BIOcean5D

MARINE BIODIVERSITY ASSESSMENT AND PREDICTION ACROSS SPATIAL, TEMPORAL AND HUMAN SCALES

D3.1 Initial Data Management Plan

Co-funded by
the European Union

Grant Agreement number	101059915
Call	HORIZON-CL6-2021-BIODIV-0
Topic	HORIZON-CL6-2021-BIODIV-01-03
Type of Action	HORIZON-RIA
Project title	MARINE BIODIVERSITY ASSESSMENT AND PREDICTION ACROSS SPATIAL, TEMPORAL AND HUMAN SCALES
Project acronym	BIOcean5D
Deliverable title	Data Management Plan
Deliverable number	D3.1
Version	v1.0
Document status	Final
Type	Document, report
Diffusion	PU - Public
Related Work Package	WP3 Data to knowledge, a digital foundation for holistic marine biodiversity assessment
Task	T3.2
Full lead beneficiary	EMBL - European Molecular Biology Laboratory
Author(s) and Affiliations	Stéphane Pesant (EMBL), Pier Luigi Buttigieg (AWI), Meike Vogt (ETHZ), Sarah Schulz (EMBL), Kerstin Leberecht (EMBL), Amandine Nunes- Jorge (EMBL), Peer Bork (EMBL), Jean-Olivier Irisson (SU), Lynn Delgat (VLIZ), Leen Van de Pitte (VLIZ)
Project Officer	Victoria Beaz Hidalgo
Due date	M6 30.04.2023
Submission date	22.12.2023
Total number of pages	43
Keywords	Data flow, data management, data specifications and standards

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Table of content

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Version history

Version	Authors	Summary of changes	Date
0.1	Amandine Nunes- Jorge (EMBL), Kerstin Leberecht (EMBL), Sarah Schulz (EMBL)	Initial draft	14.08.2023
0.2	Jean-Oliver Irisson (SU), Leen Vandepitte (VLIZ), Lynn Delgat (VLIZ)	Input to initial draft	31.08.2023
0.3	Stéphane Pesant	Review and cleaned version	06.11.2023
0.4	Meike Vogt	Rereview and revision	10.11.2023
0.5	Stéphane Pesant and Pier Luigi Buttigieg	Rereview and cleaned version	16.11.2023
0.5	Kerstin Leberecht (EMBL), Sarah Schulz (EMBL), Josipa Bilic-Zimmermann (EMBL), Amandine Nunes-Jorge (EMBL), Peer Bork (EMBL)	Review, restructuring of chapters: introduction, FAIR, design of figures, layout formatting, abbreviation and references lists	06.12.2023

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

List of Acronyms

AAI	Authentication and Authorisation Infrastructure
ABNJ	Area Beyond National Jurisdiction
AI	Artificial Intelligence
AWI	Alfred Wegener Institute
BioSamples	EBI database storing and supplying descriptions and metadata about biological samples used in research and development by academia and industry
BODC	British Oceanographic Data Center
CA	Consortium Agreement
CC	Creative Commons
CC0	Creative Commons Zero
CC-BY	Creative Commons Attribution
CDIF	Cross-domain Interoperability Framework
CFC	Climate and Forecast Convention
CNRS	Centre National de la Recherche Scientifique
DCAT	Data Catalogue Vocabulary
DMP	Data Management Plan
DMV	Deliberative Monetary Valuation
DOI	Digital Object Identifier
DwC	Darwin Core
EBI	European Bioinformatics Institute
EC	European Commission
EcoTaxa	Web application dedicated to the visual exploration and the taxonomic identification of images of plankton
EMBL	European Molecular Biology Laboratory
EMBRC	European Marine Biological Resource Centre
EMODnet	European Marine Observation and Data Network
ENA	European Nucleotide Archive
EU	European Union
EurOBIS	European node of the Ocean Biodiversity Information System (OBIS)
ETHZ	Eidgenössische Technische Hochschule Zürich
EV	Environmental Variable
FAIR	Findable Accessible Interoperable Reusable
GA	Grant Agreement
GBIF	Global Biodiversity Information Facility
GDPR	General Data Protection Regulation
GNU	GNU's not Unix
GOOS	Global Ocean Observing System
GSC	Genomic Standards Consortium
HUPO	Human Proteomics Organisation
Ifremer	Institut France pour la Recherche et l'Exploitation de la Mer
IIF	Image Interoperability Framework
INSDC	International Nucleotide Sequence Database Collaboration
IOC	Intergovernmental Oceanographic Commission

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

IODE	International Oceanographic Data and Information Exchange
IP	Intellectual Property
IPR	Intellectual Property Rights
LIMS	Laboratory Information Management System
MARCO-BOLO	MARine COastal BiOdiversity Long-term Observations
MARS	Model for Applications at Regional Scale
MARUM	Zentrum für Marine Umweltwissenschaften der Universität Bremen
MB	Megabytes
MBA	Marine Biological Association
MetaboLights	EBI database for Metabolomics experiments and derived information
MSFD	Marine Strategy Framework Directive
MIMARKS	Minimum information about a marker gene sequence (MIMARKS)
MIME	Multipurpose Internet Mail Extensions
MIxS	Minimum Information about any Sequence
MGnify	EBI Microbiome analysis resource
ML	Machine Learning
NCBI	National Center for Biotechnology Information
NVS	NERC (Natural Environment Research Council) Vocabulary Server
OBIS	Ocean Biodiversity Observation System
OBO	Open Biological and Biomedical Ontology
OBON	Ocean Biomolecular Observing Network
ODIS	Ocean Data and Information System
ORCID	Open Researcher and Contributor ID
OS	Open Science
PANGAEA	Publishing Network for Geoscientific and Environmental Data
PID	Persistent Identifier
PMB	Project Management Board
PROV	Provenance metadata standard
QC	Quality Control
ROME	Network of Integrated Environmental Microbiology Observatories
SeaDataNet	Distributed Marine Data Infrastructure for the management of large and diverse sets of data deriving from in situ of the seas and oceans
SBE	Seascape Belgium
SDM	Species Distribution Modelling
SRA	Sequence Read Archive
SU	Sorbonne University
TREC	TRaversing European Coastline
UK	United Kingdom
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
URI	Uniform Resource Locator
VLIZ	Vlaams Instituut voor de Zee
WORMS	World Register of Marine Species
WP	Work Package
WP1	Exploration to fill the marine biodiversity/ecosystem knowledge gap
WP2	

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

WP3	Understanding biodiversity in critical marine habitats and their keystone holobionts
WP4	Data to knowledge, a digital foundation for holistic marine biodiversity assessment
WP5	New theories for marine biodiversity, ecosystem function, and their relationships
WP6	Monitoring human impacts on marine biodiversity and modelling future ocean health
	Assessing biodiversity value and public values of marine natural capital for improved protective strategies

List of file formats

Proprietary formats are marked with an asterisk.

BXR4	Flow Cytometry file format*
CSV	Comma-Separated Values
DCIMG	Hamamatsu DCAM (Digital Camera) Image File*
EML	Ecological Markup Language
F90/95	Fortran 90/95 source files
FASTA	FAST-All
FASTQ	File format for sequences with quality scores
graphML	XML (Extensible Markup Language)-based file format for graphs
HDF	Hierarchical Data Format
HTML	Hypertext Markup Language
JL	Julia
JLD	Julia HDF5 Data
JSON-LD	Javascript Object Notation linked Data
JSON LR	Javascript Object Notation Left to Right
JPEG	Joint Photographic Experts Group
LMD	Linear Mode Data
LOD	Linked Open Data
NC	Numerical Control
netCDF	network Common Data Form
NPZ	array zipped archive of files named after the variables they contain
PNG	Portable Network Graphic
PY	Python script file
RAW	uncompressed and unprocessed image data
RDATA	R (statistical computing software) data
RDS	R (statistical computing software) Data Serialization
RSK	risk project data associated with the RiskMan software
TIF	Tag Image File Format
TSV	Tab-Separated Values
TXT	Text document that contains plain text in the form of lines
WAV	Waveform Audio File Format
XLSX	Microsoft Excel Spreadsheet

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Executive summary

The primary objective of BIOcean5D's data management approach is to support the integration and distribution of the project's broad range of (meta)data. In doing so, this data management plan will facilitate the development of models and indicators to understand and predict how biodiversity responds to increasingly intertwined natural and anthropogenic pressures. This approach is in line with the project's overarching ambition to establish an understanding of ocean biodiversity to support key stakeholders in valuing, protecting, and restoring marine biodiversity for the benefit of life on Earth.

This document describes the management approaches BIOcean5D will use in the integration and harmonisation of new and existing biodiversity data and knowledge from other EU, international and national projects and from long-term ecosystem and socio-ecological research infrastructures.

It describes the scientific data types produced or reused in the project and their sources, their (meta)data requirements and the infrastructures selected by the project to curate, archive and access data. It also presents the initial concepts of the BIOcean5D data flow.

This is the first version of the project's DMP, which will be regularly updated as the project develops and new opportunities arise in order to increase the FAIRness and impact of the project's data.

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

1. Introduction

1.1. Project's objectives and ambition

BIOcean5D is a Horizon Europe co-funded project that unites 31 institutes with expertise in molecular/cell biology, marine biology, sequencing, and modelling as well as in economic valuation and environmental legislation.

The project focuses on coastal ecosystems (figure 1), including estuaries and coral reefs, where marine biodiversity is highest and most threatened, but it will also explore deep-sea and open ocean habitats, including connectivity between Areas Beyond National Jurisdiction (ABNJs) and coastal/regional seas mediated by highly migratory animals (e.g. tunas, swordfish, sharks, mammals).

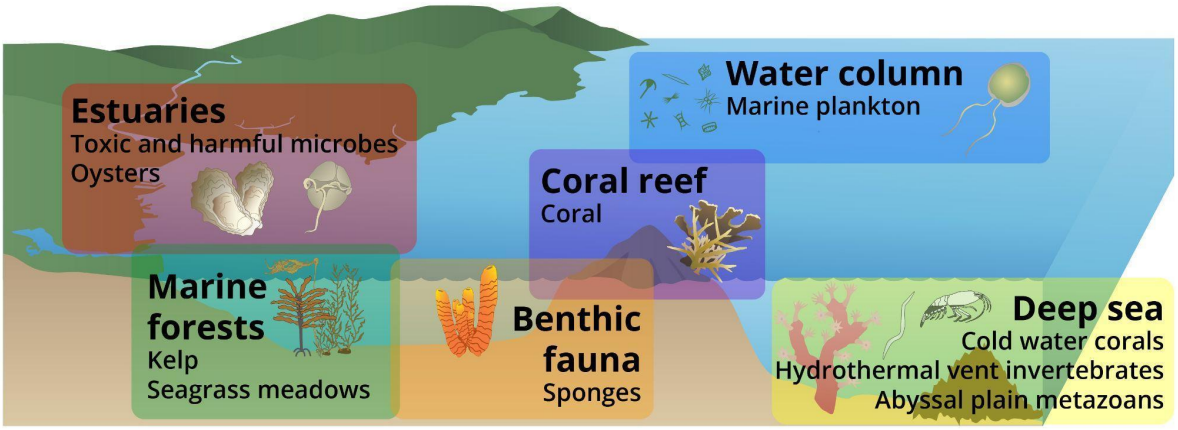

Figure 1: Overview of the ecosystems targeted in BIOcean5D

The objectives of the project are:

1. To close the gap in knowledge on the structure, dynamics, and evolution of marine biodiversity within and across marine ecosystems

2. To understand the drivers and mechanisms of biodiversity changes and degradation, assess and predict marine ecosystem health, and provide quantitative eco-systemic tools for marine biodiversity policy implementation

3. To develop and apply tools integrating the dynamics of functional biodiversity to measure the financial and non-financial value of marine life and ecosystem services and share these new concepts of marine biodiversity services to societies

To achieve these objectives, the project aims to build a unique suite of technologies, protocols, and models allowing holistic (re-)exploration of marine biodiversity, from viruses to mammals, from genomes to holobionts, across multiple spatial and temporal scales stretching from pre-industrial times to today. BIOcean5D is therefore a data intensive project that will not only generate new data, but also reuse existing data from decadal to centennial collections of European marine stations, and from recent major ocean biodiversity surveys. As such, biodiversity data is inherently heterogeneous, including sequence information, imagery data, traditional microscopy and net counts, acoustics measurements, aerosol data as well as a wide range of environmental and biogeochemical context data. Consequently,

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

biodiversity data are multi- and trans-disciplinary, stewarded by diverse organisations, and widely scattered. As noted here and in other projects, high fragmentation of data acquisition, handling, and storage inevitably create problems in data management and delivery, restricting interoperability (at different levels/scales) and (in the marine realm) limiting opportunities to advance knowledge on coastal processes and resource management.

The ambition of BIOcean5D is to gather and broker all digital assets of the projects in an open-access data hub at EMBL, using agreed international standards in terms of metadata, formatting and taxonomical references, so they can be delivered to a range of long-term data repositories such as EurOBIS, taking advantage of the expertise of project partners involved in data standardisation (e.g. WORMS at VLIZ), data curation and data archiving international trusted infrastructures (e.g. MGnify, ENA), in direct alignment with the digital knowledge management framework of the UN Decade of Ocean Science for Sustainable Development Implementation plan (UNESCO-IOC (2021)).

BIOcean5D data and digital assets will be relevant for a wide range of user communities and research fields, including, but are not limited to:

- The ocean, climate modelling and marine biodiversity research community, with its applications in marine microbiology, ocean biogeochemistry, theoretical ecology and marine macroecological and climate impact research;
- Policy makers using biodiversity data to guide decision making;
- Industrial and commercial actors, using biodiversity data to assess bio-prospecting potential, environmental or climate change impacts, sustainable development options, to quantify threats;
- Scientific groups outside of the marine and biodiversity communities, pursuing analogues to their interests in the ocean or seeking to build cross-disciplinary links;
- Coastal marine ecosystem managers engaged in ecosystem management and conservation actions, carbon mitigation or health assessments;
- The general public, powering awareness campaigns with rapidly ingestible (meta)data.

1.2. Project's commitment to Open Science and FAIR data

The BIOcean5D consortium is committed to implementing transparent and reproducible research. To deliver on this commitment, practices aligned to the Open Science (OS) framework and FAIR Principles have been integrated into the project work plan. These practices will be implemented through:

1. Detailed **data management planning**: this present Data Management Plan (DMP) describes the data collection, processing and archiving, following the template provided by the European Commission (EC), and in line with with the DMPs of other past and ongoing European Union (EU) funded projects, such that data transfer and interoperability between sister projects is guaranteed and compliance with international standards and developments is secured. The project's DMP will be continuously updated as the project develops and opportunities to increase the FAIRness and impact of the project's data clarify.

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

2. **Commitment to open and FAIR data**: (meta)data produced by BIOcean5D will be published in open access, trusted archives including European Nucleotide Archive (ENA), MetaboLights and PANGAEA, and made available to information systems such as EMODnet, OBIS and GBIF. Furthermore, the project will commit specific resources to the standardisation of (meta)data in compliance with prevailing standards in biodiversity and oceanography to promote discovery, reuse and re-analysis as broadly as possible.

3. **Open access to materials**: Whilst most samples will be fully utilised by the project's experiments, reference samples will also be collected and stored within the EMBRC infrastructure and/or the EMBL sample storage. The BIOcean5D consortium will make samples and materials collected and generated in the implementation of BIOcean5D openly accessible to the wider research community, whenever this is possible. To this end, all reference samples will be managed in a Laboratory Information Management System (LIMS) developed and maintained by partner EMBL (Heidelberg). The LIMS will reference the associated provenance metadata archived at BioSamples (EMBL-EBI).

4. **Rigorous and complete metadata collection** supporting multiple reuse scenarios: BIOcean5D will use a flexible framework to extract and reformat relevant metadata, so that it suits all major international data repositories targeted by the data types generated in the project. Metadata files will be served in JSON-LR format, and biological information will follow Darwin Core (DwC) standard notation for seamless integration with OBIS (Ocean Biodiversity Observation System), GBIF, EcoTaxa and other data type specific repositories. To this end, a metadata catalogue with machine-readable content will be produced.

5. **Digitally transparent provenance:** BIOcean5D will enhance trust in its digital assets through a FAIR-aligned identification of its actors, processes and methodologies used to create, modify, curate, redact and release metadata. To this end, each data-providing individual and each institution producing, curating, servicing or handling data needs to be identifiable with a unique identifier. We recommend BIOcean5D experts to provide their ORCID identifier (https://orcid.org), and for institutions to register themselves with the IODE-ODIS (International Oceanographic Data and Information Exchange - Ocean Data and Information System) Catalogue of Sources (https://catalogue.odis.org). Each digital asset should further ensure that EC Funding is acknowledged through the inclusion of the originators and project's name, acronym, and Grant Agreement number (GA# 101059915).

6. **Capacity building:** BIOcean5D will help all involved biodiversity scientists to meet the requirements of 21st century FAIR data by providing the relevant templates, guidelines, formats and expertise to enable scientists to integrate their data with today's high complexity ocean digital ecosystem, to analyse their data using open source software, to use cloud computing infrastructures for their analysis (e.g. the BlueCloud2026 infrastructure; https://blue-cloud.org/) where possible, and to transparently and systematically publish data arising along the scientific value chain from the raw data to models, analysis and output products using unique identifiers that guarantee access to products and developments in the long term. To this end,

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

FAIR data practices will be evaluated at multiple instances throughout the project and feedback will be directly communicated back to all parties involved. The feedback and evaluations will be the object of the D3.3: Biodiversity data assessment and roadmap towards international standardisation and operationalization (AWI).

1.3. High level principles of project knowledge management

The GA and Consortium Agreement (CA) define the main approaches regarding the ownership, protection and access to key knowledge like Intellectual Property (IP) and data. This section complements the D8.1: Project management handbook and knowledge management strategy. The major aspects are:

- **Confidentiality:** Each partner will treat results and background data information from other partners as confidential unless otherwise stated and not disclose it to third parties unless the information is publicly available (CA Article 8.4.3);
- **Intent to upload data to repositories:** Results-/ Background-/ Data-owners will notify the partnership of their planned intent to upload data sets to open-access repositories following the same prior notice procedure as is set up for publication of results (a minimum of 45 days), as detailed below; Data underpinning a scientific publication should be deposited at the latest at the time of publication, and in line with standard community practices.
- **Intellectual Property and pre-existing know-how:** Pre-existing Know How: Each partner is, and remains, the sole owner of the IP associated with their pre-existing know-how. The partners have identified and listed in the Attachment 1 of the CA the pre-existing know-how (so called “Background”) over which they grant access rights for the project under the conditions set out in the CA Article 9.1.;
- **Ownership and protection of results:** In general, partners which generate (meta)data and information products, services, or other outputs will retain ownership of these assets according to the legal frameworks their organisations operate within. **Ownership rights are often waived through open, public domain licences or through submission to archival services whose usage agreements require the relinquishing of ownership rights.** Protection within the consortium’s digital commons will be implemented appropriately. When the result is the outcome of work carried out by two or more partners and their respective share of the work cannot be ascertained, joint ownership will be agreed between the partners as it is established in the CA Article 8.2. If a partner wishes to assign any knowledge gained in the span of the project to a third party, they should do so while observing the conditions set out in the BIOcean5D CA, especially article 8-Results, 9-Access rights and 10-Non-disclosure of information, and should inform the other partners and request their written consent, which should not unreasonably be withheld. All information, agreements, and caveats/special conditions concerning the ownership of digital assets must be clearly included in metadata records associated with that asset;
- **Access Rights:** (see CA article 9) Partners grant to each other royalty-free access rights to knowledge generated in the project and to the Background knowledge identified in the Attachment 1 of the CA. Any Party may add additional Background throughout the lifetime of the project, provided they give written notice to the other

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Parties. Approval of the General Assembly is needed should a Party wish to modify or withdraw its Background listed in Attachment 1 of the CA;

- **Patents:** partners who own knowledge suitable for patent are obliged to make applications for patents or similar form of protection and shall supply details of such application to the other partners. Information relating to patents that have been registered must be submitted under the 'IPR' section of the EU Funding and Tender Opportunities Portal;
- **Use and Dissemination:** If dissemination of knowledge, information, and data does not adversely affect its protection or use and is subject to legitimate interests, the partners shall ensure further dissemination of their own knowledge as provided under the GA (see Article 17 and its Annex 5). Beneficiaries must ensure open access to peer-reviewed scientific publications relating to their results. This includes articles and long-text formats, such as monographs and other types of books. Immediate open access is required i.e. at the same time as the first publication, through a trusted repository using specific open licences. When choosing the publishing venue and the repository, beneficiaries/authors must keep in mind that licensing requirements, metadata requirements and validation requirements must also be complied with at this time. Metadata on all such products - including links to openly available data or contact information for products that are not disseminated - must be made available to the ocean community via IOC-UNESCO IODE's ODIS.

Additionally, this BIOcean5D DMP is based on the following regulations:

- This DMP follows **the definition of the obligations/mandatory practices** in the Grant Agreement Article 16-IPR, Background and Results, Access rights and rights of use and Article 17-Communication, dissemination and visibility, in particular the Open Science section, together with their complement in Annex 5. The elaboration of the DMP based on these definitions will allow BIOcean5D partners to address all issues related to IP protection and data in line with the obligations/mandatory practices. All IPR related information will be reported in the dedicated tab of the continuous reporting sheets internal to the project, which will then be reported in the EC continuous reporting platform (all details on reporting can be found in the D8.1).
- The consortium will comply with the requirements of the Data Protection Laws as defined by the CA in Article 1.2- Additional Definitions as well as GA Article 14-Ethics and Values and 15-Data protection on **the protection of natural persons with regard to the processing of personal data** ;
- Procedures surrounding **data collection, storage, access, sharing policies, protection, retention and destruction are in line with EU standards** as described in the BIOcean5D GA and CA, particularly GA Article 20.1 Keeping records and supporting documents; GA Article 16 Intellectual Property Rights; CA Article 8 Results; CA Article 9 Access Rights; CA Article 10-Confidentiality".

This DMP outlines the project's initial approach to fulfil these commitments. Subsequent versions will be released when significant changes arise, as the project develops and feedback is acquired through the consortium and internal evaluations (Task 3.1). A final DMP will be released at the close of the project, documenting the final approach used in BIOcean5D (D3.5 - M46).

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

This DMP and its updates are binding for all partners in this project. All partners are responsible for the standardisation, documentation and dissemination of their BIOcean5D outputs and data products according to the guidelines as outlined in this document.

2. Data Summary

2.1. Documenting data sets

In line with the project's commitments described in chapter 1, the following documentation/metadata will be compiled for each data set collected, harvested, processed, and/or generated in the project and prepared for encoding/serialisation in formats to be specified by WP3.

At this stage, WP3 has gathered initial intelligence on the data types and contextual metadata being harvested by the Consortium (Table 1, Annexes 1 and 2). Additional data sets may be identified and added to future versions of the DMP as necessary. Partners are required to maintain rich information on at least the elements listed as bullet points at the end of this paragraph. This information will be recorded in the "Data sets" and "IPR" tabs of the continuous reporting sheets (details on the reporting procedure available in D8.1) and will be thereafter recorded in the corresponding sections of the EC continuous reporting platform. Any information listed below but not captured by the continuous reporting system will be included in the data set reports/associated deliverables.

- **Data provenance:** A complete account of the data's life cycle, including a) rich information on the data's journey from generation onwards and b) information on what BIOcean5D partners and their collaborators have done to the data in their possession (e.g. quality control, curation, uplift);
- **Data ownership:** Description of the data owner (organisation and/or individual), to include the owner's full name, originating work package (if applicable), task and activity, the responsible researcher(s)' name and the primary contact details for enquiries regarding the data;
- **Description of IP considerations:** statement summarising whether special measures are needed to protect IP and whether an evaluation by the BIOcean5D Project Management Board (PMB) is needed;
- **Data summaries:** One or more textual abstracts (authored by one or more contributors) describing the data set;
- **Data identification:** All names, aliases, identifiers, and other forms of identification used to reference this data;
- **Diagnostic/technical metadata:** metadata stating, using appropriate international standards (e.g. MIME types for media), the type and format of the data, the expected overall storage size of the data, etc;
- **Machine-actionable licence:** each output must specify the applicable licence including a URI (e.g. Creative Commons (CC), GNU Public License, etc);

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

- **Metadata to enable implementation of the FAIR Principles and 5-star Open Data[1]:**
  Measures must be taken for each of the FAIR Principles and sub-principles in a technically sound manner. WP3 should be proactively consulted to ensure that each partner's implementation is compatible across the project and its key digital stakeholders. Measures must be taken to advance as far along the 5-star Open Data plan as possible. WP3 should be proactively consulted for clarifications and alignment with other partners. Findability: Description of domain-relevant repositories, whether the data will be made identifiable by a standard identification mechanism and the type of metadata that will be provided;
- **Allocated resources:** Description of the estimated costs required to make the data FAIR and how these costs will be covered (e.g. covered by work package budget);
- **Security and confidentiality considerations:** Full description of the data security and confidentiality measures in place and/or required when handling a given data set, including confirmation of plan for recovery, secure storage and protection over the transfer of sensitive data;
- **Ethical considerations:** Any potential ethical issues must be noted such as risks to endangered/vulnerable/rare species, sensitive habitats, indigenous rights, Access and Benefit Sharing agreements, and risks of misuse.

2.2. High level data flow

Case of reuse of historical data:

During its first phase, where new measurements from the TREC expedition and other activities (in WP1, WP2, WP3, WP4, WP5, WP6) are not yet available, BIOcean5D will focus on aggregating and harmonising existing data sets. Partners will collect, collate, harmonise and augment data sets to support their tasks, which will be made available through the project's data hub. Furthermore, the project will draw upon a wider range of data assets including new and existing time series data, model codes (ocean biogeochemical, climate or species distribution models (*.R; *.f90), biogeochemical and climate model outputs (*.nc; e.g., from the DARWIN model or the CMIP6 model suite) food web models such as EwE, high-trophic-level IBMs such as Ev-OSMOsE), model outputs, as well as integrated, climatological *in situ* observations for modelling and analysis purposes. Considerable effort has and will be made with regard to the documentation and metadata standardisation of such products. WP3 will gather, and WPs 4, 5 and 6 will generate new digital assets for distribution within the project. To this end, the following procedure (adapted from the MARCO-BOLO DMP[2] to improve interoperability across the biodiversity actions funded under the EU Horizons programmes) will be used to optimise data flow and integration into WP1 and WP2 data products.

WP1-WP6 will either gather or generate external (meta)data sets aligned to their tasks from a diverse range of sources. WP3, in consultation with the BIOcean5D community and international stakeholders, will establish how (meta)data gathered in this manner shall be internally standardised to align to the requirements of the project's key stakeholders and international digital ecosystems and interoperability initiatives. These specifications will guide

[1] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data
[2] MARCO-BOLO D7.2 - Version v1.0.0 - https://doi.org/10.5281/zenodo.8208410

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

the data generators in uplifting and harmonising their harvested digital assets for improved integration and sustainable and maximally discoverable dissemination. The following cases are anticipated:

- Should the external data already be archived in long-term, trusted repositories with public accessibility, then its metadata will be harmonised inline with the ODIS Architecture[3] and, consequently, with the emerging CDIF conventions[4].
- Should the external data not be safely archived in long-term, trusted repositories with public accessibility, then - where licensing conditions and permissions allow - the BIOcean5D WPs handling that data will encourage its deposition by the data originator or submitting BIOcean5D partner in the appropriate long-term archive (e.g. INSDC databases for sequence data) or in the BIOcean5D Zenodo space, to allow reproducibility and transparency of all knowledge generation.
- Should 1) the external data not be archived as noted above, and 2) its licensing conditions or other agreements/restrictions restrict upload therein by BIOcean5D partners, its metadata must still be harmonised, made available to declare its use and properties, as well as the agreements under which the relevant BIOcean5D partners secured access and usage rights to the data.

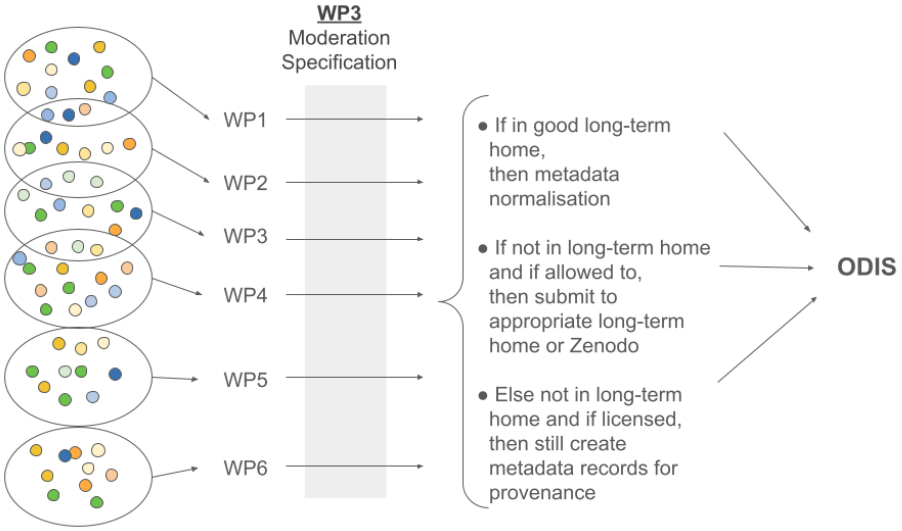

Figure 2: Data exchange between BIOcean5D and ODIS for newly generated and reused data

Case of newly generated data:
Additionally, BIOcean5D Partners will also generate data and metadata during the project. This may include primary data from biodiversity observation, but also metadata describing actions they have taken on those assets as well as legacy/external data. Similar to the case above, WPs 1-6 will - following the guidelines generated by WP3 and/or proactively seeking

[3] https://book.oceaninfohub.org/
[4] https://zenodo.org/records/7682399

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

their guidance - pool this data into the project's data hub and into trusted repositories for long-term archiving, according to the project's data flow chart (Figure 3). In consultation with stakeholder communities and the WP participants themselves, (meta)data flows will be established between the project's data lake and 1) long-term repositories such as OBIS and the INDSC, 2) IOC-UNESCO's ODIS, and 3) Zenodo for the archival of data that does not yet have a clearly mandated home (e.g. acoustics).

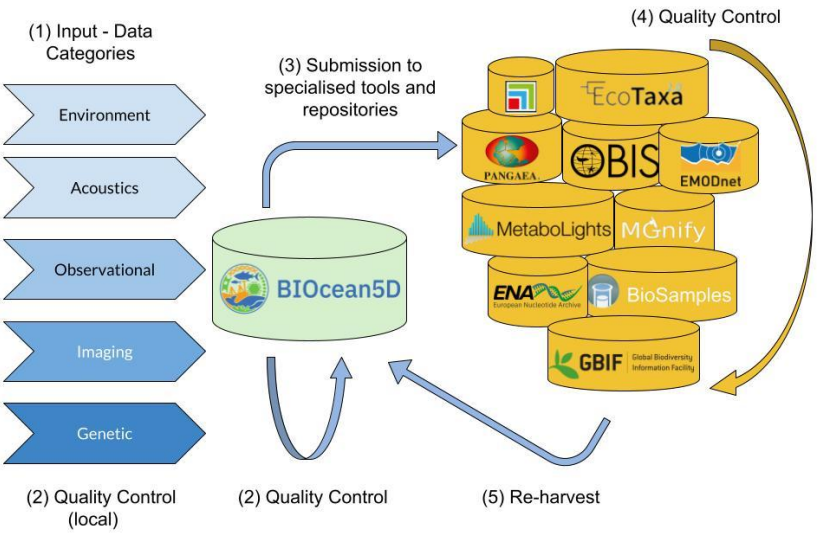

Figure 3: Data flow diagram in BIOcean5D for newly generated data. (Meta)data that is either gathered from existing sources or generated during BIOcean5D will be pooled in the project's data hub, with partners standardising their contributions along WP3 guidance (1). An initial round of quality control (2) will be performed before the (meta)data is submitted to its appropriate long-term archive (3; e.g. EBI, OBIS, PANGAEA) with a delay in the order of 3-6 months. Those archives will perform their own quality control process (4) and request changes as needed. Once the (meta)data have passed quality control and are issued definitive permanent identifiers on the web, BIOcean5D will reharvest the (meta)data for use in the generation of data products and reproducible and FAIR scientific analyses (5).

2.3. Data types and formats

The biodiversity data generated in BIOcean5D consists of a wide range of data types, subjects and formats. These data types encode descriptions of genetic elements (metabarcoding, metagenomics, -transcriptomics and proteomics), (quantitative) organismal and environmental imaging, mammal and bird sightings, active and passive acoustics, physical properties, chemicals, remotely sensed satellite derived products, citizen science activities, and socio-ecological phenomena.

A general overview of the data types being generated or reused can be found in Table 1 below including their data formats. More details on each data type can be found in the next subsections.

Table 1: Overview of data types with corresponding data formats being generated during BIOcean5D

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Data subject	Data format
Metagenomic & Metabarcoding	FASTA, FASTQ, TXT, CSV
Metagenomic and metabarcoding – Derived data formats (taxonomic annotations, sequence count tables, etc)	CSV
Environmental - Chemical Analysis	CSV, XLSX, RSK
Metabolomics - Lipodomics	CSV
Flow Cytometry	TXT, DCIMG, LMD, BXR4, CSV
Environmental - Aerosol size distribution	CSV
Environmental – Meteorological data	CSV, NC
Environmental - Lagrangian Diagnostics	CSV, NC
Acoustics - Raw data	WAV
Acoustics - Derived data	CSV, JPEG
Imaging - Raw data	JPEG, PNG, TIF
Imaging - Derived data	CSV
Traditional – Raw Data	CSV
Traditional – Derived data	CSV, NC
Metadata	CSV, XLS, TXT, TSV, JSON-LD
Modelling – Model codes and analysis software	F90, PY, R, JL, Java, etc
Modelling – Model outputs	CSV, TXT, NC, RDS, graphML, NPZ, RDATA, HTML, JLD, HDF, netCDF, etc.
Natural accounting methods	CSV, TXT
Literature review methodology	CSV, TXT
Human survey data - Quantitative raw data	CSV

As in MARCO-BOLO's DMP, we do not expect to issue prescriptive formats (i.e. restrict our partners) at this stage of the project, however the following principles will be followed in selecting data formats for data harmonisation, generation, and delivery to stakeholders:

- No new data formats shall be generated by BIOcean5D, as mandated by authorities/ custodians of MSFD indicators;
- Formats shall not require proprietary or a single form of software (e.g.created and maintained by a single organisation) to access and (re)use;

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

- Formats must default to well-adopted community standards that - as far as possible - comply with the FAIR Principles and the 5-star Open Data Plan;
- Formats should be immediately identifiable via a universally recognised extension and/or internal metadata (e.g. file headers, magic numbers[5])
- Formats must be non-proprietary with open, complete documentation and specifications publicly and freely (at no cost) available;
- Unless specifically required, with reasoning documented, formats must be unencrypted;
- Unless using a lossless and non-proprietary approach, formats must be uncompressed;
- Formats shall use common, non-proprietary character encodings.

In the following subsections, brief descriptions of the expected thematic (meta)data types will be provided for orientation.

2.3.1. Collection of cross-type metadata

Given the diversity of BIOcean5D's data types, it is essential that the metadata about them (including rich provenance) is stored in a homogenous manner, which implements the FAIR Principles in a way that data systems across the ocean community can make use of. To do so, metadata gathered on all the data types below will be serialised in JSON-LD using schema.org semantics (with less broadly understood semantic markup embedded). This rendition of our metadata will be aligned to the IOC-UNESCO Ocean Data and information System (ODIS) specifications[6], which - in turn - contribute to the UN Ocean Decade's OceanData2030 Programme through alignment with the Decade's Data and Information Strategy[7] and its subsequent and emerging Implementation Plan[8]. Particular attention shall be given to aligning our metadata with ODIS specifications being developed for the Ocean Biomolecular Observing Network (OBON)[9] and Marine Life 2030[10]. Similar approaches are being adopted by Horizon projects such as MARCO-BOLO[11] (described in its DMP[12]) and WorldFAIR[13] (described in D11.1[14]).

2.3.2. Sequencing data

Sequencing data will come from a variety of sources, a detailed list can be found in Annex 1. Samples will be sequenced through a number of different techniques such as metabarcoding, single-cell transcriptomics and whole genome sequencing. The raw and quality-controlled files will be in the standardised fasta/fastq formats. Occurrences derived from these genetic data will be documented in Darwin Core Archives (containing CSV files).

2.3.3. Imaging data

[5] https://en.wikipedia.org/wiki/Magic_number_(programming)
[6] https://book.oceaninfohub.org/index.html
[7] https://unesdoc.unesco.org/ark:/48223/pf0000385542.locale=en
[8] https://github.com/iodepo/OceanDecade-dsig/blob/main/statements/enhance-discoverability.md
[9] https://www.obon-ocean.org/
[10] https://marinelife2030.org/
[11] https://marcobolo-project.eu/
[12] https://doi.org/10.5281/zenodo.8208410
[13] https://worldfair-project.eu/
[14] https://doi.org/10.5281/zenodo.7682399

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Imaging data within BIOcean5D spans macroscopic as well as microscopic scales down to single-cell flow cytometry images. As it is a quickly evolving field, especially on the technological side, BIOcean5D is pursuing a technology-open policy to try to encourage the community to use standards that are best suited for each type of imaging. All partners will be strongly encouraged to reuse and extend (where applicable) internationally adopted and advanced metadata standards for image- and video-based data, particularly the International Image Interoperability Framework (IIIF) as well to work around proprietary formats and ensure that their data meets the open science standards.

2.3.4. Metabolomics and proteomics data

Metabolomic and proteomic data is planned to originate from *Platynereis dumerilii* atoke and epitoke samples, as well as the kelp and seagrass environments that they can be found in, which will be sampled as part of the TREC expedition as well as regular seasonal sampling stations. Mass spectrometry data is currently planned to be using the csv format for both raw and curated data, meeting the open science standards.

2.3.5. Chemical data

Chemical analysis encompasses a number of different measurements of biological, physical and/or chemical nature such as particle counts within the aerosol measurements, organic and inorganic pollution or granulometry of sediments. Measurements are planned to be provided in the open CSV and TXT formats.

2.3.6. Human economics data

The research approach in human economics consists of quantitative data (choice experiment) and qualitative data (discussions in focus groups). The quantitative raw data will be provided as CSV; the qualitative raw data will be text-based. For the literature review, natural accounting and fuzzy cognitive-models, they will be provided in CSV and text formats.

2.3.7. Acoustics data

Bio-acoustics data will be sampled during the TREC expedition. Raw acoustics files will be in the WAV format. Quality control includes visualising leading to processed files being in the .csv and .jpeg formats. As raw acoustics data are too large for available public repositories, only processed data and metadata as well as snippets of recordings which are used to generate downstream products and key findings are planned to be uploaded to repositories. The full raw acoustics data files will be available on hard drives upon request. Occurrences and diversity derived from acoustics data will be documented in Darwin Core Archives (containing CSV files).

2.3.8. Modelling data products

BIOcean5D will produce multiple types of mechanistic and statistical models, as well as model output products that are to be delivered to the project. Models themselves are prone to a substantial diversity in code formats as well as output products and pertain to a variety of research fields with divergent codes, data documentation standards and dissemination practices, including those models used in choice modelling, qualitative content analysis, species distribution modelling, network modelling, as well as marine ecosystem and climate, trait-based and trade-off modelling.

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

BIOcean5D will encourage all modellers to use open source software for all new code and analysis where possible, and to submit their model code in the native programming language used in their field (Fortran, Python, R, Julia, etc) to a GitHub repository which has a tagged release (or a series of such) archived in Zenodo using native GitHub-Zenodo integrations[15]. Where possible, geo-referenced model outputs will be submitted to EMODnet using the netCDF format and Climate and Forecast Conventions (CFC) metadata. Non-georeferenced data such as network graphs, connectivity matrices, ecological niche estimates etc. will be recorded in their native formats (graphML, NPZ arrays, RDATA files etc) and will be submitted to Zenodo or a field specific research data repository for which a DOI can be obtained. Geo-referenced model outputs will be accompanied by a readme file in text format with metadata compatible with EMODnet and Zenodo submissions. As with all other digital assets, dedicated metadata for each model will be prepared for harvesting by ODIS.

Table 2: Overview of modelling activities, corresponding data types and formats, output data type and likely field-specific data repository, generated during BIOcean5D

Model type	Expected model outputs	Model code formats	Geo-referenced output formats	Preferred data repository
Statistical niche modelling	Model objects Environmental predictor data Biological response data Response curves Extrapolated maps of marine plankton and fish biodiversity and abundance or biomass distribution patterns as well as derived diversity metrics	R, some Python R, Python, Jupyter notebooks, RShinyApps *.RData, *.R	(netCDF, *.nc) Associated raw data: *.csv	GitHub, Zenodo EMODnet (fields) Raw data: Zenodo, Pangaea or (Eur)OBIS/GBIF Output fields/adat: Pangaea
Individual-based models	Model code Functional traits and distributions Lagrangian trajectories Ecological theory	F90/95 *.py, *.R	(*.nc)	GitHub
NUM modelling approach	Model code Biodiversity-size relationships Metabolic and trait diversity	TBD	TBD	TBD

[15] https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

	Ecological theory			
Multi-layer network models	Model code Metabolic models Network graphs Ecological theory	Yes	No	GitHub, Zenodo
Regional to global scale marine ecosystem and climate models	Model code Maps of past, present and/or future plankton biomass distribution and diversity Maps of ocean physical and biogeochemistry tracers Maps of important ecosystem services Maps of past, current and future indicators of high-trophic-level species (e.g., exploited fish) abundance and diversity and associated ecosystem services (e.g. fisheries landings)	F90 (MITgcm-DARWIN) Java (Ev-OSMOSE)	netCDF *.nc netCDF *.nc	Yes Local open access ENEA repository (MITgcm-Darwin output fields), GitHub, CMIP6 archive (fields) GitHub, Zenodo
Multi-criteria risk assessment and predictive scenario model	Model: interactions and deep learning Interaction matrix Interaction probabilities Maps of vulnerability Maps of biodiversity hotspots Future projections of change	R, Python	Yes *.csv *.csv and lists netcCDF *.nc	Aquamaps
Models of biodiversity value	Models Indices	Yes	TBD	TBD
Natural accounting methodologies	Models Maps	Yes	TBD	TBD

2.3.9. Taxonomically resolved microscopy, net count and fish catch data

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Selected historic global compilations of phyto-, zooplankton and fish presence-absence, abundance, biomass and biodiversity data will be delivered to the project based on previous collections by the project partners. Raw observational data files will be supplied in CSV format using DwC standard notation. Gridded data will be provided in netCDF format and using Climate and Forecast Conventions (CFC) metadata.

2.4. Data size

As sampling and thus data generation is not yet finished, an exact data size cannot be stated at this stage of the project but is expected to range in tens to (low) hundreds of terabytes overall with acoustics and imaging expected to take up terabytes whereas chemical measurements are estimated to range in the low megabytes. Estimates of expected data sizes of each study of the project can be found in Annex 1.

2.5. Data exploitation

The knowledge collected in BIOcean5D will inform (i) new theories and models of marine biodiversity, as well as ecological and evolutionary dynamics and drivers, from both taxonomic and functional perspectives, (ii) a portfolio of novel prototype holistic indicators of marine ecosystem health, (iii) innovative methods for economic and legal valuations of marine biodiversity and services, integrating the dynamical and functional complexity of marine life. BIOcean5D will create a unique opportunity to bridge molecular/subcellular biology to organismal biology, theoretical ecology and econometrics, and marine complex systems to social sciences, toward the sustainable preservation of our oceans and seas.

3. FAIR data

BIOcean5D will ensure that all published research (meta)data[16] is made Findable, Accessible, Interoperable, and Reusable (FAIR principles – Table 3) within relevant data ecosystems[17], and that it is appropriately managed.

We plan to process and store raw data across the data types described in chapter 2 at the different partner institutes of the consortium where they will undergo first-order quality control. The resulting, curated data sets will be archived in the appropriate thematic repositories (e.g. the INSDC for sequence data) and interlinked via the BIOcean5D data hub, which will also hold full metadata records for each BIOcean5D data set. From this point, the data sets will undergo further work before they are published in open and trusted archives for long-term preservation and reuse.

Table 3: FAIR principles as described by Wilkison et al., 2016[18]

[16] excluding (meta)data which cannot be shared for legal or ethical reasons (e.g. personally identifiable data, data about endangered species)

[17] The FAIR Principles can be met in many ways, and two systems may - themselves - be entirely FAIR but also entirely disjoint/siloed from one another. Thus, the way that the FAIR Principles are implemented at the lower level (i.e. the standards chosen, the repositories used for archiving) will have considerable impact on whether BIOcean5D data is found, accessed, interoperable with, and reusable by the relevant stakeholders.

[18] The FAIR Guiding Principles for scientific data management and stewardship, Wilkinson et al., 2016 https://www.nature.com/articles/sdata201618

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

F	F1. (meta)data are assigned a globally unique and persistent identifier
	F2. data are described with rich metadata (defined by R1 below)
	F3. metadata clearly and explicitly include the identifier of the data it describes
	F4. (meta)data are registered or indexed in a searchable resource
A	A1. (meta)data are retrievable by their identifier using a standardised communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary
	A2. metadata are accessible, even when the data are no longer available
I	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
	I2. (meta)data use vocabularies that follow FAIR principles
	I3. (meta)data include qualified references to other (meta)data
R	R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage licence R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards

3.1. Making data findable, including provisions for metadata

The project will distinguish between raw data, which will be deposited in field specific, trusted repositories (e.g. EcoTaxa for imagery) or Zenodo repositories where trusted repositories do not yet exist (e.g. acoustics data), and derived products such as diversity fields, which will be recorded in BIOcean5D's data hub, in order to be delivered to more general repositories such as EurOBIS (e.g. abundance data), GBIF (e.g. presence/absence data), or EMODnet (e.g. geo-referenced model output). Models will be deposited in trusted code hosting platforms that allow version control and collaboration such as GitHub. Exhaustive documentation of pipelines/workflows (including software versions, parameters and commands executed) will be generated and provided together with the data sets in human and machine-readable formats e.g. via GitLab or GitHub. The list of depositories used in BIOcean5D is available in Table 4.

Table 4: List of Repositories used in BIOcean5D

Data type	Resource	Description of the resource
Data repositories		
Metadata	BioSamples (EMBL)	the sample provenance metadata archive
Sequencing data	ENA (EMBL)	the European Nucleotides Archive
European marine data	EMODnet (SBE)	the European Marine Observation and Data Network
Satellite data	Copernicus	the Earth observation component of the European Union's Space programme
Imaging data	BioImage Archive (EMBL) EcoTaxa (SU)	the Biological Image Archive the platform for the visual exploration and taxonomic identification of plankton images

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Geo-referenced data	PANGAEA (AWI/MARUM)	Data Publisher for Earth & Environmental Science
Mass spectrometry data	MetaboLights (EMBL), MassIVE	database for Metabolomics experiments and derived information
Data annotation		
Linkage of all sample-derived data and metadata	BioSamples (EMBL)	the sample metadata archive
Microbiome data	MGnify (EMBL)	the analysis and archiving platform for microbiome data
Plankton phenotype morphological imaging data	EcoTaxa (CNRS/SU, soon Ifremer)	the plankton image annotation platform
Data modelling		
Model codes	GitHub	Data archive for scientific software distribution
Model analysis scripts	GitHub	Data archive for scientific software distribution
Model output fields	EMODnet GitHub Zenodo	Geo-referenced data products will be delivered to EMODnet, non-georeferenced data products such as network graphs will be delivered to GitHub or Zenodo
Data dissemination		
Environmental data	PANGAEA (AWI/MARUM)	Data Publisher for Earth & Environmental Science
European marine data	EMODnet (SBE)	the European Marine Observation and Data Network
Ocean biodiversity and biogeographic data	OBIS (UNESCO)	Ocean Biodiversity Information System
Global biodiversity data	GBIF	the Global Biodiversity Information Facility
Ocean biodiversity and biogeographic data	EurOBIS	the European node of the Ocean Biodiversity Information System
Human survey data	Zenodo	Open Science platform

Persistent, dereferenceable Identifiers (PID) are critical in order to trace, link and reference complementary (meta)data. Most of the repositories listed assign each submitted data set a DOI as PID (with the exception of EcoTaxa) and allow the import of DOIs from existing data sets and are envisioned as long-term data archives.

Samples originating from the TREC and TARA Europa expeditions have been assigned PIDs via BioSamples. Other types of data will be assigned persistent identifiers from trusted registries such as EurOBIS and PANGAEA which issues DOIs for biodiversity and environmental data sets. Persistent identifiers will be associated[19] with provenance and

[19] either by their kernel information profiles or through linked metadata files

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

contextual metadata that comply as much as possible with prevailing standards, e.g. MIxS for omic data, DwC archive for biodiversity data in general, and ODIS-Arch for generic metadata.

To increase discovery and reuse of data sets, keywords will be used as part of the standard minimal metadata required as these are also used by open repositories as part of the submission process. As BIOcean5D data will be made available in such, keywords will be provided to optimise discovery and reuse of data sets after the end of the project. Where possible, these keywords will use semantic qualifiers to not only promote findability, but also semantic interoperability (Section 3.3).

Harmonised data sets with standardised metadata will be provided in accordance with the metadata standards in marine and terrestrial ecology. Metadata will be served in the open JSON-LD, CSV, EML (XML) and txt formats. Metadata will be made available via the project's data hub, which will relay BIOcean5D's catalogues to ODIS. All metadata will be harvestable by two main routes: 1) All high-level metadata will be made discoverable to the ODIS Federation and globally scoped resources such as Ocean InfoHub, which will also allow generic data set searches powered by Google and other uses of structured data on the web, 2) Where domain-specific repositories exist, the metadata will align with global biodiversity standards and conventions (e.g. DwC, MIxS), which will allow indexing via either a harvest or submission model.

BIOcean5D will rely on the range of key:value specifications from biodiversity and related ocean data standard- and convention setting organisations (e.g. GSC, TDWG, GOOS, IODE) as well as generic or thematic metadata standards (e.g. schema.org, DCAT, PROV) to establish a reasonable benchmark for richness. We will consider metadata richer when more of the relevant fields in these standards are accurately populated. In addition, any information our consortium experts believe essential to reproducibility or understanding of their digital products will be added to our metadata, using valid extensions of these standards (e.g. additionalProperty in schema.org, or MeasurementOrFact in DwC).

3.2. Making data accessible

All project data will be made accessible to the project consortium through the project's data hub, as soon as possible after quality-control has been completed. In the hub, internal Authentication and Authorisation Infrastructure (AAI) systems from the individual consortium members will be used to ascertain the identity of the person(s) accessing the data. Once data has been made openly and freely available, no identification process is envisioned.

In addition, before publication in a scientific journal, data used in the paper should be deposited in a repository and the accession number should appear in the paper. All repositories listed in 3.2-Table 4 are open access, do not need authentication and use open protocols (HTTP(S), FTP, ...), to retrieve the data. If an embargo is applied on the data to give time to publish or seek protection of the intellectual property (e.g. patents), the reason(s) and duration of the embargo must be specified, bearing in mind that research data is a public resource and should be made available as soon as possible. However, the duration of the embargo should not exceed 12 months, and metadata about embargoed data should be released within one month of its generation or harvesting.

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

All data as well as software, simulation models and code for statistical analyses will be licensed under a public domain dedication CC0. Exceptional cases may arise for sensitive data such as endangered or commercially valuable species, in which case the accessibility will be limited and the data will be submitted to restricted licensing.

No human sensitive/personal data is expected to be generated/reused as part of BIOcean5D with the exception for the socio-economics WP which uses surveys gathering personal but not sensitive data, and which will ensure handling of personal data in a GDPR compliant manner (see Ethics).

3.3. Making data interoperable

To make data and metadata interoperable, community standards, formats and methodologies will be used after assessing their fitness. All data producers are responsible to ensure a common minimal standard of (meta)data quality and completeness by following WP3 guidance before deposition or publication of their digital assets. To this end, templates with (meta)data and documentation standards are available for use within the consortium in the guidelines “Data sharing best practices” (*in prep.*). These templates will evolve following the continuous feedback circuit between WP3 experts and data generators.

ENVO is the standard ontology concerning environmental metadata and is being used by TREC/TARAEuropa for the 2023/24 expedition which will provide a third of the data used in BIOcean5D.

Sequenced, biomolecular samples will be described using Genomic Standards Consortium Standards (primarily the MixS checklists) as well as sample metadata checklist[20] to promote interoperability of the data. For metabarcoding the dada2 pipeline will be used to generate ASV tables from raw metabarcoding data.

Other data sets will adhere to discipline specific standards such as DarwinCore and BODC terms used inheaders for raw data, WoRMS & NCBI registries for taxonomy, DarwinCore (OBIS-ENV-DATA format) for biological entities occurrences, CFC for netCDF files, species distribution model documentation according to Zurell et al. 2020, standard species distribution model calibration, evaluation and QC according to new community standards implemented into new automatic pipeline currently being built in EU project BlueCloud2026 with input from AtlantECO and BIOcean5D Species Distribution Modelling teams for climate models, and the HUPO Proteomics Standards Initiative for mass spectral data.

Additionally, where relevant and accurate, semantic resources such as the vocabularies/thesauri present in the NERC Vocabulary Server (NVS) already used in SeaDataNet, EMODnet, OBIS) and the ontologies from the OBO Foundry and Library (used across multiple domains and systems) will be used to qualify keywords with more machine actionable and LOD/FAIR aligned properties, enhancing the AI-readiness of our digital assets.

No study within BIOcean5D has so far declared a need to generate study-specific ontologies and/or vocabularies.

[20] e.g. https://www.ebi.ac.uk/ena/browser/view/ERC000012

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

3.4. Increase data reuse

BIOcean5D will not only produce new data, but also produce knowledge based on the reuse, treatment and interpretation of historic data sets from decadal to centennial collections.
Historic data being reused includes sets from:

- past TARA expedition data from TARA Ocean, TARA Pacific, TARA Microbiome,
- publicly available sequence data sets and associated metadata from EBI and ENA as well as Sequence Read Archive (SRA) from NCBI Genbank,
- HYSPLIT from NOAA
- the JUVENA/BIOMAN surveys available on the AZTI platform eBegi platform: https://ebegi.azti.es/?lang=en.
- the ROME eDNA network (Ifremer)
- the AtlantECO (10.5281/zenodo.7944433),
- the MARS 3D oceanographic model data (http://doi.org/10.12770/3edee80f-5a3e-42f4-9427-9684073c87f)

Data sets from collections held at European marine stations for long term ecological research observations like those at Villefranche-sur-Mer and the Bay of Biscay, and such as samples and data from deep-sea ecosystem at Ifremer will also be provided by the partners. Further data sets to be provided include the AtlantECO-BASE data set on traditional microbiomes observations (https://zenodo.org/doi/10.5281/zenodo.7944432), as well as regridded environmental climatologies and associated data products for species distribution modelling.

More will be added as projects that run concurrently with BIOcean5D develop and while data sources will be chosen by each WP, task by task, other sources considered at this stage include:

- Open access databases (e.g. OBIS, GBIF, MGnify, EMODnet, ENA, SeaDataNet, IODE and GOOS affiliated resources, PANGAEA);
- Partners' institutional (meta)data that are not necessarily generated in the context of BIOcean5D, but are explicitly listed by the partners as part of the service offer (including citizen science data and long-term international biodiversity data collection programmes).

The provenance of existing data that are used for BiOcean5D analyses and modelling will follow the same standards as much as possible, based on existing metadata from the source data archives and the available literature. The provenance of new samples collected during BIOcean5D will be curated following the minimum information standards and formats (MixS) of the Genomics Standards Consortium (GSC). Those metadata will be archived at EMBL-EBI using the BioSamples archive.

Collectively, all BIOcean5D researchers will contribute to an inventory of reused and generated data to support documentation, discovery, and integration/synthesis throughout the project.

As all data will be made publicly available, including rich metadata, (meta)data will remain usable by third parties even after the end of the project. Where proprietary software has to

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

be used as part of the data acquisition as is common in flow cytometry, data will be converted to an open format.

To increase data reuse after the end of the project, data will be qualified by version numbers or hashsums, so that changes during peer-review of associated papers can be seamlessly traced, and in agreement with the framework laid out by the GA, by default, BIOcean5D generated data will be open with no restriction for sharing, however, exceptions may arise for sensitive data such as endangered or commercially valuable species, and scientific publication.

Additional documentation will be provided in the form of published scripts, codebooks, R markdown files and readme files which will be deposited on GitHub and GitLab. This documentation will include the quality assurance processes for the different data types. Due to the width of disciplines covered by BIOcean5D, studies within the consortium will differ in the details, please consult Annex 1 regarding the quality assurance processes of specific data types.

Partners will announce to the consortium their intent to publish data and other digital assets, and allow a period of 45 working days within which the consortium can raise concerns following the project's data sharing agreement as specified in the grant agreement. The project encourages all scientists to make their data and papers publicly available using FAIR principles upon submission of a preprint of their work in a trusted preprint repository such as biorXiv (https:// https://www.biorxiv.org/) or authorea (https://authorea.com), with each data submission characterised by a unique identifier. In cases where the above-mentioned procedure is not possible, project data will be made publicly available in open source and using a CC-BY (or more permissive) licence as the project's default, as soon as the associated paper is published, or at the latest two years after the end of the project.

4. Other research outputs

The following other research outputs are planned:

- open science skills training
- public engagement
- technical presentations
- policy briefings

In executing the project BIOcean5D will also collect a large number of water, sediment, and genetic samples. Whilst most samples will be fully utilised by the proposed experiments, reference samples will also be collected and stored within the EMBRC infrastructure and/or the EMBL biobank.

All other research outputs will be made publicly available and according to FAIR standards as far as possible.

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

5. Allocation of resources

The WP3 ("Data to knowledge, a digital foundation for holistic marine biodiversity assessment") has the following partners with their respective Person Month (PM) contribution: : AWI (35PM), AZTI (2PM), BIOBYTE (8PM), DTU (18PM), EMBL (74PM), ETHZ (106PM), NOC (3PM), NORCE (15PM), SU (26PM), SZN (12PM), VLIZ (13PM). This makes 312.00 PM total, with WP3 lead beneficiary being ETHZ.

The financial constraints do not allow for the hire of a central data manager for the project, therefore data-generating partners in WPs will standardise their (derived) data and metadata and contribute their data. EMBL/BIOBYTE will host the (meta)data, as well as develop the data hub structure and BIOBYTE will develop the corresponding user interface. ETHZ, SU, DTU will provide expertise on specific data types. AWI, EBI and VLIZ will provide expertise on meta-data structures and data exchange with existing archives.
More details on the contribution and responsibilities of each partner to WP3 can be seen in Table 5.

Table 5: Summary of contribution of partners to WP3 Tasks and Deliverables as described in the GA - Description of Action.

AWI	• T3.1: assessment of the state of biodiversity data resources and development of a roadmap towards a sustainable digital ecosystem • T3.2: provide expertise on meta-data structures and data exchange with existing archives. • T3.3 & T3.4: contribute knowledge on specific data types and experiences from other integrative projects, as well as on taxonomic standardisation and (joint) species distribution modelling, and to collaborate with other WPs on layers of interest (WP4-5) • D3.3: Biodiversity data assessment and roadmap towards international standardisation and operationalization
AZTI	• T3.3 & T3.4: contribute knowledge on specific data types and experiences from other integrative projects, as well as on taxonomic standardisation and (joint) species distribution modelling, and to collaborate with other WPs on layers of interest (WP4-5)
BIOBYTE	• T3.2: Establish a European data hub that collects and brokers existing and novel project-relevant marine biodiversity (meta)data. Codevelop the data hub structure with EMBL. Develop the respective user interface of the data hub. • T3.3: contribute knowledge on specific data types and experiences from other integrative projects, as well as on taxonomic standardisation
DTU	• T3.2: Provide expertise on specific data types • T3.3 & T3.4: contribute knowledge on specific data types and experiences from other integrative projects, as well as on taxonomic standardisation and (joint) species distribution modelling, and to collaborate with other WPs on layers of interest (WP4-5)
EMBL	• T3.2: provide expertise on meta-data structures and data exchange with existing archives (EMBL_EBI) • T3.2: Establish a European data hub that collects and brokers existing and novel project-relevant marine biodiversity (meta)data. Host the metadata, and some of the data. Codevelop the data hub structure with BIOBYTE • T3.3 & T3.4: contribute knowledge on specific data types and experiences from

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

	other integrative projects, as well as on taxonomic standardisation and (joint) species distribution modelling, and to collaborate with other WPs on layers of interest (WP4-5) ● T3.4 Integrative analysis of global biodiversity patterns across main axes of variability (taxa, spatio-temporal scales, methodology, diversity metrics) with ETHZ ● D3.2: Data Hub ● D3.1 and D3.5: Initial and final DMP
ETHZ	● T3.2: Provide expertise on specific data types ● T3.3: Use data science methods to create an integrated marine diversity data set across data types and methods, and to map the resulting diversity patterns in space and time ● T3.3 & T3.4: contribute knowledge on specific data types and experiences from other integrative projects, as well as on taxonomic standardisation and (joint) species distribution modelling, and to collaborate with other WPs on layers of interest (WP4-5) ● T3.4 Integrative analysis of global biodiversity patterns across main axes of variability (taxa, spatio-temporal scales, methodology, diversity metrics) with EMBL ● D3.4: Integrated marine biodiversity data sets and layers of added values
NOC	● T3.3: assist in the definition of community standards for (joint) species distribution modelling (SDM) ● T3.3 & T3.4: contribute knowledge on specific data types and experiences from other integrative projects, as well as on taxonomic standardisation and (joint) species distribution modelling, and to collaborate with other WPs on layers of interest (WP4-5)
NORCE	● T3.3 & T3.4: contribute knowledge on specific data types and experiences from other integrative projects, as well as on taxonomic standardisation and (joint) species distribution modelling, and to collaborate with other WPs on layers of interest (WP4-5) ● T3.4: perform diversity assessment from surface to depth
SU	● T3.2: Provide expertise on specific data types ● T3.3 & T3.4: contribute knowledge on specific data types and experiences from other integrative projects, as well as on taxonomic standardisation and (joint) species distribution modelling, and to collaborate with other WPs on layers of interest (WP4-5)
SZN	● T3.3 & T3.4: contribute knowledge on specific data types and experiences from other integrative projects, as well as on taxonomic standardisation and (joint) species distribution modelling, and to collaborate with other WPs on layers of interest (WP4-5) ● T3.4: map metabolic potential of communities.
VLIZ	● T3.2: provide expertise on meta-data structures and data exchange with existing archives. ● T3.3 & T3.4: contribute knowledge on specific data types and experiences from other integrative projects, as well as on taxonomic standardisation and (joint) species distribution modelling, and to collaborate with other WPs on layers of interest (WP4-5)

Both this initial (D3.1) as well as the final (D3.5) version of the DMP, describe all scientific data outputs by each work package as well as the data requirements to perform the required analysis and its sources. The provision of this information lies with the task leads of each work package and will be synthesised by EMBL and sent for revision to all partners before submission.

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

The costs for making BIOcean5D data available open access are expected to be limited to personnel costs. If extra costs occurs within the project lifespan, costs related to open access of research data in Horizon Europe are eligible under the conditions defined in the BIOcean5D GA Article 6 – Eligible and Ineligible Costs, such as Article 6.2.C.3 – Other goods works and services, but also other articles relevant for the cost category chosen. These include the costs of data deposit, long-term storage and cost in time and effort needed to prepare the data for sharing and preservation. Costs cannot be claimed retrospectively. Project partners will be responsible for including any relevant costs in their financial statements.

Publication fees are only eligible when publishing in full open access publishing venues (venues in which the entire scholarly content is openly accessible to all) and not in hybrid venues.

Given the budgetary constraints, BIOcean5D cannot sustain a dedicated BIOcean5D data manager, therefore the consortium as a whole will prioritise securing data flows to ODIS process to maximise delivery to key global processes and partner projects in Europe and beyond (WP3's priority), as well as flows to well-established long-term archives.

6. Data security

6.1. Hardware and network infrastructure

Overall, the primary responsibility to take necessary measures to ensure data security lies with their generators. Before publication, raw data is planned to be stored primarily on the local servers of the respective partner/data generator with institute specific plans for access management, backup and recovery. Additionally, some raw data will be stored in repositories such as SRA, which allow the storing and access of data pre-publication. Furthermore, the long-term storage of data and output in long-term repositories (which have their own back-up mechanisms and redundancies) adds another layer of security with regards to data loss. Regarding data in the BIOcean5D data hub, EMBL runs a big scale-out filesystem (powered by 11 filers) that is backuped weekly into a tape-based backup pipeline running by EMBL's central IT service department (ITS).
BIOcean5D aims to follow the guidelines for physical data security provided by the UK Data Archive, which recommends the regulation of access to areas housing data, computers, or media; to record the extraction and entry of media or hardcopy materials in storage facilities and to transport sensitive data only under exceptional circumstances.

Securing computer systems involves utilising password protection and firewall installations. If needed, confidentiality should be maintained by including the enforcement of non-disclosure agreements for managers or users handling sensitive data. It is crucial not to transmit personal or confidential information via email or other file transfer methods without prior encryption. Data disposal should be executed in a consistent manner when required. It's important to note that file-sharing services like Google Docs or Dropbox may not provide robust security measures.

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

6.2. Data access

To facilitate joint data sharing and processing within the consortium before publication, the BIOcean5D datahub is envisioned as a common hub to link the processed data and metadata together and make it available to the partners of BIOcean5D. This will ensure that data can be shared in a controlled environment and foster cooperation and collaboration within the consortium. For the data stored in the data hub, EMBL uses a big scale-out filesystem, which can store terabytes of data directly. Keeping data there also facilitates the daily easy access by users. For data security, the NFSv4-ACLs policy can be applied to ensure/manage the users' access. Enhancing data security encompasses measures such as implementing password protection and regulated access to data files, with options like no access, read-only, read-and-write, or administrator-only permissions.
For publication, both raw and processed data will be stored in trusted repositories for long term preservation (see 3.1 for a list of repositories).

The GDPR data, which is gathered as part of the Human Survey task, has additional safety measures to account for the different types of data compared to the environmental data. Raw data will be stored on secure, internal drives. Quantitative data will be collected in a way that ensures anonymity. Qualitative data will be anonymized by removing any personal/identifiable references from the transcripts. All participants in the Deliberative Monetary Valuation (DMV) workshops will be informed about our data collection and analysis. Informed consent will be a precondition of participation in the data collection. As only anonymized data will be hosted in the BIOcean5D data hub, no additional safety measurements and restrictions for accessibility have to be applied.

7. Ethics

BIOcean5D is committed to carry out its work with the highest ethical standards of the EU , national and international bodies as well as upholding the values of the EU in all aspects of the consortium's work. WP9 is solely focused on compliance with the ethics requirements as laid out in grant agreement article 14 and should issues arise it will be their responsibility to deal with them in an appropriate manner to ensure the ethical integrity of BIOcean5D. An ethics advisor has been appointed as of June 2023 (PlusEthics) and ethics reports will be delivered in M18, M36 and M48 of the BIOcean5D project.

The following ethical or legal issues have been identified that could impact data sharing

1. Human participation (workshops, qualitative interviews, questionnaires),
2. Collection of personal data (human participation as above, summer schools, hackathons),
3. Research on animals which may involve the collection of endangered species,
4. Participation of non-EU countries (Norway, Switzerland, UK), which may involve the transfer of biological material and / or personal data from and / or to the EU,
5. Environmental concerns, as research activities within marine protected areas including areas in Switzerland and UK, as well as artificial intelligence that combines species genetic data, ecosystem network data and machine learning (ML) algorithms (however, AI/ML techniques do not raise ethical concerns related to human rights and values).

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

The personal data that will be gathered as part of the surveys within the socio-economics WP (WP6) will be anonymised. These human surveys will not gather sensitive personal information and will comply with GDPR regulations for non-sensitive personal data. All participants in the DMV workshops will be informed about our data collection and analysis. Informed consent will be a precondition of participation in the data collection.

As of writing of this initial version of the DMP, a deep screening of BIOcean5D tasks is being performed to identify potential Ethics bottlenecks.

8. Other issues

No other issues are known at this point.

Annexes

Annex 1 - Overview of data types collected in BIOcean5D

(see next page)

This table lists the sample types and data types to be used in the project with mainly information on generators, data formats (proprietary are highlighted in yellow), analyses, expected sizes, quality control steps, long term-storage platforms and access model.

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Project area	Project (core vs plug in)	WP/Task/ Subtask	Task Lead	Sample type	Sample fraction	Analysis type	Specific analysis (only 1 per row)	Method	# samples per normal site	# samples per super site	Seq tech.	Nb of reads/sample (million)	Primer ?	Raw data type produced	(Expected) Generated data size/sample	Raw data format	Raw data storage (internal)	QC by	QC processes	Curated data type produced	Curated data size produced	Curated data storage	Format of metadata	Re-use of historic/ existing data?	Re-use: Which data?	Open access - which depository?	Persistent identifier	Sharing data via BSD data hub	Access for all BSD members ok?
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	0.22-3 µm	Seq	MetaG	Illumina Novaseq	2	2		100		Short Read Seq	30 Gb	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	0.22-3 µm	Seq	MetaT (Ribo depletion for prokaryotic organisms)	Illumina Novaseq	2	2		100		Short Read Seq	30 Gb	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	0.22-3 µm	Seq	MetaB (SSU V4V5 Prok/Euk)	Illumina Novaseq	2	2		0.25		Short Read Seq	40 Mb	fastq	local server	Nicolas Henry	Amplicon sequencing analysis workflow (dada2)	fastq and csv	~ 50 Mb for all samples	local server	csv	TRUE	Tara expeditions	ENA (raw sequences) / Zenodo (asv table)	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	0.22-3 µm	Seq	MetaB (18S V9)	Illumina Novaseq	2	2		0.25		Short Read Seq	40 Mb	fastq	local server	Nicolas Henry	Amplicon sequencing analysis workflow (dada2)	fastq and csv	~ 50 Mb for all samples	local server	csv	TRUE	Tara expeditions	ENA (raw sequences) / Zenodo (asv table)	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	0.22-3 µm	Seq	MetaG	Oxford Nanopore	2	2		TBD		Long Read Seq	TBD	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	0.22-3 µm	Seq	HiC Chromatin Conformation	Illumina Novaseq	2	2		TBD		Short Read Seq	TBD	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	3-20 µm	Seq	MetaG	Illumina Novaseq	2	2		100		Short Read Seq	30 Gb	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	3-20 µm	Seq	MetaT (PolyA selection for Eukaryotic organisms)	Illumina Novaseq	2	2		100		Short Read Seq	30 Gb	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	3-20 µm	Seq	MetaB (SSU V4V5 Prok/Euk)	Illumina Novaseq	2	2		0.25		Short Read Seq	40 Mb	fastq	local server	Nicolas Henry	Amplicon sequencing analysis workflow (dada2)	fastq and csv	~ 50 Mb for all samples	local server	csv	TRUE	Tara expeditions	ENA (raw sequences) / Zenodo (asv table)	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	3-20 µm	Seq	MetaB (18S V9)	Illumina Novaseq	2	2		0.25		Short Read Seq	40 Mb	fastq	local server	Nicolas Henry	Amplicon sequencing analysis workflow (dada2)	fastq and csv	~ 50 Mb for all samples	local server	csv	TRUE	Tara expeditions	ENA (raw sequences) / Zenodo (asv table)	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	3-20 µm	Seq	MetaG	Oxford Nanopore	2	2		TBD		Long Read Seq	TBD	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	3-20 µm	Seq	HiC Chromatin Conformation	Illumina Novaseq	2	2		TBD		Short Read Seq	TBD	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	20-200 µm	Seq	MetaG	Illumina Novaseq	2	2		100		Short Read Seq	30 Gb	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	20-200 µm	Seq	MetaT (PolyA selection for Eukaryotic organisms)	Illumina Novaseq	2	2		100		Short Read Seq	30 Gb	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	20-200 µm	Seq	MetaB (SSU V4V5 Prok/Euk)	Illumina Novaseq	2	2		0.25		Short Read Seq	40 Mb	fastq	local server	Nicolas Henry	Amplicon sequencing analysis workflow (dada2)	fastq and csv	~ 50 Mb for all samples	local server	csv	TRUE	Tara expeditions	ENA (raw sequences) / Zenodo (asv table)	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	20-200 µm	Seq	MetaB (18S V9)	Illumina Novaseq	2	2		0.25		Short Read Seq	40 Mb	fastq	local server	Nicolas Henry	Amplicon sequencing analysis workflow (dada2)	fastq and csv	~ 50 Mb for all samples	local server	csv	TRUE	Tara expeditions	ENA (raw sequences) / Zenodo (asv table)	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	20-200 µm	Seq	MetaG	Oxford Nanopore	2	2		TBD		Long Read Seq	TBD	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Tara SW Plankton	Plug-in	WP1	de Vargas /Poulain?	Shallow waters	20-200 µm	Seq	HiC Chromatin Conformation	Illumina Novaseq	2	2		TBD		Short Read Seq	TBD	fastq	local server	TBD	TBD	fastq	TBD	local server	csv	FALSE		ENA	Biosamples	FALSE	FALSE
Shallow waters	Core	WP1.1	Berthelot	Shallow waters	-	Analytics	Nutrient analyses (DOC, DON, nitrate, nitrite, silicate, ammonium, phosphate)	Segmented Flow Anlayzer	3	3	-	-		-	small	.csv	local server	Hugo Berthelot		.csv	few MB	local server	.csv	FALSE				FALSE	FALSE
	Core	WP1.1	Berthelot	Water column	-	Analytics	Nutrient analyses (DOC, DON, nitrate, nitrite, silicate, ammonium, phosphate)	Segmented Flow Anlayzer	3	3	-	-		-	small	.csv	local server	Hugo Berthelot		.csv	few MB	local server	.csv	FALSE			Biosamples	FALSE	FALSE
	Core	WP1.1	Berthelot	Water column	1µm-200µm+	Analytics	cytosense (semi continuous flow cytometry analyser)	flow cytometry	-	-	-	-		-	1 TB	.csv	local server	Hugo Berthelot		.csv	1 TB	local server	.csv	FALSE			Biosamples	FALSE	FALSE
Shallow waters	Plug-in	WP1.1	Vincent	Shallow waters	-	Analytics	CTD measurements	CTD	1	1	-	-		-	small	.nsk	local server			.nsk: .xlsx	few MB	local server	txt, csv	FALSE		Biosamples	Biosamples	FALSE	FALSE
Shallow waters	Plug-in	WP1.1	Vincent	Shallow waters	20-200 µm	Imaging	Confocal on single cells	CM	2	10	-	-		tif	TBD	tif	local server			tif	TBD	local server	txt, csv	FALSE	no	EcoTaxa?	Biosamples	FALSE	FALSE
Shallow waters	Plug-in	WP1.1	Vincent	Shallow waters	20-200 µm	Imaging	Imaging flow cytometry on single cells	IFC	2	10	-	-		see q	TBD	txt, dcimg, lmd and lsm4	local server				TBD	local server	txt, csv	FALSE	no	EcoTaxa?	Biosamples	FALSE	FALSE
Shallow waters	Plug-in	WP1.1	Vincent	Shallow waters	5-20 µm	Imaging	Confocal on single cells	CM	2	10	-	-		-	TBD	tif	local server				TBD	local server	txt, csv	FALSE	no	EcoTaxa?	Biosamples	FALSE	FALSE
Shallow waters	Plug-in	WP1.1	Vincent	Shallow waters	5-20 µm	Imaging	Imaging flow cytometry on single cells	IFC	2	10	-	-		-	TBD	txt, dcimg, lmd and lsm4	local server				TBD	local server	txt, csv	FALSE	no	EcoTaxa?	Biosamples	FALSE	FALSE
	Plug-in	WP1.1	Vincent	Water column	20-200 µm	Imaging	Confocal on single cells	CM	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	FALSE			Biosamples	FALSE	FALSE
	Plug-in	WP1.1	Vincent	Water column	20-200 µm	Imaging	Imaging flow cytometry on single cells	IFC	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	FALSE			Biosamples	FALSE	FALSE
	Plug-in	WP1.1	Vincent	Water column	5-20 µm	Imaging	Confocal on single cells	CM	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	FALSE			Biosamples	FALSE	FALSE
	Plug-in	WP1.1	Vincent	Water column	5-20 µm	Imaging	Imaging flow cytometry on single cells	IFC	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	FALSE			Biosamples	FALSE	FALSE
Shallow waters	Plug-in	WP1.1	Vincent	Shallow waters	5-20 µm	Seq	MetaB (16V4V5 Prok/Euk) on single cells	MetaB	2	10		Illumina	TBD		fastq	TBD	fasta	local server			fasta	TBD	local server	txt, csv	FALSE	no	ENA	Biosamples	FALSE	FALSE
	Plug-in	WP1.1	Vincent	Water column	20-200 µm	Seq	MetaB (16V4V5 Prok/Euk) on single cells	MetaB	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	FALSE			Biosamples	FALSE	FALSE
	Plug-in	WP1.1	Vincent	Water column	5-20 µm	Seq	MetaB (16V4V5 Prok/Euk) on single cells	MetaB	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	FALSE			Biosamples	FALSE	FALSE
Shallow waters	Plug-in	WP1.1	Raina	Shallow waters	-	Seq	MetaG	-	TBD	30		Illumina NovaSeq	100 M		Short Read Seq	30 Gb	fastq	local server		quality and adapter trimming	fasta	TBD	local server	csv	TRUE	ENA data	ENA	Biosamples	TRUE	TRUE
Intertidal annelids	Plug-in	WP1.1.3	Bleidorn	Annelids	-	cox1-barcode	Sanger seq		transect / TBD	transect / TBD	Sanger	1		Sanger reads	650 bp	fasta	local server	TBD	qc and trimming	fasta	TBD	local server	csv	FALSE		BOLD	Biosamples	TRUE	TRUE
Intertidal annelids	Plug-in	WP1.1.3	Bleidorn	Annelids	-	Seq	Illumina genome skimming		transect / TBD	transect / TBD	Illumina NovaSeq	TBD		Short Read Seq	8 Gbp	fastq	local server	TBD	quality and adapter trimming	fasta	TBD	local server	csv	TRUE	SRA	ENA	Biosamples	TRUE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	Biogeochemics l	BVOC profile ON TARA		0	TBD	-				TBD	.csv	local server	James O'Brien	Identification via NIST08 library in NIST MS Search v 2.2f	.csv	TBD	local server	.csv	FALSE			Biosamples	FALSE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	physio-chem. Prof	Particle counts (size: 0.25 um - 35 um) on TARA		1	1	-				0.5 Mb	.csv	local server	Michel Flores		.csv	TBD	local server	.csv	FALSE			Biosamples	FALSE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	physio-chem. Prof	Meterological data (temp., relative humidity, wind speed, wind direction, pressure, precipitation) ON TARA		1	1	-				0.5 Mb	.csv	local server	Michel Flores		.csv	TBD	local server	.csv	FALSE			Biosamples	FALSE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	physio-chem. Prof	Particle counts (size: 0.25 um - 35 um)		1	1	-				0.5 Mb	.csv	local server	Michel Flores		.csv	TBD	local server	.csv	FALSE			Biosamples	FALSE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	physio-chem. Prof	Meterological data (temp., relative humidity, wind speed, wind direction, pressure, precipitation)		1	1	-				0.5 Mb	.csv	local server	Michel Flores		.csv	TBD	local server	.csv	FALSE			Biosamples	FALSE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	physio-chem. Prof	Morphology (SEM, elemental analysis)		2	2	-				0.5 Mb	.xlsx	local server	Michel Flores		.csv	TBD	local server	.csv	FALSE			Biosamples	FALSE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	Seq	Meta B (16V4V5 Prok/Euk, total > 0.45um)		2	2	Illumina NovaSeq	0.5 M		Short Read Seq	300 Mb	fastq	local server	Hans-Joachim Ruscheweyh	quality and adapter trimming	fastq	TBD	local server	.csv	FALSE		Biosamples	Biosamples	TRUE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	Seq	Size segregated Meta B (16V4V5 Prok/Euk, size: 0.26, 0.61, 1.2 and 3 µm)		2	2	Illumina NovaSeq	0.5 M		Short Read Seq	300 Mb	fastq	local server	Hans-Joachim Ruscheweyh	quality and adapter trimming	fastq	TBD	local server	.csv	FALSE		Biosamples	Biosamples	TRUE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	Sea surface microlayer	-	Seq	Meta B (16V4V5 Prok/Euk, Surface microlayer, 0.2um and 3um)		0	30	Illumina NovaSeq	0.5 M		Short Read Seq	300 Mb	fastq	local server	Hans-Joachim Ruscheweyh	quality and adapter trimming	fastq	TBD	local server	.csv	FALSE		Biosamples	Biosamples	TRUE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	Seq	MetaB (16V4V5 Prok/Euk) (12h sample - day night)		2	2	Illumina NovaSeq	0.5 M		Short Read Seq	300 Mb	fastq	local server	Hans-Joachim Ruscheweyh	quality and adapter trimming	fastq	TBD	local server	.csv	FALSE		Biosamples	Biosamples	TRUE	TRUE
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	Seq	MetaG		TBD	TBD	Illumina NovaSeq	100 M		Short Read Seq	30 Gb	fastq	local server	Hans-Joachim Ruscheweyh	quality and adapter trimming	fastq	TBD	local server	.csv	FALSE		Biosamples	Biosamples	TRUE	TRUE

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Project area	Project (core vs plug in)	WP/Task/ Subtask	Task Lead	Sample type	Sample fraction	Analysis type	Specific analysis (only 1 per row)	Method	# samples per normal site	# samples per super site	Seq tech.	Nb of reads/sample (million)	Primer ?	Raw data type produced	(Expected) Generated data size/sample	Raw data format	Raw data storage (Internal)	QC by	QC processes	Curated data type produced	Curated data size produced	Curated data storage	Format of metadata	Re-use of historic/ existing data?	Re-use: Which data?	Open access - which depository?	Persistent identifier	Sharing data via BSD data hub	Access for all BSD members ok?
Aerosols	Plug-in	WP1.1.4	Sunagawa /Flores	Aerosols	-	Seq	MetaB (16V4V5 Prok/Euk) (4h sample)		1	30	Illumina Novaseq	0.5 M		Short Read Seq	300 Mb	fastq	local server	Hans-Joachim Ruscheweyh	quality and adapter trimming	fastq	TBD	local server	.csv	FALSE		Biosamples	Biosamples	TRUE	TRUE
Bioacoustics		WP1.1.5	Di Iorio	passive acoustic data		Acoustics	soundscape, spectra & long-term spectrograms		2-3	20				acoustic	several TB	.wav			visualization	.jpg	several TB	internal	csv	FALSE	-	- data too large	Biosamples	not relevant	not relevant
Bioacoustics		WP1.1.5	Di Iorio	passive acoustic data		Acoustics	sound type diversity and classification		2-3	20				acoustic	several TB	.wav			visualization	.jpg & .csv .wav clips	several TB	internal	csv	FALSE		to be defined	Biosamples	FALSE	FALSE
Bioacoustics		WP1.1.5	Di Iorio	passive acoustic data		Acoustics	acoustic mass phaeoconvexen, invertebrate & fish mass choruses		2-3	20				acoustic	several TB	.wav			visualization	.jpg & .csv	several TB	internal	csv	FALSE	-	- data too large	Biosamples	FALSE	FALSE
Bioacoustics		WP1.1.5	Di Iorio	passive acoustic data		Acoustics	noise, anthropogenic noise quantification		2-3	20				acoustic	several TB	.wav			visualization	.jpg & .csv	several TB	internal	csv	FALSE	-		Biosamples	FALSE	FALSE
Paleocore	Plug-in	WP1.2	Siano	Sediments	-	Biogeochemical	Organic Pollution	MS	6	6													FALSE			Biosamples	FALSE	FALSE
Paleocore	Plug-in	WP1.2	Siano	Sediments	-	Biogeochemical	Inorganic Pollution	MS	1 x100 core layers	1 x100 core layers													FALSE			Biosamples	FALSE	FALSE
Paleocore	Plug-in	WP1.2	Siano	Sediments	-	Biogeochemical	Inorganic Pollution	MS	1 x100 core layers	1 x100 core layers													FALSE			Biosamples	FALSE	FALSE
Paleocore	Plug-in	WP1.2	Siano	Sediments	-	Chem Prof	Organic Carbon on paleocores	Carbon analyser	1 x100 core layers	1 x100 core layers													FALSE			Biosamples	FALSE	FALSE
Paleocore	Plug-in	WP1.2	Siano	Sediments	-	Chem Prof	Granulometry on paleocores	Laser granulometr y	1 x100 core layers	1 x100 core layers													FALSE			Biosamples	FALSE	FALSE
Paleocore	Plug-in	WP1.2	Siano	Sediments	-	Imaging	Meiofauna Microscopy on paleocores	Microscopy	1 x100 core layers	1 x100 core layers													FALSE			Biosamples	FALSE	FALSE
Sediments	Plug-in	WP1.2.2	Briant	Sediments		Organic Pollutants	Chemical profiling																FALSE				FALSE	FALSE
Sediments	Plug-in	WP1.2.2	Castro	Sediments		Organic Pollutants	Chemical profiling	LC/MS-MS	TBD	TBD							local server	Charles Pollonomy wolf		.csv	very small	local server	txt, csv	FALSE	no	TBD	Biosamples	TRUE	TRUE
Paleocores		WP1.2	Siano	acient sediment	-	Seq	MetaB ancient DNA	Illumina	TBD	TBD	Illumina Novaseq		18s V4	Short Read	TBD	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fast.q	TBD	local server	CSV	FALSE				TRUE	TRUE
Paleocores		WP1.2	Siano	acient sediment	-	Seq	MetaB ancient DNA	Illumina	TBD	TBD	Illumina Novaseq		COI Metazoa	Short Read	TBD	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fast.q	TBD	local server	CSV	FALSE				TRUE	TRUE
Paleocore	Plug-in	WP1.2	Siano	Sediments	-	Seq	MetaB (18V4 Euk) with ancient DNA	Illumina Sequencing (MetaB)	1 x100 core layers	1 x100 core layers	Illumina Sequencing	TBD		fastq	TBD	fastq	local server	Jan Pawlowski		fasta		local server	csv	FALSE		TBD	Biosamples	FALSE	FALSE
Paleocore	Plug-in	WP1.2	Siano	Sediments	-	Seq	MetaG with ancient DNA	Shotgun sequencing (MetaG)	1 x100 core layers	1 x100 core layers	Shotgun sequencing	TBD		fastq	TBD	fastq	local server	Antonio Fernandez Guerra		fasta		local server	csv	FALSE		TBD	Biosamples	FALSE	FALSE
Paleocores		WP1.2	Siano/Guerra	acient sediment	-	Seq	MetaG ancient DNA	Illumina	TBD	TBD	Shot Gun		-	Short Read	TBD	fastq	local server	internal	quality and adapter trimming, demultiplexing	fast.q	TBD	local server	CSV	FALSE				TRUE	TRUE
VALCOT	Plug in	WP1.3	Doxaran	Water column	>0.7µm	HPLC	identification & quantification of phytoplankton pigments	HPLC	3	7						.xlsx	local server	David		.xlsx	.xlsx		csv, txt	FALSE				TRUE	TRUE
VALCOT	Plug in	WP1.3	Doxaran, Dimier	Shallow waters	>0.7µm	HPLC	identification & quantification of phytoplankton pigments (HP protocol)	HPLC	3	7						.xlsx	local server	David		.xlsx	.xlsx		csv, txt	FALSE				TRUE	TRUE
Fish/marine mammals/birds		WP 1.3	Soccodato	fish/marine mammal/bird eDNA	0,45um	Seq	metaB, eDNA	Illumina	NA (not TREC)	NA (not TREC)	NovaSeq	-	takeo	Short Read Seq	GB	fastq	SRA	internal	eg. https://github.com/rodriguez-ezpeleta/deepFish_eDNA	.csv	TBD	Zenodo	txt	yes	observations : trawls, acoustics	ebegil (https://aztidata.es/ebegil/)	doi	TBD	TBD
Dissolved gases	Plug in	WP1.4	Cardini	Water column	-	Biogeochemical	O2/Ar	MIMS	3	3	-	-	-	csv	TBD	.csv	local server	Ulisse Cardini	Sample quality, signal quality and metadata quality	.csv	TBD	local server	.csv	TRUE	Tara MMA	to be defined	Biosamples	TRUE	TRUE
Dissolved gases	Plug in	WP1.4	Cardini	Water column	-	Biogeochemical	N2/Ar	MIMS	3	3	-	-	-	csv	TBD	.csv	local server	Ulisse Cardini	Sample quality, signal quality and metadata quality	.csv	TBD	local server	.csv	TRUE	Tara MMA	to be defined	Biosamples	TRUE	TRUE
Dissolved gases	Plug in	WP1.4	Cardini	Water column	-	Biogeochemical	DMS	MIMS	3	3	-	-	-	csv	TBD	.csv	local server	Ulisse Cardini	Sample quality, signal quality and metadata quality	.csv	TBD	local server	.csv	TRUE	Tara MMA	to be defined	Biosamples	TRUE	TRUE
water chemical analyses		WP1.4	Jeanthon /Cardini	water and filters	-	chemical analyses	segmented flow analysis			NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
water chemical analyses		WP1.4	Jeanthon /Cardini	water and filters	-	chemical analyses	spectrofluorimetry			NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
water chemical analyses		WP1.4	Jeanthon /Cardini	water and filters	-	chemical analyses	MIMS			NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
water chemical analyses		WP1.4	Jeanthon /Cardini	water and filters	-	chemical analyses	EA			NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
water chemical analyses		WP1.4	Jeanthon /Cardini	water and filters	-	chemical analyses	TOC			NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
water chemical analyses		WP1.4	Jeanthon /Cardini	water and filters	-	chemical analyses	IRMS			NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
water chemical analyses		WP1.4	Jeanthon /Cardini	water and filters	-	chemical analyses	pollutants			NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
sediment chemical analyses		WP1.4	Jeanthon /Cardini	sediment	-	Chemical analyses	pollutants			NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
water HPLC and flow cytometry		WP1.4	Jeanthon /Cardini	water and filters	-	flow cytometry	flow cytometry	FC	NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
water HPLC and flow cytometry		WP1.4	Jeanthon /Cardini	water and filters	-	HPLC	HPLC	HPLC	NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
plankton community analyses		WP1.4	Jeanthon /Cardini	water and filters	-	Imaging	microscopy			NA	TBD		TBD	TBD	csv, text	TBD	.csv, .txt, fasta	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt, fasta	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
sediment community analyses		WP1.4	Jeanthon /Cardini	sediment	-	Imaging	microscopy			NA	TBD		TBD	TBD	csv, text	TBD	.csv, .txt, fasta	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt, fasta	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
water productivity and n fixation		WP1.4	Jeanthon /Cardini	water and filters	-	mass spectroscopy	IRMS	MS	NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
water productivity and n fixation		WP1.4	Jeanthon /Cardini	water and filters	-	mass spectroscopy	MIMS	MS	NA	TBD	-	-	-	csv, text	TBD	.csv, .txt	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
plankton community analyses		WP1.4	Jeanthon /Cardini	water and filters	-	Seq	MetaB	Illumina	NA	TBD	Illumina Novaseq	TBD	TBD	csv, text	TBD	.csv, .txt, fasta	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt, fasta	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
plankton community analyses		WP1.4	Jeanthon /Cardini	water and filters	-	Seq	MetaG	Illumina	NA	TBD	Illumina Novaseq	TBD	TBD	csv, text	TBD	.csv, .txt, fasta	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt, fasta	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
plankton community analyses		WP1.4	Jeanthon /Cardini	water and filters	-	Seq	MetaT	Illumina	NA	TBD	Illumina Novaseq	TBD	TBD	csv, text	TBD	.csv, .txt, fasta	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt, fasta	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
sediment community analyses		WP1.4	Jeanthon /Cardini	sediment	-	Seq	MetaB	Illumina	NA	TBD	Illumina Novaseq	TBD	TBD	csv, text	TBD	.csv, .txt, fasta	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt, fasta	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
sediment community analyses		WP1.4	Jeanthon /Cardini	sediment	-	Seq	MetaG	Illumina	NA	TBD	Illumina Novaseq	TBD	TBD	csv, text	TBD	.csv, .txt, fasta	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt, fasta	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Project area	Project (core vs plug in)	WP/Task/ Subtask	Task Lead	Sample type	Sample fraction	Analysis type	Specific analysis (only 1 per row)	Method	# samples per normal site	# samples per super site	Seq tech.	Nb of reads/sample (million)	Primer ?	Raw data type produced	(Expected) Generated data size/sample	Raw data format	Raw data storage (internal)	QC by	QC processes	Curated data type produced	Curated data size produced	Curated data storage	Format of metadata	Re-use of historic/ existing data?	Re-use: Which data?	Open access - which depository?	Persistent identifier	Sharing data via BSD data hub	Access for all BSD members ok?
sediment community analyses		WP1.4	Jeanthon /Cardini	sediment	-	Seq	MetaT	Illumina	NA	TBD	Illumina Novaseq	TBD	TBD	csv, text	TBD	.csv, .txt, fasta	local servers	internal	Sample quality, signal quality and metadata quality	.csv, .txt, fasta	TBD	local server	csv	yes	LTER data	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
Plankton		WP 1.5.2	Lee-Karp	Water column	2-100µm	imaging	abundance, species and morphology of plankton (imaged with ifcb)	flowcam	?	?					3000 images/sample	.jpg .tsv	local server	PiqV	QC procedure of the plateforme d'imagerie quantitative from villefranche sur mer	.csv	few mb	local server (+ ecotaxa) + obis + obifs	csv, txt	FALSE		Ecotaxa		TRUE	TRUE
ifcb	Core	WP 1.5.2	Lee-Karp	Water column	2-100µm	imaging	abundance, species and morphology of plankton (imaged with ifcb)	flowcam	?	?					3000 images/sample	.jpg .tsv	local server	PiqV	QC procedure of the plateforme d'imagerie quantitative from villefranche sur mer	.csv	few mb	local server (+ ecotaxa) + obis + obifs	csv, txt	FALSE		Ecotaxa		TRUE	TRUE
Plankton		WP 1.5.2	Lombard	Water column	>200µm	imaging	abundance, species and morphology of plankton (imaged with zooscan)	Zooscan	?	?					3000 images/sample	.jpg .tsv	local server	PiqV	QC procedure of the plateforme d'imagerie quantitative from villefranche sur mer	.csv	few mb	local server (+ ecotaxa) + obis + obifs	csv, txt	FALSE		Ecotaxa		TRUE	TRUE
Plankton		WP 1.5.2	Lombard	Water column	>680µm	imaging	abundance, species and morphology of plankton (imaged with zooscan)	Zooscan	?	?					3000 images/sample	.jpg .tsv	local server	PiqV	QC procedure of the plateforme d'imagerie quantitative from villefranche sur mer	.csv	few mb	local server (+ ecotaxa) + obis + obifs	csv, txt	FALSE		Ecotaxa		TRUE	TRUE
Plankton		WP 1.5.2	Lombard	Water column	20-200µm	imaging	abundance, species and morphology of plankton (imaged with flowcam)	flowcam	?	?					3000 images/sample	.jpg .tsv	local server	PiqV	QC procedure of the plateforme d'imagerie quantitative from villefranche sur mer	.csv	few mb	local server (+ ecotaxa) + obis + obifs	csv, txt	FALSE		Ecotaxa		TRUE	TRUE
zooscan WP2	Core	WP 1.5.2	Lombard	Water column	>200µm	imaging	abundance, species and morphology of plankton (imaged with zooscan)	Zooscan	?	?					3000 images/sample	.jpg .tsv	local server	PiqV	QC procedure of the plateforme d'imagerie quantitative from villefranche sur mer	.csv	few mb	local server (+ ecotaxa) + obis + obifs	csv, txt	FALSE		Ecotaxa		TRUE	TRUE
zooscan regent	Core	WP 1.5.2	Lombard	Water column	>680µm	imaging	abundance, species and morphology of plankton (imaged with zooscan)	Zooscan	?	?					3000 images/sample	.jpg .tsv	local server	PiqV	QC procedure of the plateforme d'imagerie quantitative from villefranche sur mer	.csv	few mb	local server (+ ecotaxa) + obis + obifs	csv, txt	FALSE		Ecotaxa		TRUE	TRUE
flowcam	Core	WP 1.5.2	Lombard	Water column	20-200µm	imaging	abundance, species and morphology of plankton (imaged with flowcam)	flowcam	?	?					3000 images/sample	.jpg .tsv	local server	PiqV	QC procedure of the plateforme d'imagerie quantitative from villefranche sur mer	.csv	few mb	local server (+ ecotaxa) + obis + obifs	csv, txt	FALSE		Ecotaxa		TRUE	TRUE
Shallow waters	Plug-in	WP1.6	Vincent /Peiccerlok	Shallow waters	20-200 µm	Seq	MetaB (16V4V5 Prok/Euk) on sindle cells	MetaB	2	10	Illumina	TBD		fastq	TBD	fasta	local server			fasta	TBD	local server	txt, csv	FALSE	no	ENA	Biosamples	TRUE	TRUE
Land-sea interphase	Plug-in	WP 1.6.3	Pepperkok /Bethelot	Shallow waters		Imaging	Morphology – e-HCFM/SEM/eaM nanci		2	2														FALSE					
Land-sea interphase	Plug-in	WP 1.6.3	Pepperkok /Bethelot	Shallow waters	5-20 µm	Imaging	Morphology – e-HCFM/SEM/eaM nanci		2	2																	FALSE	FALSE
Tara sampling	Plug-in	WP 1.6.3	Pepperkok /Bethelot	Water column	-	Imaging	Morphology – e-HCFM/SEM/eaM nanci		2	2																		
Sediment	Core	WP2.1.1	Siano	Sediments	-	Chem Prof	Porosity on superficial sediment		3	3	-	-	-	-								csv	FALSE			Biosamples	TRUE	TRUE
Estuaries		WP2.1.1	Siano	Water filter (TREC)	all fractions	Seq	Metabarcoding, interaction network analyses in estuary ecosystems	Illumina	cf TREC	Cf. trec	Illumina Novaseq		18s V4	Short Read Seq	TBD	fastq	local server	internal	quality and adapter trimming, demultiplexing	fast.q	TBD	local server	CSV	FALSE				TRUE	TRUE
Estuaries		WP2.1.1	Siano	Water filter (TREC)	all fractions	Seq	Metabarcoding, interaction network analyses in estuary ecosystems	Illumina	cf TREC	Cf. trec	Illumina Novaseq		16s V4-V5	Short Read Seq	TBD	fastq	local server	internal	quality and adapter trimming, demultiplexing	fast.q	TBD	local server	CSV	FALSE				TRUE	TRUE
Estuaries		WP2.1.1	Siano	Water filter (TREC)	all fractions	Seq	Metabarcoding, interaction network analyses in estuary ecosystems	Illumina	cf TREC	Cf. trec	Illumina Novaseq		COI Metazoa	Short Read Seq	TBD	fastq	local server	internal	quality and adapter trimming, demultiplexing	fast.q	TBD	local server	CSV	FALSE				TRUE	TRUE
Estuaries		WP2.1.1	Siano/Nunes	water filter (ROME network)	>20 µm	Seq	Metabarcoding, interaction network analyses in estuary ecosystems	Illumina	600 metaB	NA (not TREC)	Illumina Novaseq		COI Metazoa	Short Read Seq	TBD	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fast.q	TBD	local server	CSV	TRUE				TRUE	TRUE
selected species/algae & inverts		WP 2.1.4	Hambleton	dinoflagellate algae symbiotic in invertebrate anemone	-	Seq	metaB	Illumina	30	30	Illumina Novaseq	35 (sample subset tecmics)	- (metaB in-house)	Short Read Seq	TBD	fastq	local server	internal	quality and adapter trimming	fastq/fasta	TBD	local server	text	no	-	NCBI or similar public	Biosamples	TRUE	TRUE
selected species/algae & inverts		WP 2.1.4	Hambleton	dinoflagellate algae symbiotic in invertebrate anemone	-	Seq	transcriptomics on subset of samples	Illumina	9	9	Illumina Novaseq	35 (sample subset tecmics)	-	Short Read Seq	TBD	fastq	local server	internal	quality and adapter trimming	fastq/fasta	TBD	local server	text	no	-	NCBI or similar public	Biosamples	TRUE	TRUE
anemone	Plug in	WP2.1.4	Hambleton	dinoflagellate algae symbiotic in invertebrate anemone	-	Seq	metaB	Illumina	30 metaB	31 metaB	Illumina Novaseq	35 (sample subset tecmics)		Short Read Seq	TBD	fastq	local server	internal	quality and adapter trimming	fastq/fasta	TBD	local server	txt	FALSE	no		Biosamples	TRUE	TRUE
anemone	Plug in	WP2.1.4	Hambleton	dinoflagellate algae symbiotic in invertebrate anemone	-	Seq	transcriptomics on subset of samples	Illumina	30 metaB, 9 of which +tecmic	31 metaB, 9 of which +tecmic	Illumina Novaseq	35 (sample subset tecmics)		Short Read Seq	TBD	fastq	local server	internal	quality and adapter trimming	fastq/fasta	TBD	local server	txt	FALSE	no		Biosamples	TRUE	TRUE
Seagrasses - Selected species		WP 2.2	Arnaud-Haond	Sea grasses	-	Seq	WGS	WGS	25	25	Illumina Novaseq	TBD		Short Read Seq	TBD	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fasta	TBD	local server	csv	FALSE	no	ENA/EBI-Dataref (Ifremer)	Biosamples	TRUE	TRUE
ADNe-TARA		WP 2.2	Arnaud-Haond	Water column	>0.45µm	Seq	Meta B (16V4V5 Prok/Euk)	MetaB	1	1	Illumina Novaseq	0.25		Seq	250 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBD	local server	csv	FALSE		TBD	Biosamples	TRUE	TRUE
ADNe-TARA		WP 2.2	Arnaud-Haond	Water column	>0.45µm	Seq	Meta B (COI Metazoa)	MetaB	1	1	Illumina Novaseq	0.5		Seq	350 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBD	local server	csv	FALSE		TBD	Biosamples	TRUE	TRUE
ADNe-TARA		WP 2.2	Arnaud-Haond	Water column	>0.45µm	Seq	Meta B (18S V7V2 Metazoa)	MetaB	1	1	Illumina Novaseq	0.5		Seq	500 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBD	local server	csv	FALSE		TBD	Biosamples	TRUE	TRUE
ADNe-TARA		WP 2.2	Arnaud-Haond	Water column	>0.45µm	Seq	Meta B (12S Teleo 04)	MetaB	1	1	Illumina Novaseq	0.25		Seq	60 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBD	local server	csv	FALSE		TBD	Biosamples	TRUE	TRUE
ADNe-TARA		WP 2.2	Arnaud-Haond	Water column	>0.45µm	Seq	Meta B (12S MiFishE Elasmobranchi)	MetaB	1	1	Illumina Novaseq	0.25		Seq	100 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBD	local server	csv	FALSE		TBD	Biosamples	TRUE	TRUE
Kelps	Plug in	WP2.2.1	Coelho	Kelps sp 1	-	Seq	DNAseq (illumina)	-	30		Illumina NovaSeq PE 2x250	20		Short Read Seq		Fastq.gz	local server	TbD	quality and adapter trimming	fasta	TbD	local server		FALSE			Biosamples	FALSE	FALSE
Kelps	Plug in	WP2.2.1	Coelho	Kelps sp 1	-	Seq	DNAseq (illumina)	-	30		Illumina NovaSeq PE 2x250	20		Short Read Seq		Fastq.gz	local server	TbD	quality and adapter trimming	fasta	TbD	local server		FALSE			Biosamples	FALSE	FALSE
Kelps	Plug in	WP2.2.1	Coelho	Kelp swabs	-	Seq	MetaB (16V4V5 Prok/Euk; fungal ITS2)	-	30	30	Illumina NovaSeq PE 2x250	1.25		Short Read Seq	150 mM (compressed	Fastq.gz	local server	TbD	quality and adapter trimming		TbD	local server	csv	FALSE		ENA	Biosamples	TRUE	TRUE
Kelps	Plug in	WP2.2.1	Coelho	Water/Kepa	-	Seq	MetaB (16V4V5 Prok/Euk; fungal ITS2)	-	45080	45080	Illumina NovaSeq PE 2x250	1.25		Short Read Seq	150 mM (compressed	Fastq.gz	local server	TbD	quality and adapter trimming		TbD	local server	csv	FALSE		ENA	Biosamples	TRUE	TRUE
Seagrasses	Plug in	WP2.2.2	Arnaud-Haond	Sea grasses	-	Seq	WGS	WGS	25	25	Illumina Novaseq	TBD		Short Read Seq	TBD	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fasta	TBD	local server	csv	FALSE	no	ENA/EBI-Dataref (Ifremer)	Biosamples	TRUE	TRUE
Seagrasses	Plug in	WP2.2.2	Petersen	Sea grasses		Seq	nested PCR, MetaB (16V4V5 Prok)	Illumina			Illumina Novaseq												FALSE				FALSE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke	-	Imaging	RNA FISH	Confocal	?	?	-	-		-	TBD	.tif	local server	TBD			?	local server	-	FALSE			Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke	-	Imaging	SEM	SEM	?	?	-	-		-	TBD		local server	TBD			?	local server	-	FALSE			Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke Larvae	-	Imaging	RNA FISH	Confocal	?	?	-	-			?	.tif	local server	Luca	-	-	?	local server	-	FALSE		?	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke Larvae	-	Imaging	EM/ Xray	-	?	?	-	-			?	.tif	local server	TBD	-	-	?	local server	-	FALSE		?	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke Larvae	-	Imaging	Gaviscope	-	?	?	-	-			?	.tif	local server	TBD	-	-	?	local server	-	FALSE		?	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Atoke	-	Metabolomics	Lipidomics	-	?	?	-	-		MassSpec	TBD	csv	local server	TBD		csv	?	local server	-	FALSE			Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke	-	Metabolomics	Lipidomics	-	?	?	-	-		MassSpec	TBD	csv	local server	TBD		csv	?	local server	-	FALSE			Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Platy Atoke_Seagrass	-	Metabolomics	Lipidomics	-	?	?	-	-		MassSpec	TBD	csv	local server	TBD		csv	?	local server	-	FALSE			Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Platy Atoke Kelp	-	Metabolomics	Lipidomics	-	?	?	-	-		MassSpec	TBD	csv	local server	TBD		csv	?	local server	-	FALSE			Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke Larvae	-	Metabolomics	Lipidomics	-	>5	>5	-	-		MassSpec	?	csv	local server	TBD	-	csv	?	local server	-	FALSE		?	Biosamples	TRUE	TRUE

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Project area	Project (core vs plug in)	WP/Task/ Subtask	Task Lead	Sample type	Sample fraction	Analysis type	Specific analysis (only 1 per row)	Method	# samples per normal site	# samples per super site	Seq tech.	Nb of reads/sample (million)	Primer ?	Raw data type produced	(Expected) Generated data size/sample	Raw data format	Raw data storage (internal)	QC by	QC processes	Curated data type produced	Curated data size produced	Curated data storage	Format of metadata	Re-use of historic/ existing data?	Re-use: Which data?	Open access - which depository?	Persistent identifier	Sharing data via BSD data hub	Access for all BSD members ok?
Platy	Core	WP2.2.3	Arendt	Platy Epitoke_Larvae	-	Proteomics	Phosphoglyco proteome	-	?	?	-	-		MassSpec	?	csv	local server	TBD	-	csv	?	local server	-	FALSE		?	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Atoke	-	Seq	DNAseq	-	?	?	Nextseq/ NovaSeq	?		Short Read Seq	TBD	fastq	local server	TBD		fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Platy Atoke	-	Seq	Barcoding	-	?	?	Sanger/Nextseq/NovaSeq	?		Sanger/Short Read Seq	TBD	fastq	local server	Leslie	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Platy Atoke	-	Seq	Gut Microbiome (Metall)	-	?	?	Nextseq/ NovaSeq	?		Short Read Seq	TBD	fastq	local server	TBD		fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke	-	Seq	Barcoding	-	?	?	Sanger/Nextseq/NovaSeq	?		Sanger/Short Read Seq	TBD	fastq	local server	Leslie	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke	-	Seq	scRNAseq	-	?	?	Nextseq/ NovaSeq (Pacbio	250mil/sample		Short/Long Read Seq	TBD	fastq	local server	Phil	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke	-	Seq	scATACseq	-	?	?	Nextseq/ NovaSeq	50mil/sample		Short Read Seq	TBD	fastq	local server	Leslie	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke_Water	-	Seq	Microbiome (Metall)	-	?	?	Nextseq/ NovaSeq	?		Short Read Seq	TBD	fastq	local server	TBD		fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke_Water	-	Seq	scRNAseq	-	?	?	Nextseq/ NovaSeq (Pacbio	250mil/sample		Short/Long Read Seq	TBD	fastq	local server	Phil	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke_Larvae	-	Seq	scRNAseq	-	>10	>10	Nextseq/ NovaSeq (Pacbio	250mil/sample		Short/Long Read Seq	?	fastq	local server	Tobi	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke_Larvae	-	Seq	ATACseq	-	>10	>10	Nextseq/ NovaSeq	50mil/sample		Short Read Seq	?	fastq	local server	Leslie	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke_Mated	-	Seq	DNAseq/ RNAseq (female epitoke)	-	20	20	Nextseq/ NovaSeq	TBD		Short Read Seq	?	fastq	local server	TBD	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke_Unmated males	-	Seq	DNAseq	-	10	10	Nextseq/ NovaSeq	TBD		Short Read Seq	?	fastq	local server	TBD	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	TRUE
coral holobiont (host & algae)		WP 2.3.1	Planes	coral host tissues including symbiotic dinoflagellate algae	-	Seq	metaB (16S V4V5)	Illumina	340 metaB	NA (not TREC)	Illumina Novaseq	100		Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
coral holobiont (host & algae)		WP 2.3.1	Planes	coral host tissues including symbiotic dinoflagellate algae	-	Seq	metaG	Illumina	20 metaG	NA (not TREC)	Illumina Novaseq	100		Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
coral holobiont (host & algae)		WP 2.3.1	Planes	coral host tissues including symbiotic dinoflagellate algae	-	Seq	metaT	Illumina	340 metaT	NA (not TREC)	Illumina Novaseq	100		Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
sponge holobiont (host & bacteria-archea-eukaryotes)		WP 2.4	Montoya	sponge host tissue including symbiotic bacteria and archaea	-	Seq	metaB	Illumina	240 metaB	NA	Illumina Novaseq	0.4	16S V4/V5	Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
sponge holobiont (host & bacteria-archea-eukaryotes)		WP 2.4	Montoya	sponge host tissue including symbiotic bacteria and archaea	-	Seq	metaG	Illumina	40-60 metaG	NA	Illumina Novaseq	0.4	16S V4/V5	Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
sponge holobiont (host & bacteria-archea-eukaryotes)		WP 2.4	Montoya	sponge host tissue including symbiotic bacteria and archaea	-	Seq	metaT	Illumina	40-60 metaT	NA	Illumina Novaseq	0.4	16S V4/V5	Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
selected species/plankton	Core	WP 2.5	Decelle	plankton phaeocystis	-	Imaging	TEM of HPF samples	TEM		depends on the presence	-	-		-	TBD	.tif	local server			.tif	TBD	local server		FALSE			Biosamples	FALSE	FALSE
selected species/plankton	Core	WP 2.6	Decelle	plankton phaeocystis	-	Imaging	Volume EM (FIB-SEM, SBF-SEM) of HPF samples	Volume EM		depends on the presence	-	-		-	TBD	.tif	local server			.tif	TBD	local server		FALSE			Biosamples	FALSE	FALSE
selected species/plankton	Core	WP 2.6	Decelle	plankton phaeocystis	-	Seq	metabarcording (16S)	Illumina		depends on the presence	Illumina Novaseq	30		Short Read Seq	TBD	fastq	local server		quality and adapter trimming		TBD	local server		FALSE			Biosamples	FALSE	FALSE
selected species/plankton	Core	WP 2.7	Decelle	plankton phaeocystis	-	Seq	single-cell transcriptomics scRNAseq	Illumina		depends on the presence	Illumina Novaseq	30		Short Read Seq	TBD	fastq	local server		quality and adapter trimming		TBD	local server		FALSE			Biosamples	FALSE	FALSE
global plankton/ecosystem data		WP 3.3, 3.4	Sunagawa	historic and new omics data, multiple methods	-	Modelling	metaomics	Sequence data analysis	NA (not TREC)	NA (not TREC)	-	-	-	csv, text	some GB-TB	.csv; .txt, fasta	local servers	internal peer-review	community standards across data types	.csv	some GB-TB	WP3 data hub, Zenodo	text	TRUE	marine metaomics data	ENA, Zenodo, data type specific repositories	doi	TRUE	TRUE
global plankton/ecosystem data		WP 3.3, 3.4	Vogt	historic obs, multiple methods	-	Modelling	SDMs	SDMs	NA (not TREC)	NA (not TREC)	multiple	~20 million		csv	some GB-TB	.csv	local server, Zenodo, AtlantECO GeoNode, implementation into EMODnet/EurOBIS/GBIF in other EU projects (AtlantECO, BlueCloud2026), cloud servers and data lake in (BlueCloud2026)	peer-review, internal	community standards across data types	.csv, .nc	some GB-TB	WP3 data hub, EMODnet, EurOBIS, GBIF, Zenodo...	text	yes	AtlantECO data collection	Zenodo, data type specific repositories	doi	yes	yes
Plankton		WP4.1	Ken H Andersen	Plankton abundance, taxonomic composition and trait measurments by imaging techniques. Rec to be filled by Fabien Lombard when he is back from cruise ultimo July		Imaging	Zooscan	quantitative imaging methods	TBD	TBD	-	-	-	-	TBD	.tsv	Ecotaxa	internal	Sample quality, signal quality and metadata quality	.tsv	TBD	Ecotaxa	text	TRUE	AtlantECO data collection	Ecotaxa, OBIS/EMODNET	doi in OBIS/EMODNET, data accessibility in ecotaxa	TBD	TRUE

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Project area	Project (core vs plug in)	WP/Task/ Subtask	Task Lead	Sample type	Sample fraction	Analysis type	Specific analysis (only 1 per row)	Method	# samples per normal site	# samples per super site	Seq tech.	Nb of reads/sample (million)	Primer ?	Raw data type produced	(Expected) Generated data size/sample	Raw data format	Raw data storage (internal)	QC by	QC processes	Curated data type produced	Curated data size produced	Curated data storage	Format of metadata	Re-use of historic/ existing data?	Re-use: Which data?	Open access - which depository?	Persistent identifier	Sharing data via BSD data hub	Access for all BSD members ok?
Platy	Core	WP2.2.3	Arendt	Platy Epitoke Larvae	-	Proteomics	Phosphoglyco proteome	-	?	?	-	-		MassSpec	?	csv	local server	TBD	-	csv	?	local server	-	FALSE		?	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Atoke	-	Seq	DNAseq	-	?	?	Nextseq/ NovaSeq	?		Short Read Seq	TBD	fastq	local server	TBD		fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Platy Atoke	-	Seq	Barcoding	-	?	?	Sanger/Nextseq/NovaSeq	?		Sanger/Short Read Seq	TBD	fastq	local server	Leslie	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Platy Atoke	-	Seq	Gut Microbiome (MetaB)	-	?	?	Nextseq/ NovaSeq	?		Short Read Seq	TBD	fastq	local server	TBD		fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke	-	Seq	Barcoding	-	?	?	Sanger/Nextseq/NovaSeq	?		Sanger/Short Read Seq	TBD	fastq	local server	Leslie	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke	-	Seq	scRNAseq	-	?	?	Nextseq/ NovaSeq /Pacbio	250mil/sample		Short/Long Read Seq	TBD	fastq	local server	Phil	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke	-	Seq	scATACseq	-	?	?	Nextseq/ NovaSeq	50mil/sample		Short Read Seq	TBD	fastq	local server	Leslie	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke_Water	-	Seq	Microbiome (MetaB)	-	?	?	Nextseq/ NovaSeq	?		Short Read Seq	TBD	fastq	local server	TBD		fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Annelid Atoke_Water	-	Seq	scRNAseq	-	?	?	Nextseq/ NovaSeq /Pacbio	250mil/sample		Short/Long Read Seq	TBD	fastq	local server	Phil	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	FALSE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke_Larvae	-	Seq	scRNAseq	-	>10	>10	Nextseq/ NovaSeq /Pacbio	250mil/sample		Short/Long Read Seq	?	fastq	local server	Tobi	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke_Larvae	-	Seq	ATACseq	-	>10	>10	Nextseq/ NovaSeq	50mil/sample		Short Read Seq	?	fastq	local server	Leslie	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke_Mated	-	Seq	DNAseq/ RNAseq (female epitoke)	-	20	20	TBD			Short Read Seq	?	fastq	local server	TBD	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	TRUE
Platy	Core	WP2.2.3	Arendt	Platy Epitoke_Unmated males	-	Seq	DNAseq	-	10	10	Nextseq/ NovaSeq	TBD		Short Read Seq	?	fastq	local server	TBD	quality and adapter trimming	fasta	?	local server	-	FALSE		ENA	Biosamples	TRUE	TRUE
coral holobiont (host & algae)		WP 2.3.1	Planes	coral host tissues including symbiotic dinoflagellate algae	-	Seq	metaB (16S V4V5)	Illumina	340 metaB	NA (not TREC)	Illumina Novaseq	100		Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
coral holobiont (host & algae)		WP 2.3.1	Planes	coral host tissues including symbiotic dinoflagellate algae	-	Seq	metaG	Illumina	20 metaG	NA (not TREC)	Illumina Novaseq	100		Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
coral holobiont (host & algae)		WP 2.3.1	Planes	coral host tissues including symbiotic dinoflagellate algae	-	Seq	metaT	Illumina	340 metaT	NA (not TREC)	Illumina Novaseq	100		Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
sponge holobiont (host & bacteria-archea-eukaryotes)		WP 2.4	Montoya	sponge host tissue including symbiotic bacteria and archaea	-	Seq	metaB	Illumina	240 metaB	NA	Illumina Novaseq	0.4	16S V4/V5	Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
sponge holobiont (host & bacteria-archea-eukaryotes)		WP 2.4	Montoya	sponge host tissue including symbiotic bacteria and archaea	-	Seq	metaG	Illumina	40-60 metaG	NA	Illumina Novaseq	0.4	16S V4/V5	Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
sponge holobiont (host & bacteria-archea-eukaryotes)		WP 2.4	Montoya	sponge host tissue including symbiotic bacteria and archaea	-	Seq	metaT	Illumina	40-60 metaT	NA	Illumina Novaseq	0.4	16S V4/V5	Short Read Seq	TBD	fastq	Zenodo/local server	internal	quality and adapter trimming	fastq/fasta	TBD	Zenodo	text	no	-	NCBI, Zenodo or similar public	Biosamples	TRUE	TRUE
selected species/plankton	Core	WP 2.5	Decelle	plankton phaeocystis	-	Imaging	TEM of HPF samples	TEM		depends on the presence	-	-		-	TBD	.tif	local server			.tif	TBD	local server		FALSE			Biosamples	FALSE	FALSE
selected species/plankton	Core	WP 2.6	Decelle	plankton phaeocystis	-	Imaging	Volume EM (FIB-SEM, SBF-SEM) of HPF samples	Volume EM		depends on the presence	-	-		-	TBD	.tif	local server			.tif	TBD	local server		FALSE			Biosamples	FALSE	FALSE
selected species/plankton	Core	WP 2.6	Decelle	plankton phaeocystis	-	Seq	metabarcording (16S)	Illumina		depends on the presence	Illumina Novaseq	30		Short Read Seq	TBD	fastq	local server		quality and adapter trimming		TBD	local server		FALSE			Biosamples	FALSE	FALSE
selected species/plankton	Core	WP 2.7	Decelle	plankton phaeocystis	-	Seq	single-cell transcriptomics scRNAseq	Illumina		depends on the presence	Illumina Novaseq	30		Short Read Seq	TBD	fastq	local server		quality and adapter trimming		TBD	local server		FALSE			Biosamples	FALSE	FALSE
global plankton/ecosystem data		WP 3.3, 3.4	Sunagawa	historic and new omics data, multiple methods	-	Modelling	metaomics	Sequence data analysis	NA (not TREC)	NA (not TREC)	-	-	-	csv, text	some GB-TB	.csv; .txt, fasta	local servers	internal peer-review	community standards across data types	.csv	some GB-TB	WP3 data hub, Zenodo	text	TRUE	marine metaf/omics data	ENA, Zenodo, data type specific repositories	doi	TRUE	TRUE
global plankton/ecosystem data		WP 3.3, 3.4	Vogt	historic obs, multiple methods	-	Modelling	SDMs	SDMs	NA (not TREC)	NA (not TREC)	multiple	~20 million		csv	some GB-TB	.csv	local server, Zenodo, AtlantECO GeoNode, implementation into EMODnet/EurOBIS/GBIF in other EU projects (AtlantECO, BlueCloud2026), cloud servers and data lake in (BlueCloud2026)	peer-review, internal	community standards across data types	.csv, .nc	some GB-TB	WP3 data hub, EMODnet, EurOBIS, GBIF, Zenodo...	text	yes	AtlantECO data collection	Zenodo, data type specific repositories	doi	yes	yes
Plankton		WP4.1	Kiin H Andersen	Plankton abundance, taxonomic composition and trait measurments by imaging techniques. Rest to be filled by Fabien Lombard when he is back from cruise ultimo July.		Imaging	Zooscan	quantitative imaging methods	TBD	TBD	-	-	-	TBD	TBD	.tsv	Ecotaxa	internal	Sample quality, signal quality and metadata quality	.tsv	TBD	Ecotaxa	text	TRUE	AtlantECO data collection	Ecotaxa, OBIS/EMOCNET	doi in OBIS/EMODNET, data accessibility in ecotaxa	TBD	TRUE

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Project area	Project (core vs plug in)	WP/Task/ Subtask	Task Lead	Sample type	Sample fraction	Analysis type	Specific analysis (only 1 per row)	Method	# samples per normal site	# samples per super site	Seq tech.	Nb of reads/sample (million)	Primer ?	Raw data type produced	(Expected) Generated data size/sample	Raw data format	Raw data storage (internal)	QC by	QC processes	Curated data type produced	Curated data size produced	Curated data storage	Format of metadata	Re-use of historic/ existing data?	Re-use: Which data?	Open access - which depository?	Persistent identifier	Sharing data via BSD data hub	Access for all BSD members ok?
Plankton		WP4.1	Ken H Andersen	Plankton abundance, taxonomic composition and trait measurments by imaging techniques. Rest to be filled by Fabien Lombard when he is back from cruise ultimo July		Imaging	Flowcam	quantitative imaging methods	TBD	TBD	-	-	-	-	TBD	.tsv	Ecotaxa	Internal	Sample quality, signal quality and metadata quality	.tsv	TBD	Ecotaxa	text	TRUE	AtlantECO data collection	Ecotaxa, OBIS/EMODNET	doi in OBIS/EMODNET, data accessibility in ecotaxa	TBD	TRUE
Plankton		WP4.2	Ward	locations and times of open ocean samples	-	Modelling	Individual-based modelling of plankton biodiversity	Simulations	NA (not TREC)	NA (not TREC)	-	-	-	-	Order of 1 Tb	tbd	local server	Internal	tbd	tbd	tbd	tbd	text	no		tbd		TRUE	TRUE
ADNe-TARA	Plug-in	WP5.1	Arnaud-Haond	Water column	>0.45µm	Seq	Meta B (16V4V5 Prok/Euk)	MetaB	1	1	Illumina Novaseq	0.25		Seq	250 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBC	local server	csv	FALSE		TBD	Biosamples	TRUE	TRUE
ADNe-TARA	Plug-in	WP5.1	Arnaud-Haond	Water column	>0.45µm	Seq	Meta B (COI Metazoa)	MetaB	1	1	Illumina Novaseq	0.5		Seq	350 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBC	local server	csv	FALSE		TBD	Biosamples	TRUE	TRUE
ADNe-TARA	Plug-in	WP5.1	Arnaud-Haond	Water column	>0.45µm	Seq	Meta B (18S V1V2 Metazoa)	MetaB	1	1	Illumina Novaseq	0.5		Seq	500 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBC	local server	csv	FALSE		TBD	Biosamples	TRUE	TRUE
ADNe-TARA	Plug-in	WP5.1	Arnaud-Haond	Water column	>0.45µm	Seq	Meta B (12S Teleo 04)	MetaB	1	1	Illumina Novaseq	0.25		Seq	60 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBC	local server	csv	FALSE		TBD	Biosamples	TRUE	TRUE
ADNe-TARA	Plug-in	WP5.1	Arnaud-Haond	Water column	>0.45µm	Seq	Meta B (12S MifishE Elasmobranchii)	MetaB	1	1	Illumina Novaseq	0.25		Seq	100 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBC	local server	csv	FALSE		TBD	Biosamples	TRUE	TRUE
Surface sediment		WP 5.1	Cordier	Surface sediment		Seq	metaB, 18S V9	Illumina	NA (not TREC)	NA (not TREC)	NovaSeq	0.5-1	V9	Short Read Seq	50-300 Gb	fastq	locasl server	Genoscope + internal	Basecalling, primers detection	fastq/fasta	50-250Gb	local server	text	yes	local datasets	SRA	BioSample accessions	TBD	TBD
Surface sediment		WP 5.1	Cordier	Surface sediment		Seq	metaB, V1V2 nema	Illumina	NA (not TREC)	NA (not TREC)	NovaSeq	0.5-1	V1V2 nema	Short Read Seq	50-300 Gb	fastq	locasl server	Genoscope + internal	Basecalling, primers detection	fastq/fasta	50-250Gb	local server	text	yes	local datasets	SRA	BioSample accessions	TBD	TBD
Surface sediment		WP 5.1	Cordier	Surface sediment		Seq	metaB, V1V2 forams	Illumina	NA (not TREC)	NA (not TREC)	NovaSeq	0.5-1	V1V2 forams	Short Read Seq	50-300 Gb	fastq	locasl server	Genoscope + internal	Basecalling, primers detection	fastq/fasta	50-250Gb	local server	text	yes	local datasets	SRA	BioSample accessions	TBD	TBD
Alien-Harbour	Plug-in	WP5.1	Viard	Water column	>0.45µm	Seq	Meta B (16V4V5 Prok/Euk)	MetaB	9	9	Illumina Novaseq	0.25		Seq	250 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBC	local server	csv	TRUE	unpublished data or SRA	TBD	Biosamples	FALSE	FALSE
Alien-Harbour	Plug-in	WP5.1	Viard	Water column	>0.45µm	Seq	Meta B (COI Metazoa)	MetaB	9	9	Illumina Novaseq	0.5		Seq	350 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBC	local server	csv	TRUE	unpublished data or SRA	TBD	Biosamples	FALSE	FALSE
Alien-Harbour	Plug-in	WP5.1	Viard	Water column	>0.45µm	Seq	Meta B (18S V1V2 Metazoa)	MetaB	9	9	Illumina Novaseq	0.5		Seq	500 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBC	local server	csv	TRUE	unpublished data or SRA	TBD	Biosamples	FALSE	FALSE
Alien-Harbour	Plug-in	WP5.1	Viard	Water column	>0.45µm	Seq	Meta B (12S Teleo 04)	MetaB	9	9	Illumina Novaseq	0.25		Seq	60 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBC	local server	csv	TRUE	unpublished data or SRA	TBD	Biosamples	FALSE	FALSE
Alien-Harbour	Plug-in	WP5.1	Viard	Water column	>0.45µm	Seq	Meta B (12S MifishE Elasmobranchii)	MetaB	9	9	Illumina Novaseq	0.25		Seq	100 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastQ	TBC	local server	csv	TRUE	unpublished data or SRA	TBD	Biosamples	FALSE	FALSE
Aliens-Harbour (water)		WP 5.1.3	Viard/Turon/ Arnaud-Haond	Filtered seawater	>0.45µm	Seq	Meta B (16V4V5 Prok/Euk)	MetaB	9	9	Illumina Novaseq	0.25-0.5	16V4V5 Prok/Euk	Short Read Seq	1260 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastq	TBC	local server	csv	yes	unpublished data or SRA	TBD	Biosamples	TBD	FALSE
Aliens-Harbour (water)		WP 5.1.4	Viard/Turon/ Arnaud-Haond	Filtered seawater	>0.45µm	Seq	Meta B (COI Metazoa)	MetaB	9	9	Illumina Novaseq	0.25-0.6	COI Metazoa	Short Read Seq	1261 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastq	TBC	local server	csv	yes	unpublished data or SRA	TBD	Biosamples	TBD	FALSE
Aliens-Harbour (water)		WP 5.1.5	Viard/Turon/ Arnaud-Haond	Filtered seawater	>0.45µm	Seq	Meta B (18S V1V2 Metazoa)	MetaB	9	9	Illumina Novaseq	0.25-0.7	18S V1V2 Metazoa	Short Read Seq	1262 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastq	TBC	local server	csv	yes	unpublished data or SRA	TBD	Biosamples	TBD	FALSE
Aliens-Harbour (water)		WP 5.1.6	Viard/Turon/ Arnaud-Haond	Filtered seawater	>0.45µm	Seq	Meta B (12S Teleo 04)	MetaB	9	9	Illumina Novaseq	0.25-0.8	12S Teleo 04	Short Read Seq	1263 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastq	TBC	local server	csv	yes	unpublished data or SRA	TBD	Biosamples	TBD	FALSE
Aliens-Harbour (water)		WP 5.1.7	Viard/Turon/ Arnaud-Haond	Filtered seawater	>0.45µm	Seq	Meta B (2S MifishE Elasmobranchii)	MetaB	9	9	Illumina Novaseq	0.25-0.9	2S MifishE Elasmobranchii	Short Read Seq	1264 Mb	fastq	local server	Genoscope	quality and adapter trimming, demultiplexing	fastq	TBC	local server	csv	yes	unpublished data or SRA	TBD	Biosamples	TBD	FALSE
global plankton/ecosystem data		WP 5.2	Cael	locations and times of open ocean samples	-	Modelling	SDMs	SDMs	NA (not TREC)	NA (not TREC)	-	-	-	csv	MB	csv	Zenodo	Internal	comparison with ocean mask	csv	MB	Zenodo	text	yes	any and all available data compilations	Zenodo	doi	yes	yes
global plankton/ecosystem data		WP 5.2.3	Rubio / Manso	Connectivities between different areas	-	Lagrangian diagnostics	Connectivity maps/matrices or other diagnostics	Lagrangian diagnostics	NA (not TREC)	NA (not TREC)	-	-	-	csv	MB-GB	csv	Zenodo/local server	Internal	comparison with literature or observations		MB-GB	Zenodo	text	yes	historical MARS3D model data and biological observations	Zenodo	doi	yes	yes
Selected species		WP 5.3	Barausse	Traits and diet informations	-	Modelling	Trait based modelling	Simulations	NA (not TREC)	NA (not TREC)	-	-	-	csv	TBD	csv	Aquamaps	Internal / TBD	TBD	csv	TBD	Aquamaps	text	yes	Global	TBD	TBD	TRUE	TRUE
Natural capital accounting of selected habitats, functions, and services		WP6.2.1	Murillas/ D'Alelio	-	-	Literature review methodology	Literature review methodology	-	-	-	-	-	-	csv, text	TBD	.csv, .txt	Zenodo/local server	-	-	.csv, .txt	TBC	local server		no	-		Biosamples	TRUE	TRUE
Natural capital accounting of selected habitats, functions, and services		WP6.2.2.	Murillas/D'Alelio	-	-	Natural accounting methods	Natural accounting methods	-	-	-	-	-	-	csv, text	TBD	.csv, .txt	Zenodo/local server	-	-	.csv, .txt	TBC	local server	csv	yes?			Biosamples	TRUE	TRUE
Natural capital accounting of selected habitats, functions, and services		WP6.2.3	Murillas/D'Alelio	-	-	Trade-offs models - Fuzzy Cognitive Moodels, Nets, ...	Trade-offs models - Fuzzy Cognitive Moodels, Nets, ...	-	-	-	-	-	-	csv, text	TBD	.csv, .txt	Zenodo/local server	-	-	.csv, .txt	TBC	local server	csv	no	-		Biosamples	TRUE	TRUE
OMV		WP 6.3	Bartkowski	preference data	-	Modelling	Choice modelling	mixed logit	NA (not TREC)	NA (not TREC)	-	-		text	10 MB	.csv	local server	Internal		.csv	10 MB	local server	text	FALSE		Zenodo	-	FALSE	TRUE
OMV		WP 6.3	Bartkowski	qualitative discussion data	-	Modelling	Qualitative content analysis		NA (not TREC)	NA (not TREC)	-	-		text	1 MB	.txt	local server	Internal		.txt	1 MB	local server	text	FALSE		TBD	-	FALSE	TRUE

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Annex 2 - Contextual and metadata for new and historic geo-referenced observational data submitted to BIOcean5D Data Hub

Provenance and Overarching fields	ProjectID	Name of the overarching project
	File name	File name of the uploaded file
	ProjectWP	Work Package within the overarching project
	ContactNames (min. 2)	String with names of the people in charge of the dataset
	ContactAdress (min. 2)	String with email addresses of the people in charge of the dataset
	occurrenceID	TBD
Geography	decimalLatitude	Geographic Latitude in decimal degree, following the -180/+180 WGS84 Spatial Reference System (SRS).
	decimalLongitude	Geographic Longitude in decimal degree, following the -180/+180 WGS84 SRS
	geodeticDatum	SRS of the spatial coordinates; give WGS84 only.
	CoordUncertainty	Uncertainty estimate of the decimal coordinates; in meters
	CountryCode	ISO3166-1-alpha-2 code for the country the observations belongs to, e.g., IT for Italy, DE for Germany.
	SamplingDate	Date of the sampling event. (e.g. 'YYYY-MM-DD'). If possible, add time of the day, after the data, as follows: 'THH:MM:SS' (e.g. '2017-09-23T12:04:23').
	TimeZone	time zone to which the time of day refers. If time of the day was not recorded, then just add 'NA' for not available'.If no time zone is specified, local time at sampling site is assumed.
	Bathymetry	Depth of the seafloor at sampling event, in meters, ≤ 0. Will helps us inform whether the observation stems from a coastal environment or not.
	BathySource	String indicating whether Bathymetry was measured at sampling event or inferred a posteriori. Enter 'in-situ' or 'post' for posteriori.
	HabitatType	String indicating the type of habitat the sample was taken from (e.g. shallow waters, sediment, open ocean water column, river plume, river, coral reef, mangrove…)
	LonghurstProvince	Longhurst Province the sample was taken from (one of 56 possible four-letter geocodes. Can be attributed a posteriori from https://github.com/thechisholmlab/Longhurst-Province-Finder.
	Depth	Sample depth (in meters below the local sea surface); > 0; = 0 is surface.
	DepthAccuracy	Single term that describes the accuracy of the collection depth, in meters
	DepthIntegral	Depth span below sea surface, in meters; > 0; = 0 if surface
	MinDepth	minimum depth for depth-integrated quantities, in meters, > 0.
	MaxDepth	maximum depth for depth-integrated quantities, in meters, > 0.
Source identifying	ParentEventID	Describes the parent event, which is composed of one or more sub-sampling (child) events (eventID in next column). (e.g. for TREC:

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

fields		SamplingSite_LSI_# → Aarhus_LSI_1 etc.)
	EventID	Labels the replicate samples (or sub-samples) from a ParentEventID (e.g. a sample number from a station if multiple samples were taken at same sampling station). Make sure each replicate sample receives a unique eventID, which could be based on the unique sample ID in your dataset. (e.g. SamplingSite_LSI_#_SampleType_Transect#_replicate#_sample/protocol → Aarhus_LSI_1_soil_1_1_Ions).
	BioSamplesID	ID of samples registered in the BioSamples depository.
	SampleBarcode	Internal sample barcode number in the stocks and the Sample Hub in Heidelberg (only applicable to TREC-related samples). Alternatively, provide the unique sample name/ID.
	SamplingProtocol	Protocol used to collect the sample (e.g. ShallowWater_MB320, Sediment_Metals, in-situ measurement).
	InstitutionCode	Custodian institution for the data record (ex: EMBL, Ifremer, IMEV, MBA, UU, FUSP etc.)
	SourceDepository	Describes the online archive where the data is stored (e.g. PANGAEA, ENA, MetaboLights, BioImages, OBIS, GBIF, DRYAD etc.)
	OrigCollectionCode	Code given to the collection or the dataset within the SourceDepository (e.g. the code given to the CPR collection in OBIS/GBIF)
	OrigCollectionID	occurrenceID given to the record/measurement in the SourceDepository. Retained to ensure traceability
	BiblioCitation	String indicating the bibliographic citation associated with the data (when possible; can be a dataset, a paper's DOI, a report) if possible
	DateDataAccess	Date at which the data was downloaded from the SourceDepository if it is not your own data; in (e.g. YYYY-MM-DD).
Measurement fields (type, quantity and methodology)	MeasurementID	if applicable- identifier for the measurement or fact (information pertaining to measurements, facts, characteristics, or assertions).
	MeasurementType	The nature of the measurement, fact, characteristic, or assertion (e.g. presence, absence, length, size, abundance, concentration, microplastic counts, carbon flux rate etc.).
	MeasurementTypeID	An identifier for the MeasurementType (global unique identifier, URI). The identifier should reference the MeasurementType in a vocabulary. Where possible, use an URI to identify the quantity you want to describe (e.g. S1228 for 'copies of the nifHgene', or S1230 for 'blood'). List of URIs available here: http://vocab.nerc.ac.uk/collection/S12/current/
	MeasurementValue	The numeric value of the measurement, fact, characteristic, or assertion (e.g. '42' for a microscope count, or '1' for an occurrence, '0' for an absence etc.). No units.
	MeasurementUnit	The unit associated with the MeasurementValue. Recommended best practice is to use the International System of Units (SI). Examples: m, mg, cells.m-3, mgC.m-3… NA in case MeasurementValue is presence/absence.
	MeasurementAcurracy	Numeric value of the potential error associated with the MeasurementValue. Must be in same unit.
	MeasurementValueID	An identifier for facts stored in the column measurementValue (global unique identifier, URI). This identifier can reference a controlled vocabulary (e.g. for sampling instrument names, methodologies, life stages). When the measurementValue refers to a value and not to a fact, the measurementvalueID has no meaning and should remain empty
	MeasurementMethod	Indicates the protocol used to make the measurement.
	SampleAmount	Numeric value indicating the volume, or mass, of sample analysed to make the measurement (e.g. a volume of seawater filtered).
	SampleAmountUnit	Unit corresponding to the SampleAmount (e.g. liter, ml, m3…).

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

	DeterminedBy	Name(s) of the people, groups or organizations who made the measurement, or identified the organism.
	DeterminedDate	The date on which the measurement/identification was made (can differ from the eventDate).
	Note	Any note or comment on the measurement event (e.g. 'Potential contamination', 'Heavy net clogging', 'Rough sea' etc.).
Oberservation Classification fields	OrigScientificName	Scientific name of the observed taxon as reported in the source dataset. Might already be the correct one.
	ScientificName	Corrected scientific name of the taxon, as given in WoRMS.
	WoRMS_ID	Uniform resource identifier issued from WoRMS for the biological taxon recorded.
	TaxonRank	Taxonomic rank at which the taxon was identified and recorded (e.g. species, family, order…).
	Kingdom	Either 'Animalia' (aphiaID #2), 'Plantae' (#3), 'Fungi' (#4), 'Protozoa' (#5), 'Bacteria' (#6), 'Chromista' (#7), or 'Archaea' (#8)
	Phylum	Phylum of the taxon recorded.
	Class	Class of the taxon recorded
	Order	Order of the taxon recorded
	Family	Family of the taxon recorded.
	Genus	Genus of the taxon recorded.
	Species	Species name of the taxon recorded.
	Subspecies	Subspecies, or variety, of the taxon recorded.
	LifeForm	The type of population organisation- and the ecological organisation of the organism recorded (e.g."singular","colonial","symbiotic","free living" etc.).
	AssocTaxa	The function and scientific name of any taxon associated with the biological unit recorded (e.g."HOST_Rhizosolenia").

Co funded by the European Union (GA# 101059915). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.