grabs, these projects focus on various organisms, including cetaceans, plankton, and benthos. Data is being collected from regions as diverse as the Arctic Ocean, the Mediterranean Sea, and the Atlantic Ocean, providing a broad spectrum of data sources.

Marine Megafauna Data for EU Digital Twin Ocean

Institution: Bangor University

Type of data: cetaceans, seals and other marine megafauna, citizen science observations

Collection area: British Isles, Nothern Sea, Scottish water, Irish Waters, Celtic Sea, Norwegian Sea, Bay of Biscay

Futurismo: linking whale watching tourism with cetacean research in the Azores

Institution: Futurismo Azores Adventure

Type of data: cetaceans, citizen science observations

Collection area: Azores – Canaries Madeira

KAIROS - ZooplanKton data from Arctic marIne time-seRies to understand biOdiversity dynamicS

Institution: Institute of Polar Science (CNR), Institute of Research on Terrestrial Ecosystems → (ISP-CNR) (IRET)

Type of data: zooplankton data

Collection area: Arctic Ocean, Svalbard archipelago

Integration of southeastern Mediterranean long-term biodiversity data into EU-DTO

Institution: Israel Oceanographic and Limnological Research

Type of data: bio-geo-graphical database

Collection area: Mediterranean waters of Israel, Southeastern part of the Mediterranean Levantine Sea.

Management and publication of Marine Characterisation Research Project data

Institution: Menter Mon

Type of data: observation data, photograph, acoustic data

Collection area: Anglesey - North Wales

Pipeline for biodiversity data from the British Oceanographic Data Centre (BODC) to the OBIS network and EMODnet

Institution: National Oceanography Centre

Type of data: plankton data

Collection area: Faroe Plateau, North Sea, Celtic Seas

Managing and publishing biodiversity data from Nord University

Institution: Nord University

Type of data: zooplankton

Collection area: North of the Artic circle - three Norway Fiords

Strandaanspoelsel (beach washup) Monitoring Project (SMP)

Institution: stichting ANEMOON

Type of data: citizen science observations

Collection area: North Sea

Plankton Imaging Data Flow: Establishing a European data flow for phyto- and microzooplankton data from automated AI-assisted imaging in flow analyses

Institution: Swedish Meteorological and Hydrological Institute Sweden

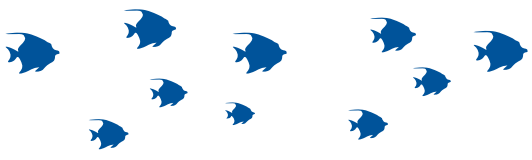
Type of data: Plankton imaging data

Collection area: Baltic Sea, Kattegat, Skagerrak


©Thanos Dailianis | HCMR

Image Credits:

Thanos Dailianis | HCMR



Funded by the
European Union