Ref. Ares (2020)1980204 – 17/05/2023

AtlantECO

Atlantic Ecosystems Assessment, Forecasting & Sustainability

Deliverable number 2.1 – AtlantECO-BASE

Dissemination level: Public

Related Work Package	WP2 Marine ecosystem structures and functions
Related task(s)	2.1 & 2.2, 2.3
Lead beneficiary	ETHZ/UFSCar
Author(s)	Meike Vogt (ETHZ), Fabio Benedetti (ETHZ), Hugo Sarmento (UFSCar), Paula Huber (UFSCar), Clara Arboleda-Baena (UFSCar), Ruby Rose Bader (ETHZ), Dominic Eriksson (ETHZ), Nielja Knecht (ETHZ), Noémy Chénier (ETHZ), Clara Arboleda-Baena (UFSCar), Olivier Jaillon (GENOSCOPE), Paul Frémont (GENOSCOPE/CEA), Fabien Lombard (SU), Lionel Guidi (SU), Florian Ricour (SU/ULiege), Erik Van Sebille (UU), Sophie Schmiz (UU), Darshika Manral (UU), Corentin Clerc (ENS), Gleice Santos (USP), Luigi Maiorano (UNIROMA1), Daniele De Angelis (UNIROMA1), Samuel Chaffron (UNantes), Damien Eveillard (UNantes), Linda Amaral-Zettler (NIOZ), Marion Gehlen (CEA), Germain Benard (CEA), Thomas Frölicher (UBern)
Due date	31.08.2022 as written in the GA
Submission date	17.05.2023 actual submission date
Type	Datasets; Report
Status and version	Final submitted version

Declaration: *Any work or result described therein is a genuine output of the AtlantECO project. Any other source will be properly referenced where and when relevant.*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862923. This output reflects only the author's view and the European Union cannot be held responsible for any use that may be made of the information contained therein.

Table of Content

Add tables of figures and of tables if used in the document

1 Version History

Version	Authors	Summary of changes	Date
1	**Lead authors:** Meike Vogt (ETHZ), Fabio Benedetti (ETHZ), Hugo Sarmento (UFSCar), Paula Huber (UFSCar) **Contributing authors:** Ruby Rose Bader (ETHZ), Dominic Eriksson (ETHZ), Nielja Knecht (ETHZ), Noémy Chénier (ETHZ), Clara Arboleda-Baena (UFSCar), Olivier Jaillon (CEA/GENOSCOPE), Paul Frémont (CEA/ GENOSCOPE), Fabien Lombard (SU), Lionel Guidi (SU), Florian Ricour (SU/ULiege), Erik Van Sebille (UU), Darshika Manral (UU), Sophie Schmiz (UU), Corentin Clerc (ENS), Gleice Santos (USP), Luigi Maiorano (UNIROMA1), Daniele De Angelis (UNIROMA1), Sam Chaffron (UNantes), Damien Eveillard (UNantes), Linda Amaral-Zettler (NIOZ), Marion Gehlen (CEA), Germain Benard (CEA), Thomas Frölicher (UBern)		17.05.2023

2 Executive summary

This deliverable reports on Task 2.2 'Assembly of observations about microbiomes, plastics, the plastisphere and carbon fluxes'. It used protocols established in task 2.1 'Definition of common standards for the assembly of spatially explicit data' to compile, quality-control and grid existing high-quality observations into a knowledge base of observations (D2.1). Data included into AtlantECO-BASE1 consisted of contributions from the five following data sources and tasks: Task 2.2.1 'Microbiome data from traditional microscopy (presence-absence, abundance and biomass)', Task 2.2.2 'Microbiome data from state-of-the-art optical/imaging analysis', Task 2.2.3 'Microbiome and plastisphere data from state-of-the-art genetic analyses', Task 2.2.4 'Nano-, micro and macroplastics data from state-of-the-art sampling methods', and Task 2.2.5 'Carbon flux data from estimated from high resolution bio-optical sensors'. Additional data contributions and mapping efforts from other partners and work packages (Task 2.3) are also included. A comprehensive list and description of all data sets collected can be found in the Appendix Tables to this document.

3 Introduction

In recent years, the amount of data generated that is targeted at an improved understanding of marine ecosystem structuring has exponentially increased, yet efforts to reconcile this new information with existing historic data have been scarce, even for relatively well-observed ocean basins such as the Atlantic Ocean. However, data collection has been patchy and biased to target locations such as the North Atlantic, with other areas scarcely sampled, or devoid of publicly available data (South Atlantic Ocean). Initial efforts to generate a knowledge base for plankton functional type biomasses have been made at the global scale (MAREDAT, Buitenhuis et al., 2013), but these datasets do not include the data generated during the past decade, nor do they include the wealth of observations made with novel state-of-the-art sampling methods such as optical imaging or genetic methods. More recent efforts have been made to describe global plankton diversity and distribution using presence-absence data (PHYTOBASE; Righetti et al. 2020; ZOOBASE, Benedetti et al. 2021), or semi-quantitative methods (e.g. Continuous Plankton Recorder CPR; Richardson et al., 2006), and there are now first global collections of metagenomic (e.g. Tara Expeditions) and imaging data (Kiko et al. 2022; Drago et al. 2022), plus a wealth of ever-increasing global data repositories (e.g. COPEPOD; O'Brien, 2010), and a range of individual cruises and expeditions that make their data publicly available.

Here, we aim to bring together and harmonise publicly available geo-referenced data on the marine microbiome and its functions and stressors/pollutants from major data repositories with large-scale expeditions gathered during the past decade, across taxonomic groups and sampling methods to generate global data products (for use in statistical mapping and extrapolation) with a particular emphasis on the Atlantic Ocean (All-Atlantic Microbiome Atlas). We gathered marine microbiome data from traditional (task 2.2.1), imaging (task 2.2.2), and genetic (task 2.2.3) methods, as well as information on ecosystem functioning (carbon fluxes, task 2.2.4) and ecosystem pollutants (plastics, task 2.2.3). Presence-absence, abundance, biomass and diversity data sets are available for use within the project in WP3 (marine ecosystem health and services), WP5 (advances in systems ecology), WP6 (advances in model predictability), comparable with assessments carried out in WP7 (ecosystem change) and will serve as inputs to WP8 (predictions of and for ecosystem services). Vice versa, we also record a range of output products generated in the other work packages in our collection, such as data on connectivity, plankton networks and a range of additional 'omics resources.

All geo-referenced data are provided as a collection of data csv and netcdf files (AtlantECO-BASE1, D2.1). Data that are either not geo-referenced and/or unsuitable for re-formatting into csv and netcdf formats are retained in their native data formats and recorded as links to published material (section 4.8; AtlantECO-ELSE), and an initial collection of CMIP6 simulations for future projections is also made available for use (section 4.11; AtlantECO-CMIP6). The collection further includes preliminary links to a first set of globally extrapolated gridded data fields based on the raw data generated using state-of-the-art species distribution modelling (AtlantECO-MAPS; D2.2, section 4.9). This first set of globally extrapolated gridded data fields has been be generated from AtlantECO-BASE1 using state-of-the-art species distribution modelling (AtlantECO-MAPS1; D2.3 due on month 36, August 2023). A range of visualisation tools has been created to showcase these data sets that will be of use in WP9 (communication, capacity building, and outreach; section 4.10). A second version of AtlantECO-BASE (D2.3) and of AtlantECO-MAPS (D2.4) are due towards the end of the project. All data sets will be made available to partners belonging to our sister projects and the wider science community via WP10 or upon e-mail request. The dissemination of the gridded and extrapolated data products as public data layers via AtlantECO's GeoNode (https://atlanteco-geonode.eu/) is ongoing (SBE).

4 AtlantECO-BASE1: Data products

4.1 Data standards, quality control and formatting (task 2.1; lead: ETHZ)

All geo-referenced raw data sets have been formatted as long table using the Darwin Core (https://dwc.tdwg.org/) standard vocabulary (DwC) for transmitting information about biodiversity, including a standard list of data and meta-data descriptors (data column headers), which makes the data compatible with major international data repositories such as the EMODnet Biology Portal (see https://www.emodnet-biology.eu/data-infrastructure), the Global Biodiversity Facility (GBIF) and the Ocean Biodiversity Information System (OBIS). In line with the convention of these data repositories, all taxonomically annotated microbiome data uses the World Register of Marine Species (WoRMS) taxonomic classification for reference (https://www.marinespecies.org/). Gridded data files have been created in NetCDF format using CF metadata (https://cfconventions.org/) for the dissemination on AtlantECO's GeoNode. Each data set generated anew within the framework of AtlantECO by AtlantECO partners is accompanied by a README file that details authors, variable names, sources, metadata included, methodology used to create added value (e.g. biomass conversion from abundance to biomass), as well as references to related scientific publications. Data sets that were generated either by or in collaboration with third parties but serviced to AtlantECO are reported in their original form.

All traditional microbiome data gathered have been taxonomically matched to the WoRMS taxonomic backbone for reference, quality controlled (as detailed in Righetti et al., 2020 for phytoplankton; and Benedetti et al., 2021 for phyto- and zooplankton) and recorded with all meta-data available. Biological records associated with non-marine organisms were discarded. Biological records (presence-absence or abundances) with missing or erroneous spatial coordinates or sampling date (day/month/year) were removed. Biological records missing information about their discrete sampling depth, or sampling depth interval ('MinDepth' and 'MaxDepth'), were also removed. For the datasets synthesising quantitative counts (abundances and biomass records), the records associated with negative or missing abundance/cell concentration values were also considered dubious and therefore discarded. Microbiome/plankton taxa whose scientific name could not be associated with any accepted name (e.g., an accepted AphiaID numeric code from WoRMS) were flagged as such in the 'WoRMS_ID' header. Data quality flags and/or comments inherited from the original data sources were retained (in the 'Note' and 'Flag' headers). All metadata headers were reformatted according to the DwC conventions. Data that were recorded in public long-term data repositories such as OBIS or GBIF have retained their full provenance (e.g., original records' ID and metadata), and all original literature references were given where data could be matched against a peer-reviewed publication. All available absences (i.e., null abundances) were retained, but not all original sources consistently recorded absences (MAREDAT; CPR).

All metagenomic data gathered have been submitted as an AtlantECO Superstudy on EMBL-EBI's MGnify platform (https://www.ebi.ac.uk/metagenomics/super-studies/atlanteco) and are being processed using MGnify's bioinformatics analysis pipelines (version 5.0). Only georeferenced samples with sampling date and water column depth information were considered. The raw data underwent a quality control and contamination removal process using *bwa-mem* algorithm as described in Mitchell et al. (2020). A subset of metagenomes already analyzed was selected to explore the microbiome functional diversity. The dataset contains only samples from the 0.22-3 µm size fraction. Samples with <10,000 reads were removed and total reads number was normalized to equalize sampling depth. Each data record is accompanied with sample and contextual environmental metadata. All metadata headers were reformatted according to the DwC conventions. The metabarcoding dataset consists of an integration of quality-controlled sequencing data from the Tara Ocean and Malaspina expeditions. The primers used for the V4-V5 region of the 16S rRNA gene were 515FB: GTGYCAGCMGCCGCGGTAA and 926R: CCGYCAATTYMTTTRAGTTT (Quince et al., 2011; Parada et al., 2016) and for the V9 region of the 18S rRNA gene were 1389F: 5'-TTGTACACACCGCCC-3' and 1510R: 5'-

CCTTCYGCAGGTTCACCTAC-3' (Amaral-Zettler et al., 2009). Amplicon sequences were processed using the DADA2 pipeline (Callahan et al., 2016; Lee, 2019) to characterize Amplicon Sequence Variants (ASVs) that were used as a proxy of microbial species (Callahan et al., 2017). Each sequencing project was analyzed separately because different runs can have different error profiles following Callahan et al. (2016). The quality of the samples was explored, and the trimming and filtering parameters were chosen according to Callahan *et al.* (2016). After merging the runs, the taxonomic classification was performed using the IDTAXA algorithm implemented in the DECIPHER package for the R programming language (Murali et al*.,* 2018) and the SILVA database (SILVA SSU r138 2019) for 16S rRNA gene primers (Quast et al., 2012) and PR2 database (PR2 v4.13) for18S rRNA gene primers as a reference (Guillou et al., 2012). Only samples with more than 10,000 reads were analyzed, and we kept ASVs with 50 reads distributed in at least three samples or those that have less than 50 reads distributed in more than three samples.

All optical imaging data processed within AtlantECO (including Tara Oceans and Tara Pacific) is recorded as images on ECOTAXA (https://ecotaxa.obs-vlfr.fr) with meta-data standards complying with Pesant et al., (2015) and sample registry (https://doi.pangaea.de/10.1594/PANGAEA.875582), station registry (http://doi.pangaea.de/10.1594/PANGAEA.842237) and event registry (http://doi.pangaea.de/10.1594/PANGAEA.842227) located within Pangaea for the Tara Ocean cruise and ongoing similar effort for Tara Pacific (Lombard et al., accepted, 2022, and cited registry therein). All positions, depth and sampled volumes were triple checked against registries and logsheets. Protocols and QC had followed protocols from the Quantitative Imaging Platform of Villefranche (PIQv https://sites.google.com/view/piqv, Jalabert et al., 2022). Zooscan, UVP and Flowcam data were segmented after background removal by zooprocess, multiples objects on zooscan were QC and further segmented (https://sites.google.com/view/piqv/softwares/flowcamzooscan?authuser=0). Planktoscope data were acquired according to newly established protocols and imported in Ecotaxa (Lombard et al., 2022a; Lombard et al., 2023). All images were annotated to a given taxonomic level (taxonomically matched to the WoRMS taxonomic backbone for reference) by taxonomic expert using an IA assisted method using standard Ecotaxa protocols (Irisson et al., 2022).

All carbon data acquired within AtlantECO have been gathered from published data collections (Underwater Vision Profiler 5 (UVP5) data from https://doi.pangaea.de/10.1594/PANGAEA.924375, sediment traps data (non-exhaustive) from https://doi.pangaea.de/10.1594/PANGAEA.855600 and Thorium-234 data (non-exhaustive) from https://doi.pangaea.de/10.1594/PANGAEA.809717). UVP5 data were quality checked following Kiko et al., (2022). Sediment traps and Thorium-234 data were quality checked (geolocalization error, duplicates, range and outlier checks) using all meta-data available. The effort to derive a general relationship to convert particle size distributions (PSD) into estimates of particulate organic carbon (POC) fluxes is ongoing (Ricour et al., PhD thesis, ongoing). Hence, we first provide UVP5 PSD instead of POC flux estimations in this first data release.

All microplastics data were retrieved from the literature and combined with the global collection by Sebille et al., (2015) for small plastics (microplastics mostly < 5 mm in diameter). The description below closely follows that of Schmiz, MSc thesis, (2021; doi:10.5281/zenodo.5338790). Beyond the approximately 11,000 data records already in Van Sebille et al., (2015), new datasets were collected using searches on Google, Google Scholar, Google Dataset Search, and LITTERBASE. This resulted in 648 new data points from 14 data sets with particles sizes ranging from 32 µm to 25 mm. If plastic particles were sampled by nets, they were trawled mostly at the sea surface with bongo, neuston or manta nets with mesh sizes between 100 µm and 500 µm, although the typical mesh size for these trawls is 300 µm. One data set was collected in a multi-level trawl in the water column. Other samples in the water column were obtained by pumping seawater through a mesh at the bottom of the vessel. Because of contamination concerns, most data sets excluded fibres, except for Brach et al., (2018), Kanhai et al., (2017) and Montoto-Martınez, Hernandez-Brito, and Gelado-Caballero (2018). Although the mesh size of the trawling nets and in the pump filters generally sets a lower limit to the

detectable particle size, it is not uncommon that smaller particles end up in the final sample as well, either as aggregates or by clogging the mesh. Different data sets are often given in different units that need to be converted before the data sets can be combined. To this end, data were converted to units of number of particles per km2 and total mass per km2 using conversion routines following Kukulka et al., (2012), as detailed in section 2.3 of Schmiz, (2021), doi:10.5281/zenodo.5338790.

Where AtlantECO-specific model runs for AtlantECO-MAPS1 (D2.3) were made (e.g. Benedetti et al., (2021)), statistical species distribution models for the extrapolation of geo-referenced data in space and time have been developed in accordance with the model quality protocols developed for the modelling of scarce and biased data (task 2.1), as detailed in Righetti et al., submitted (doi: https://doi.org/10.1101/2023.02.28.530497), and species distribution models have been documented in agreement with current community standards as documented in Zurell et al., (2020). Outputs of species distribution models and other analysis products that have been developed by external partners to the project, or prior to the project start are recorded as published.

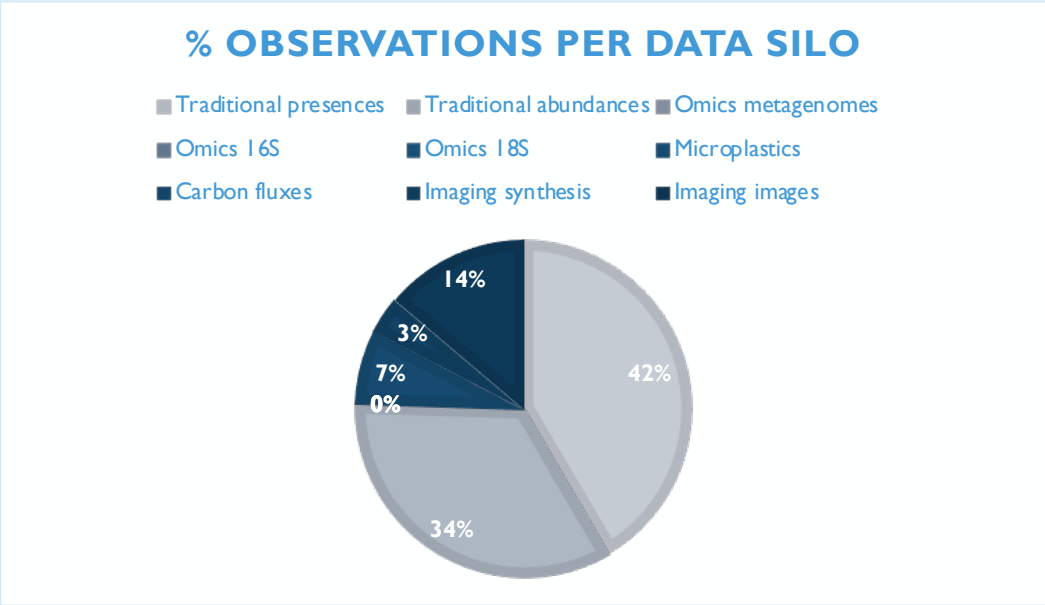
4.2 Data access

All data can already be accessed by the project team from a dedicated public transfer folder located at partner ETHZ. Raw data is accessible at https://data.up.ethz.ch/shared/AtlantECO/BASE/, gridded raw data in netcdf format at https://data.up.ethz.ch/shared/AtlantECO/GRID/, and fields extrapolated using statistical and machine learning tools at https://data.up.ethz.ch/shared/AtlantECO/MAPS/. MAPS will partly be hosted by and distributed through AtlantECO's GeoNode (https://atlanteco-geonode.eu/). Data submissions from previous publications are referenced with their original references and data dois, so that all data published in international data repositories can also directly be sourced from the respective original data providers.

A comprehensive list of all data sets collected for overview can be found in the **Appendix Tables** to this document.

SUMMARY STATISTICS:

Data silo	Data subset	Number of observations
Traditional	Presence-absence	24,556,001
Traditional	Abundance/Biomass	20,026,191
Omics	Metagenomes	835
Omics	16S	428
Omics	18S	1405
Imaging	Synthesis	2,104,712
Imaging	Images	8,183,578
Carbon fluxes	134Th/Sediment traps	15,426
Carbon fluxes	UVP particle sizes	2,083,352
Microplastics		13,158



4.3 Microbiomes data from traditional microscopy (task 2.2.1; lead: ETHZ)

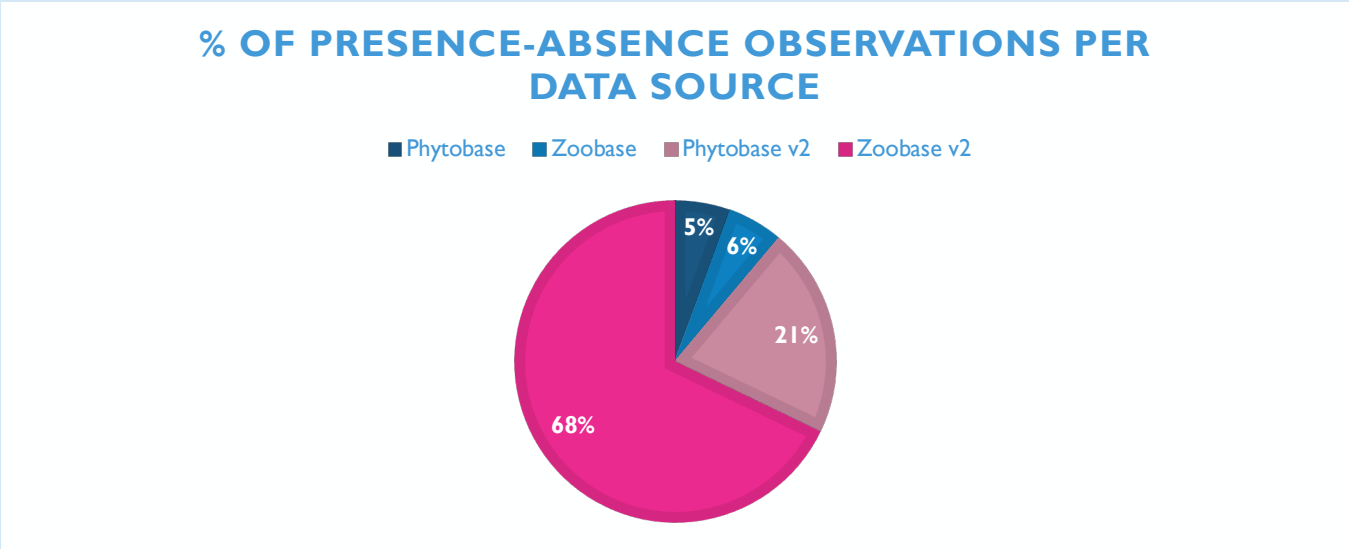
This collection brought together existing observations on the diversity and distribution of phyto- and zooplankton functional types from a range of publicly available resources and AtlantECO target expeditions on presence-absences (including updated versions of PHYTOBASE and ZOOBASE; Righetti et al., 2020; Benedetti et al., 2021), abundances (including observations from MAREDAT; Buitenhuis et al., 2013), the CPR surveys, BODC, COPEPOD, JeDI, KRILLBASE, Malaspina, Tara, among other efforts), and biomasses. Data have been quality controlled according to Benedetti et al., (2021), using the WoRMS' taxonomic classification. Biomasses have been calculated according to community standard protocols, e.g. those used in MAREDAT (Buitenhuis et al., 2013) or as described in the relevant readme file and accompanying literature (see Clerc et al., doi:10.5194/bg-20-869-2023; Knecht et al., doi:10.2254/essoar.167283650.05543210/v1; Eriksson et al., in preparation; Chénier et al., in preparation).

All in all, the data collection brought together > 21 million presence-absence and abundance observations, including 5,167,282 presence-absence observations for phytoplankton (1,691 accepted taxonomic names), 16,574,064 presence-absence observations for zooplankton (4,062 accepted taxonomic names), 4,756,047 phytoplankton abundance and biomass observations (1,072 accepted taxonomic names), 15,294,171 zooplankton abundance and biomass observations (1,262 accepted taxonomic names) and 22,461 nitrogen fixer abundance and biomass observations (20 species from 15 genera). These data sets are documented in a range of MSc theses (N. Knecht, MSc thesis, 2022 (pteropods and foraminifera); N. Chénier, MSc thesis, 2022 (dinoflagellates)) and new papers (Benedetti et al., 2021; Benedetti et al., 2022 (phytoplankton and zooplankton); N. Knecht et al., submitted (pteropods and foraminifera; preprint: doi:10.2254/essoar.167283650.05543210/v1; Clerc et al., 2023 (salps; doi:10.5194/bg-20-869-2023); Eriksson et al., in preparation (diazotrophs); Chénier et al., in preparation (dinoflagellates)). Existing data on diatoms (Leblanc et al., 2012) and coccolithophores (O'Brien et al., 2013) from MAREDAT (Buitenhuis et al., 2013), phytoplankton presence absence data from PHYTOBASE (Righetti et al., 2020) and zooplankton data from ZOOBASE (Benedetti et al., 2021) has been reformatted and delivered in its original form.

A list of all data products can be found in **Tables A 4.3.1 and 4.3.2**.

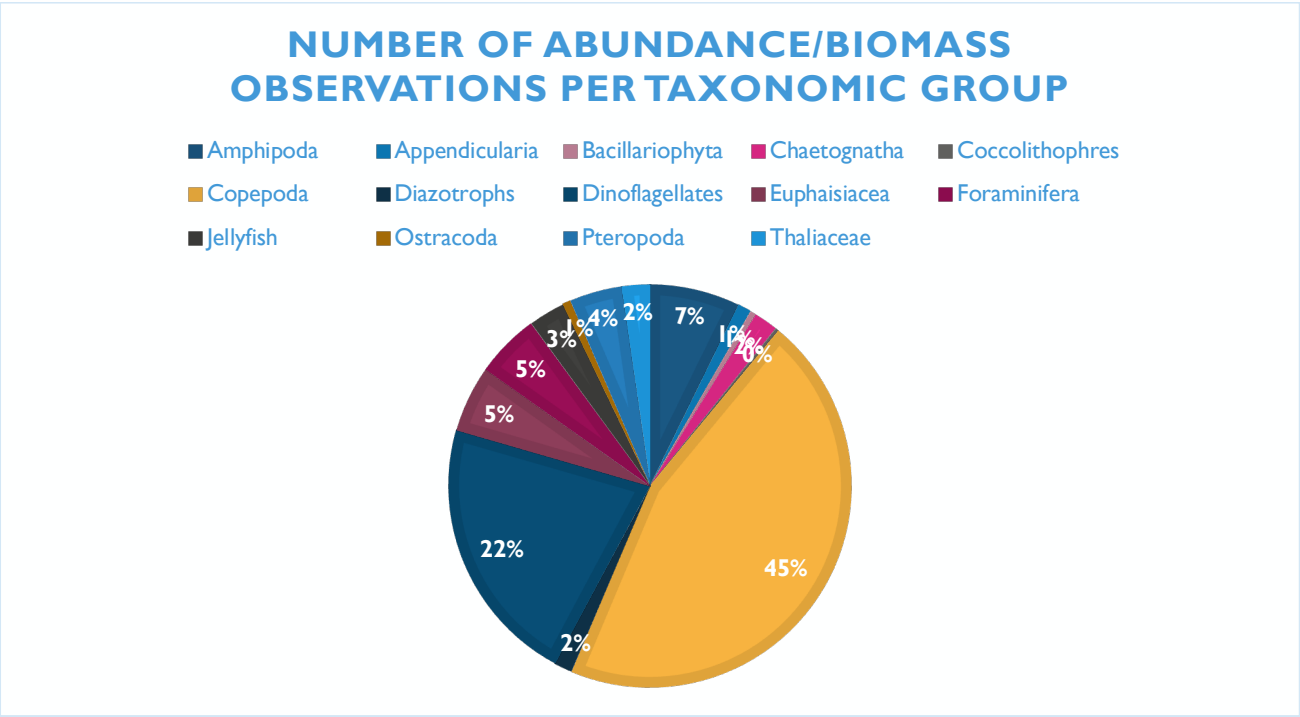
SUMMARY STATISTICS:

Presence-absence observations:



Total number of data sets: 4
Total number of observations: 24,556,001
Fraction of publicly available data sets from global data repositories: 50% (2,725,852 observations; 11% of obs)
Fraction of data sets available by download/upon e-mail request: 50% (21,830,149 observations; 89% of obs.)

Abundance and biomass observations:



Total number of data sets: 14
Total number of observations: 20,026,191
Fraction of publicly available data sets from global data repositories: 36% (2,465,117 observations; 12% of obs)
Fraction of data sets available by download/upon e-mail request: 64% (17,561,074 observations; 87% of obs.)

4.4 Microbiomes data from optical/imaging analysis (task 2.2.2; lead: SU)

The AtlantECO-BASE collection of imaging data consists of (a) the integration of new quality-controlled data from historic cruises (Tara Oceans, Tara Pacific, and 147 others cruises for UVP data), and (b) published synthesis products on particle size distribution and zooplankton community structure and biomass. The former data sets have been integrated with Ecotaxa (https://ecotaxa.obs-vlfr.fr/; Picheral et al., 2017) and EcoPart (https://ecopart.obs-vlfr.fr/ for UVP -particles data), which allows an automatic prediction of the taxonomic classification of each single image followed by a manual validation/correction. Full methodological details on data treatment and processing using machine learning are given in Irisson et al. (2022). Published synthesis products consist of a range of project-associated and or related integrated data products based on imaging data that have been published in collaboration with AtlantECO partners (Kiko et al. 2022; Drago et al. 2022; Brandao et al. 2021). Published data were quality-controlled and annotated according to community standards and using standard Ecotaxa protocols (Irisson et al., 2022), see also https://sites.google.com/view/piqv/piqv-manuals/ecotaxaecopart-manuals?authuser=0. All new Ecotaxa data is directly integrated with EMODnet Biology, and thus OBIS and GBIF, according to new exchange protocols (Martin-Cabrera et al. 2022; doi:10.3897/biss.6.94196), as developed within the H2020 project JERICO SE in collaboration with EMODnet.

All in all, the data collection brought together 15,908 observations of zooplankton abundance and community composition from Tara Oceans (Brandao et al. 2021; Soviadan et al., 2022). It also delivered the compilation of 8, 805 vertical profiles of particle size and concentration distributions based on a compilation of multiple UVP5 camera systems (Kiko et al. 2022), and 466,872 observations of zooplankton abundances from 3,449 vertical UVP5 profiles (Drago et al. 2022), in the form originally published by their project-external first authors. In addition, the collection includes 8.1 million new quality-controlled plankton images from 3 cruises integrated into the Ecotaxa imaging platform (https://ecotaxa.obs-vlfr.fr/), from the following projects/studies: Tara Ocean (3.9 million images), Tara Pacific (2.9 million images) and the Flagship cruise from AtlantECO Tara Microbiomes (>1.2 million images, sample analysis on-going) across a range of different imaging instruments and nets. Finally, the data set includes a compilation of UVP JERICO data (74,769 images). Novel imaging data has been documented and interpreted in a range of scientific publications (Drago et al. 2022; Kiko et al. 2022; Mériguet et al. 2022; Lombard et al., in prep.; Brandao et al. 2021; Soviadan et al., 2022).

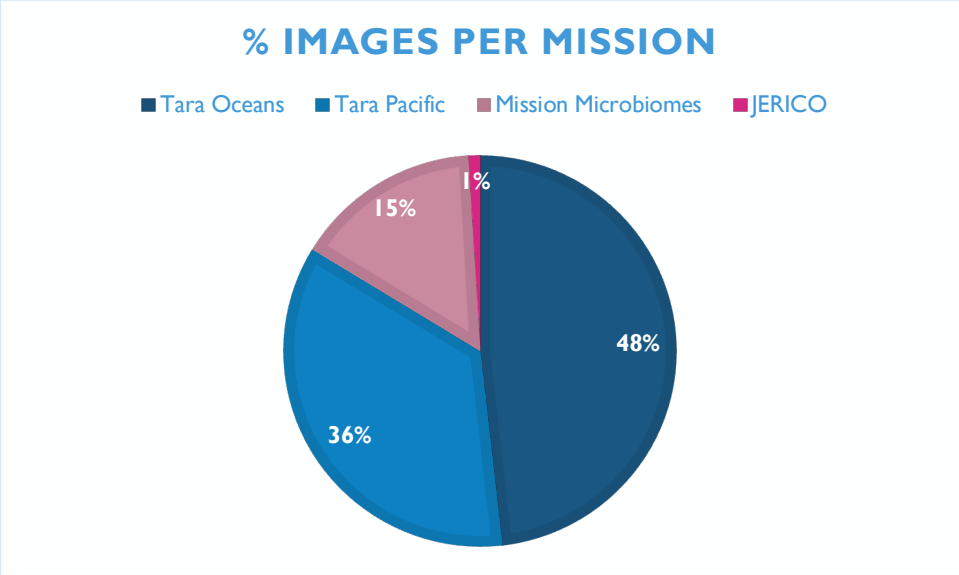
A list of all data products can be found in **Table A 4.4**.

SUMMARY STATISTICS:

Synthesis products:

Total number of integrated data sets: 4
Total number of observations: 2,104,712
Fraction of publicly available data sets in global data repositories: 75% (2,104,675 observations; ~100% obs.)
Fraction of data sets available by download/upon e-mail request: 25% (37 stations; >1% of obs.)
Fraction of data sets currently not available: 0%

Raw images processed and uploaded onto Ecotaxa:



Total number of Ecotaxa projects: 57
Total number of new quality controlled Ecotaxa images: 8,183,578
Fraction of currently publicly viewable validated Ecotaxa projects as recorded below: 43%
(Tara Oceans: 87%, Tara Pacific: 31%, Mission Microbiome: 88%, JERICO: 100%)

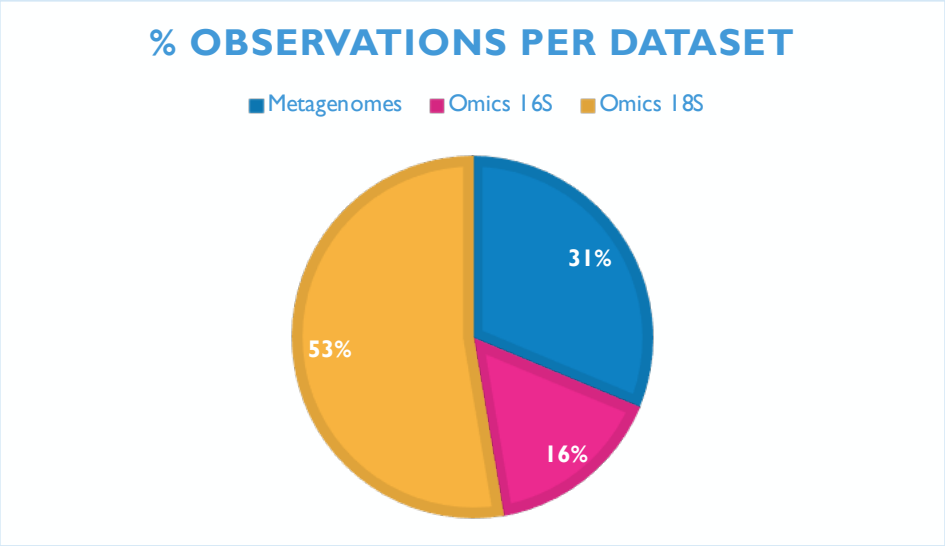
4.5 Microbiomes and plastisphere data from genetic analyses (task 2.2.3; lead: UFSCar)

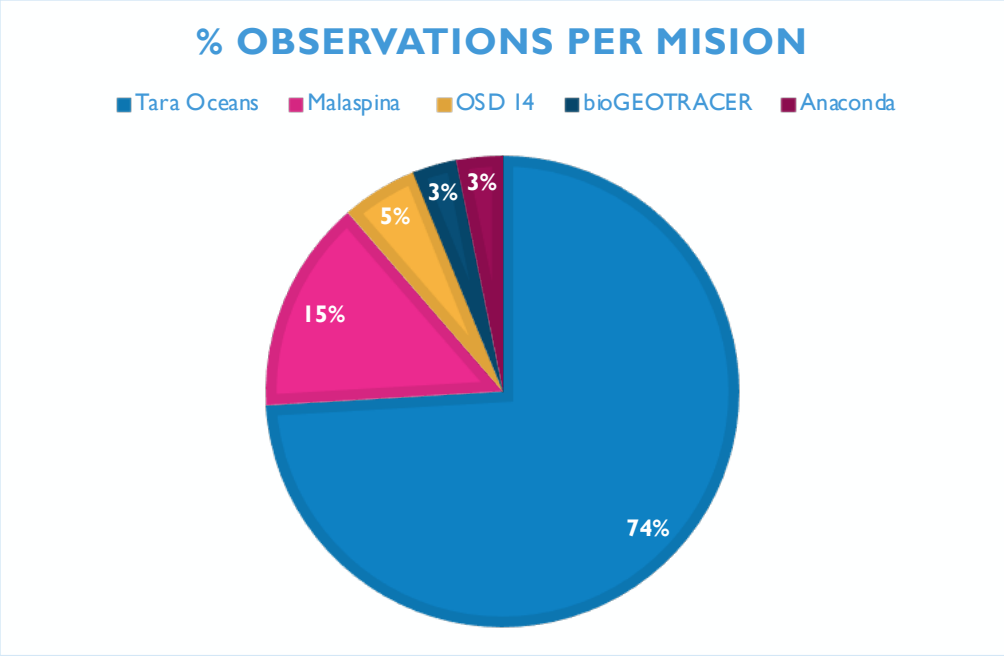
The AtlantECO-BASE collection of genetic data consists of metabarcoding and metagenomics datasets from planktonic communities and plastisphere across different aquatic environments. Publicly available m etagenomics data were obtained from the European Nucleotide Archive (ENA), corresponding to eight expeditions, including Tara Oceans, Malaspina, GO-SHIP, GEOTRACES, and OSD, as well as 16 oceanic cruises. Raw data have been submitted to MGnify and organized under the umbrella of the 'AtlantECO' superstudy. The raw data analysis is currently being carried out using the last version of MGnify's pipeline (V5.0). Data have been quality controlled as described in Mitchell et al. (2020). Amplicon data from planktonic communities have been collected from Tara Oceans and Malaspina expeditions. All data has been quality-controlled according to Callahan et al., (2016).

All in all, the compilation brought together 3,597 marine metagenome observations, 428 16S rRNA and 1,405 18S rRNA amplicon observations, 357 16S rRNA plastisphere amplicon observations, and 162 plastisphere metagenomes. The compilation includes: (1) microbial functional diversity (from 451 metagenomes of 0.2-3µm size fraction), (2) counts of representative genes of Nitrogen metabolism (from 451 metagenomes of 0.2-3µm size fraction), (3) taxonomic diversity of prokaryotes (from 428 samples with a total of 13378 ASVs), (4) autotrophic prokaryotes abundances (from 428 samples with a total of 332 ASVs), (5) heterotrophic prokaryotes abundances (from 428 samples with a total of 13046 ASVs), (6) taxonomic diversity and abundance of eukaryotes (from 1405 samples with a total of 9175 ASVs). A subset of the metagenomic data collected in AtlantECO (Tara) has been documented in Frémont et al. (2022), and a new publication that will analyse the full AtlantECO data set is in preparation (Huber et al. in preparation).

A list of all data products can be found in **Table A 4.5**.

SUMMARY STATISTICS:





Total number of data sets: 3
Total number of observations: 2668
Fraction of publicly available data sets from global data repositories: 100% (2668 observations; 100% of obs)
(Tara Oceans: 74.1%, Malaspina 14.6%, OSD 14: 5.2%, bioGEOTRACER: 2.9% Anaconda: 3.1%)

4.5.1 Plastisphere Community from Amaral-Zettler et al. (2015)

Amplicon data from the plastisphere were collected in different environments from the North Atlantic and Pacific Oceans. Data quality was controlled as described in Amaral-Zettler et al. (2015). Raw data is freely available in the ENA repository. This submission consists of a dataset of bacterial communities from plastic

samples (Plastisphere) from the North Pacific and North Atlantic subtropical gyres. The bacterial community composition was studied by amplicon sequencing Illumina technology using the V6 hypervariable region of the 16S rRNA gene. The sampling strategy and sample processing are described in Amaral-Zettler et al. (2015). The raw data was submitted in MGnify (MGYS00001767) and analyzed using version 3.0 of the amplicon pipeline.

A list of all data products can be found in **Table A 4.5.1**.

4.6 Microplastics data (task 2.2.1; lead: UU)

Microplastics data were obtained from an extensive literature review (Sebille et al. 2015) which was augmented with data from 16 additional cruises (Schmiz, 2021, MSc thesis). Data include counts of small plastics with a size < 5 mm, which were gathered using a range of nets with varying mesh sizes. All data were converted from numbers and concentrations per volume or area into concentrations per km^2 using the conversion routines as documented in Kulkera et al. (2012). All in all, the data compilations contain 13,159 small and microplastics observations.

A list of all data products can be found in **Table A 4.6**.

4.7 Carbon flux data from bio-optical sensors (task 2.2.1; lead: SU)

The compilation of particulate organic carbon fluxes results from the compilation of historical data obtained with sediment traps and derived from Thorium-234 deficits for a total of 15,425 observations. These data were QC,d as described above. The Underwater Vision Profiler (UVP5) measures the size distribution of particles in the 0.05 - 26 mm range (27 size classes). 8,728 UVP5 profiles were quality controlled according to Kiko et al. (2022) for a total of 2,083,'352 observations. Carbon fluxes were derived from UVP data according to the methodology as detailed in Guidi et al. (2008). In brief mass and settling rates of particles, *m(d)* and *w(d)*, respectively, are often described as power law functions of their diameter (*d*) obtained by fitting observed data, $m(d)x\ w(d)=Ad^B$. The particles carbon flux can then be estimated using an approximation of the integral of $n(d)xm(d)x\ w(d)$ (n being the number of particles) over a finite number of small logarithmic intervals for diameter *d* spanning from 250 µm to 1.5 mm (particles <250 µm and >1.5 mm are not considered, consistent with the method presented Guidi et al. 2008.

SUMMARY STATISTICS:

Total number of data sets: 3
Total number of observations: 4,182,130

Fraction of publicly available data sets from global data repositories: 0% (0 observations; 0% of obs)
Fraction of data sets available by download/upon e-mail request: 100% (4,182,130 observations; 100% of obs.)

A list of all data products can be found in **Table A 4.7**.

4.8 AtlantECO-ELSE: Other data sets with data-type specific formatting (D2.1; UU; GENOSCOPE; UNantes; NIOZ)

We further collected a range of other derived data products that cannot easily be converted into the template long-table format due to their underlying structure and data types. These data will be recorded as links to publicly available data repositories.

4.8.1 Connectivity data from Manral et al. (2022)

The connectivity dataset is generated from 120 lagrangian simulations of virtual particles for 30 days across the Atlantic Ocean at different depths using the GLOB16 ocean model data provided by CMCC. The connectivity between all the grid cells in the Atlantic Ocean region is represented in the form of a binary matrix (.npz file). In addition, different statistics of minimum and maximum temperatures experienced by particles connecting grid cells have been exported in matrices (.npz files), similar to the connectivity matrices. We also performed sensitivity analysis for grid cell resolution and number of particles using surface simulations. This dataset can be used to produce passive or thermally constrained connectivity maps between any set of locations within the Atlantic region.

4.8.2 Further metagenomic resources from Tara Oceans distributed via GENOSCOPE

This submission consists of a large range of analysis products based on genomic and transcriptomic information from the stations of the Tara Oceans expedition, as processed at GENOSCOPE (www.genoscope.cns.fr/tara), and documented in the related publications (Delmont et al. 2022a, 2022b, Vorobev et al. 2018, Carradec et al. 2020, Seeleuthner et al., 2018). Eukaryotic metagenomes have been assembled into metagenome-assembled genomes (MAGs) and added to a collection of genomes obtained from individual cells (SAGs), giving a total of 683 genomes. This collection covers a large range of the eukaryote radiation. Bacterial and archeal metagenome data (MAGs) contains 1888 curated MAGs, including 48 curated diazotroph MAGs. All these MAGs, eukaryotes and prokaryotes, have been manually curated using the Anvio platform (https://anvio.org/ ref: doi.org/10.1038/s41564-020-00834-3). Positions of genes were predicted on all these genomes. MATOU is a collection of assembled sequences of transcribed mRNA named unigene sequences. Different annotations have been computed (taxonomic affiliations, protein domain identification; Carradec et al. 2020). MGT is a collection of sequences obtained through a clustering of the MATOU dataset. Therefore, each MGT corresponds to a partial collection of expressed genes, likely corresponding to the same species (Vorobev et al. 2018).

4.8.3 Community networks from Chaffron et al. (2021)

This submission contains a global ocean cross-domain plankton co-occurrence network—the community interactome based on taxonomically annotated OTUs from 115 stations from the Tara Oceans expeditions (2009–2013), covering several organismal size fractions and all major oceanic provinces, with depths corresponding to the surface ocean and the deep chlorophyll maximum only. Co-occurrence graphs were derived from the integration of organismal abundances from two prokaryote-enriched size fractions and four eukaryote size fractions. For prokaryotes, taxonomic profiling was performed using the 16S ribosomal gene fragments in Illumina-sequenced metagenomes and the SILVA database (Sunagawa et al. 2015) for referencing. For eukaryotes, OTUs were derived based on 18S rRNA gene V9 amplicons clustered with the Swarm version 2.1.1 (De Vargas et al. 2015) and taxonomically and functionally annotated by global pairwise alignment against an updated version (available at http://doi.org/10.5281/zenodo.3768951) of the PR2_V9 reference database. On the basis of these taxonomic affiliations, we classified all taxa into Marine Plankton

Groups as in (de Vargas et al. 2015). Local graphs were computed for each sampling station and related to environmental predictor variables. The graph inference was performed using FW v0.13.1 using default parameters (Chaffron et al. 2021).

The resulting integrated species association network [referred to as the Global Ocean Plankton Interactome (GPI)] counts a total of 20,810 nodes corresponding to operational taxonomic units (OTUs) and 86,026 edges corresponding to potential biotic interactions. In comparison to a previous plankton interactome generated from Tara Oceans data (Lima-Mendez et al. 2015), the GPI doubled the number of recovered known interactions from the literature. A vast majority of positive associations (98.5%) were predicted, likely underlying a prevalent role for biotic interactions in shaping marine plankton communities (Lima-Mendez et al. 2015). Using a deterministic eigenvector-based network community detection algorithm (Newman, 2006), five communities emerged from the GPI, which were enriched in OTUs assigned to specific biomes, and displayed distinct predicted biotic associations. Through comparison of community abundance profiles, these five communities were preferentially observed in specific biomes as defined by Longhurst's primary biome partitioning (Polar, Westerlies and Trades biomes) (Longhurst, 2010).

A list of all data products can be found in **Tables A 4.8.1, A 4.8.2, and A 4.8.3**.

4.9 Sneak Preview: Initial collection of mapped and extrapolated data products AtlantECO-MAPS1 (task 2.3; D2.2; lead: ETHZ/SBE)

The data recorded below is part of D2.2. ('AtlantECO-MAPS). It is recorded here in advance of the deadline of the deliverable in order to facilitate collaboration within and beyond the project. The submissions consist of geo-referenced fields of plankton biogeography, abundance, biomass, community composition and diversity, as derived from AtlantECO-BASE and other data, extrapolated to the global scale using species distribution modelling (Guisan and Zimmermann, 2000).

A list of all data products can be found in **Table A 4.9**.

4.10 Sneak preview: Data visualisation tools (web applications) for the visualisation of geo-referenced point and interpolated data (task 2.3; D2.2; lead: ETHZ/SU/UNIROMA1)

In order to process, interpret and visualise data, a range of web applications have been developed by the WP2 team. These tools include:

A ShinyApp for the visualisation of particulate organic carbon and carbon flux data:
https://fricour.shinyapps.io/carbon_fluxes_app/

A PlotlyApp to serve biodiversity data from species distribution modelling to policy makers:
https://mapmaker.ethz.ch/

A ShinyApp to assist species distribution modellers with a guided pipeline for the modelling of presence-absence data:
https://dandeangelis.shinyapps.io/MarEM_1_0/

A ShinyApp for the visualisation of present day and end of century (RCP8.5 climate change scenario) niche areas of eukaryote plankton MAGs from Fremont et al. (2022) and Delmont et al. (2022):
https://gigaplankton.shinyapps.io/TOENDB/

A prototype for the automatized modelling of quantitative (abundance and biomass) data is under development for deployment within the framework of BlueCloud (open science platform for the collaborative marine research): https://blue-cloud.org/, with a prototype designed and tested during the blue-cloud hackathon 2021 'Decode the Ocean' (https://hackathon.blue-cloud.org/).

4.11 Sneak preview: AtlantECO-CMIP6: Initial AtlantECO compilation of regridded CMIP6 model data for the extrapolation of biological data to different future climates (task 2.3; D2.2; lead: ETHZ/CEA/UBern)

An initial compilation of regridded future projections for selected CMIP6 models, variables and future scenarios can be found under: https://data.up.ethz.ch/shared/AtlantECO/AtlantECO-CMIP6/

Models: GFDL-ESM4, IPSL-CM6A-LR, MPI-ESM1-2-HR,
Scenarios: historical, piControl, SSP126, SSP245
Time period: 1850-2014 (historical), 2015 - 2100 (future projections)
Variables: intpn2, intpp, mlotst, no3, o2, ph, phyc, so, spco2, thetao, tos, zooc, zos
Grid: regular degree grid; 360 (lon)x180 (lat) x 12 (months), with variable-specific depth resolution (surface versus full water column)

For further information about variables, grids, simulation set-ups, please consult the CMIP6 website (https://wcrp-cmip.org/cmip-phase-6-cmip6/) as well as Eyring et al., (2016) for simulation design and naming conventions.

5 Conclusion

This first version of AtlantECO-BASE provides an impressive valuable knowledge base of the marine (micro)biome across taxonomic groups, data streams and all five data silos considered in this project. Data will be hosted at the AtlantECO GeoNode (https://atlanteco-geonode.eu/), and thus be made available to the project partners and the general public using FAIR publishing principles. While some of our historic target cruises could not be mined at present due to data protection matters, we compiled more than 21 million data points in a wide diversity of environments, which is far more than what was anticipated. Many scientific projects exploring and discovering patterns in these data sets have been started, both within and beyond the project, and results will feed into AtlantECO-MAPS, to be delivered during M36.

6 References

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W. and Huse, S.M., 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PloS one*, *4*(7), p.6372.

Benedetti, F., Vogt, M., Elizondo, U.H., Righetti, D., Zimmermann, N.E. and Gruber, N., 2021. Major restructuring of marine plankton assemblages under global warming. *Nature communications*, *12*(1), p.5226.

Benedetti, F., Wydler, J. and Vogt, M., 2023. Copepod functional traits and groups show divergent biogeographies in the global ocean. Journal of Biogeography, 50(1), pp.8-22.

Brandão, M.C., Benedetti, F., Martini, S., Soviadan, Y.D., Irisson, J.O., Romagnan, J.B., Elineau, A., Desnos, C., Jalabert, L., Freire, A.S. and Picheral, M., 2021. Macroscale patterns of oceanic zooplankton composition and size structure. Scientific Reports, 11(1), p.15714.

Buitenhuis, E.T., Vogt, M., Moriarty, R., Bednaršek, N., Doney, S.C., Leblanc, K., Le Quéré, C., Luo, Y.W., O'Brien, C., O'Brien, T. and Peloquin, J., 2013. MAREDAT: towards a world atlas of MARine Ecosystem DATa. *Earth System Science Data*, *5*(2), pp.227-239.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nature Methods 13. pp.581-583.

Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal 11. pp.2639-2643.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K. and Kirilovsky, A., 2018. A global ocean atlas of eukaryotic genes. Nature communications, 9(1), p.373.

Chaffron, S., Delage, E., Budinich, M., Vintache, D., Henry, N., Nef, C., Ardyna, M., Zayed, A.A., Junger, P.C., Galand, P.E. and Lovejoy, C., 2021. Environmental vulnerability of the global ocean epipelagic plankton community interactome. Science Advances, 7(35), p.eabg1921.

Chenier, N., 2021. Assessing global-scale biomass patterns of dinoflagellates using a data-driven approach, MSc thesis.

Clerc, C., Bopp, L., Benedetti, F., Vogt, M., and Aumont, O., 2023. Including filter-feeding gelatinous macrozooplankton in a global marine biogeochemical model: model–data comparison and impact on the ocean carbon cycle, Biogeosciences, 20, 869–895, https://doi.org/10.5194/bg-20-869-2023.

Delmont, T.O., Gaia, M., Hinsinger, D.D., Frémont, P., Vanni, C., Fernandez-Guerra, A., Eren, A.M., Kourlaiev, A., d'Agata, L., Clayssen, Q. and Villar, E., 2022a. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. Cell Genomics, 2(5), p.100123.

Delmont, T.O., Pierella Karlusich, J.J., Veseli, I., Fuessel, J., Eren, A.M., Foster, R.A., Bowler, C., Wincker, P. and Pelletier, E., 2022b. Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. The ISME Journal, 16(4), pp.927-936.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I. and Carmichael, M., 2015. Eukaryotic plankton diversity in the sunlit ocean. Science, 348(6237), p.1261605.

Drago, L., Panaïotis, T., Irisson, J.O., Babin, M., Biard, T., Carlotti, F., Coppola, L., Guidi, L., Hauss, H., Karp-Boss, L. and Lombard, F., 2022. Global Distribution of Zooplankton Biomass Estimated by In Situ Imaging and Machine Learning. *Frontiers in Marine Science*, *9*.

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J. and Taylor, K.E., 2016. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), pp.1937-1958.

Frémont, P., Gehlen, M., Vrac, M., Leconte, J., Delmont, T.O., Wincker, P., Iudicone, D. and Jaillon, O., 2022. Restructuring of plankton genomic biogeography in the surface ocean under climate change. Nature Climate Change, 12(4), pp.393-401.

Guidi, L., Jackson, G.A., Stemmann, L., Miquel, J.C., Picheral, M. and Gorsky, G., 2008. Relationship between particle size distribution and flux in the mesopelagic zone. Deep Sea Research Part I: Oceanographic Research Papers, 55(10), pp.1364-1374.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J. and Del Campo, J., 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic acids research*, *41*(1), pp.597-604.

Guisan, A. and Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological modelling, 135(2-3), pp.147-186.

Irisson, J.O., Ayata, S.D., Lindsay, D.J., Karp-Boss, L. and Stemmann, L., 2022. Machine learning for the study of plankton and marine snow from images. Annual review of marine science, 14, pp.277-301.

Jalabert, L., Picheral, M., Desnos, C., Elineau, A., 2022. ZooScan Protocol. protocols.io https://dx.doi.org/10.17504/protocols.io.yxmvmk8j9g3p/v1.

Kiko, R., Picheral, M., Antoine, D., Babin, M., Berline, L., Biard, T., Boss, E., Brandt, P., Carlotti, F., Christiansen, S. and Coppola, L., 2022. A global marine particle size distribution dataset obtained with the Underwater Vision Profiler 5. *Earth System Science Data*, *14*(9), pp.4315-4337.

Knecht, N., 2021. The impact of zooplankton calcifiers on the marine carbon cycle, MSc thesis.

Knecht, N.S., Benedetti, F., Hofmann, U., Chaabane, S., de Weerd, C., Peijnenburg, K., Schiebel, R. and Vogt, M.,The impact of zooplankton calcifiers on the marine carbon cycle. *Authorea Preprints*. DOI: 10.22541/essoar.167283650.05543210/v1.

Kukulka, T., Proskurowski, G., Morét-Ferguson, S., Meyer, D.W. and Law, K.L., 2012. The effect of wind mixing on the vertical distribution of buoyant plastic debris. Geophysical research letters, 39(7).

Leblanc, K., Arístegui, J., Armand, L., Assmy, P., Beker, B., Bode, A., Breton, E., Cornet, V., Gibson, J., Gosselin, M.P. and Kopczynska, E., 2012. A global diatom database–abundance, biovolume and biomass in the world ocean. Earth System Science Data, 4(1), pp.149-165.

Lee M. 2019. Happy belly bioinformatics: an open-source resource dedicated to helping biologists utilize bioinformatics. Journal of Open Source Education 2. p.53.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., Roux, S., Vincent, F. and Bittner, L., 2015. Determinants of community structure in the global plankton interactome. Science, 348(6237), p.1262073.

Lombard F., Bourdin G., Pesant S., Agostini S., Baudena A., Boissin E., Cassar N., Clampitt M., Conan P., Silva O.D., Dimier C., Douville E., Elineau A., Fin J., Flores J.M., Ghiglione J.F., Hume B.C.C., Jalabert L., John S.G., Kelly R.L., Koren I., Lin Y., Marie D., McMinds R., Mériguet Z., Metzl N., Paz-García D.A., Pedrotti M.L., Poulain J., Pujo-Pay M., Ras J., Reverdin G., Romac S., Rouan A., Röttinger E., Vardi A., Voolstra C.R., Moulin C., Iwankow G., Banaigs B., Bowler C., de Vargas C., Forcioli D., Furla P., Galand P.E., Gilson E., Reynaud S., Sunagawa S., Sullivan M.B., Thomas O., Troublé R., Thurber R.V., Wincker P., Zoccola D., Allemand D., Planes S., Boss E., Gorsky G., 2022. Open science resources from the Tara Pacific expedition across coral reef and surface ocean ecosystems. bioRxiv 2022.05.25.493210. doi: 10.1101/2022.05.25.493210.

Lombard F., 2022a. Planktoscope protocol for plankton imaging. protocols.io. doi: https://dx.doi.org/10.17504/protocols.io.bp2l6bq3zgqe/v1

Lombard F., Major W., Oddone A., Clausse C., 2023. Planktoscope protocol for plankton imaging. protocols.io. doi: https://dx.doi.org/10.17504/protocols.io.bp2l6bq3zgqe/v2

Longhurst, A.R., 2010. Ecological geography of the sea. Elsevier.

Manral, D., Iovino, D., Jaillon, O., Masina, S., Sarmento, H., Iudicone, D., Amaral-Zettler, L. and Sebille, E.V., Computing Marine Plankton Connectivity under Thermal Constraints. Frontiers in Marine Science, 10, p.39.

Martin-Cabrera P., Perez Perez R., Irisson J.-O., Lombard F., Möller K.O., Rühl S., Creach V., Lindh M., Stemmann L., Schepers L., 2022. Best practices and recommendations for plankton imaging data management: Ensuring effective data flow towards European data infrastructures. Version 1.

Mériguet Z., Oddone A., Le Guen D., Pollina T., Bazile R., Moulin C., Troublé R., Prakash M., de Vargas C., Lombard F., 2022. Basin-Scale Underway Quantitative Survey of Surface Microplankton Using Affordable Collection and Imaging Tools Deployed From Tara. Frontiers in Marine Science 9. doi: 10.3389/fmars.2022.916025

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G.,Crusoe M. R., Kale, V., Potter, S.C.,Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., Finn, R.D., MGnify: the microbiome analysis resource in 2020, Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D570–D578, https://doi.org/10.1093/nar/gkz1035.

Murali, A., Bhargava, A. and Wright, E.S., 2018. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, *6*(1), pp.1-14.

Newman, M.E., 2006. Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), pp.8577-8582.

O'Brien, C.J., Peloquin, J.A., Vogt, M., Heinle, M., Gruber, N., Ajani, P., Andruleit, H., Arístegui, J., Beaufort, L., Estrada, M. and Karentz, D., 2013. Global marine plankton functional type biomass distributions: coccolithophores. Earth System Science Data, 5(2), pp.259-276.

O'Brien, T.D., 2010. COPEPOD, a global plankton database: a review of the 2010 database contents, processing methods, and access interface.

Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples: primers for marine microbiome studies. Environmental Microbiology 18. pp.1403-1414.

Pesant S., Not F., Picheral M. Kandels-Lewis S., Le Bescot N., Gorsky G., Iudicone D., Karsenti E., Speich S., Troublé R., 2015. Open science resources for the discovery and analysis of Tara Oceans data. Scientific data 2:150023.

Picheral, M., Colin, S. and Irisson, J.O., 2017. EcoTaxa, a tool for the taxonomic classification of images.
Richardson, A.J., Walne, A.W., John, A.W.G., Jonas, T.D., Lindley, J.A., Sims, D.W., Stevens, D. and Witt, M., 2006. Using continuous plankton recorder data. *Progress in Oceanography*, *68*(1), pp.27-74.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F.O., 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, *41*(1), pp.590-596.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. 2011. Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12. p.38.

Righetti, D., Vogt, M., Gruber, N., Zimmermann, N.E., 2023. Mapping global marine biodiversity under sparse data conditions. bioRxiv, doi: https://doi.org/10.1101/2023.02.28.530497.

Righetti, D., Vogt, M., Zimmermann, N.E., Guiry, M.D. and Gruber, N., 2020. PhytoBase: a global synthesis of open-ocean phytoplankton occurrences. *Earth System Science Data*, *12*(2), pp.907-933

Righetti, D., Vogt, M., Gruber, N., Psomas, A. and Zimmermann, N.E., 2019. Global pattern of phytoplankton diversity driven by temperature and environmental variability. *Science Advances, 5*(5), p.eaau6253.

Schmiz, S. 2021. Comparing models and observations of the surface accumulation zone of floating plastic in the North Atlantic subtropical gyre. MSc thesis, doi:10.5281/zenodo.5338790.

Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., Wessner, M., Leconte, J., Mangot, J.F., Poulain, J., Labadie, K. and Logares, R., 2018. Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. Nature Communications, 9(1), p.310.

Soviadan, Y.D., Benedetti, F., Brandão, M.C., Ayata, S.D., Irisson, J.O., Jamet, J.L., Kiko, R., Lombard, F., Gnandi, K. and Stemmann, L., 2022. Patterns of mesozooplankton community composition and vertical fluxes in the global ocean. Progress in Oceanography, 200, p.102717.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A. and Cornejo-Castillo, F.M., 2015. Structure and function of the global ocean microbiome. Science, 348(6237), p.1261359.

Van Sebille, E., Wilcox, C., Lebreton, L., Maximenko, N., Hardesty, B.D., Van Franeker, J.A., Eriksen, M., Siegel, D., Galgani, F. and Law, K.L., 2015. A global inventory of small floating plastic debris. Environmental Research Letters, 10(12), p.124006.

Vorobev, A., Dupouy, M., Carradec, Q., Delmont, T.O., Annamalé, A., Wincker, P. and Pelletier, E., 2020. Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. Genome research, 30(4), pp.647-659.

Zurell, D., Franklin, J., König, C., Bouchet, P.J., Dormann, C.F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A. and Lahoz-Monfort, J.J., 2020. A standard protocol for reporting species distribution models. Ecography, 43(9), pp.1261-1277.

7 Appendix

Table A 4.3.1 Traditional presence-absence data sets gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields (AtlantECO-MAPS); M: readme; N: published on AtlantECO's GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request; c: confidential); y:yes, n:no, o:ongoing/preliminary results available, d:derived fields only.

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Data contact & Related reference/s	Data doi	Diss L
Phytoplankton species occurrences	PhytoBase (v1) (presence-absence; no.obs: 1,360,621)	AtlantECO-BASE-v1_microbiome_traditional_phytoplankton_species_occurrences_PhytoBasev1_20200313.csv **Monthly modelled phytoplankton species richness (from Righetti et al. 2019):** https://data.up.ethz.ch/shared/AtlantECO/MAPS/Phytoplankton_diversity_Righettietal.2019/ AtlantECO-MAPS-v1_microbiome_traditional_phytoplankton_richness_Righetti_et_al_2019_aau6253_data_file_s1.nc **Readme:** see corresponding journal articles and supplementary material	Presence- absence observations for 1704 species	y	d	y	n	n	Contact: Meike Vogt (meike.vogt@env.ethz.ch) Righetti et al., 2020; doi:10.5194/essd-12-907-2020 Righetti et al., 2019; doi: 10.1126/sciadv.aau6253	10.1594/PANGAEA.904397	p

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 862923. This output reflects only the author's view and the European Union cannot be held responsible for any use that may be made of the information contained therein.

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Data contact & Related reference/s	Data doi	Diss L
Zoo- plankton species occurrences	ZooBase (v1) (no. obs: 1,365,231)	AtlantECO-BASE-v1_microbiome_traditional_zooplankton_species_occurrences_ZooBasev1_20210714.csv **Species richness for 14 functional groups:** AtlantECO/MAPS/Groups_species_richness_Benedettietal.2021/ **Monthly species habitat suitabilities for phyto- and zooplankton species included in Benedetti et al. 2021:** https://data.up.ethz.ch/shared/AtlantECO/MAPS/Species_habitat_suitability_Benedettietal.2021/phytoplankton/ https://data.up.ethz.ch/shared/AtlantECO/MAPS/Species_habitat_suitability_Benedettietal.2021/zooplankton/ **Readme:** see corresponding journal article and supplementary material	Presence-absence observations for 3 371 species	y	d	y	n	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch) Benedetti et al., 2021; doi:10.1038/s41467-021-25385-x	10.5281/zenodo.5101518 10.5281/zenodo.5101349	p
Updated PhytoBase	PhytoBasev2 (no. obs: 5,167,282)	AtlantECO-BASE-v1_microbiome_traditional_phytoplankton_species_occurrences_PhytoBasev2_20220905.csv	Presence-absence observations for 1 691 species	y	n	n	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch)	-	u
Updated ZooBase	ZooBasev2 (no. obs: 16,662,867)	AtlantECO-BASE-v1_microbiome_traditional_zooplankton_species_occurrences_ZooBasev2_20220909.csv	Presence-absence observations for 4 072 species	y	n	n	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch)	-	u

Table A 4.3.2 Traditional abundance and biomass data sets gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields (AtlantECO-MAPS); M: readme; N: published on AtlantECO's GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request); y: yes; n: no; o: ongoing work/preliminary results available

Taxon/ Type	Description (name; no. obs)	Filename(s)	Vars	R	G	E	M	N	Data contact & Related reference/s	Data doi	Diss L
Amphipoda	Global compilation of field observations of abundances/biomass for Amphipods (no obs.: 1,449,666)	AtlantECO-BASE-v1_microbiome_traditional_Amphipoda_abund+biomass_20221220.csv **Carbon conversion tables per taxonomic unit:** AtlantECO-BASE-v1_microbiome_traditional_Amphipoda_ind_carbon_values_20220930.xlsx **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_traditional_Amphipoda_abund+biomass_20220930.nc	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	n	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch)		u
Appendicularia	Global compilation of field observations of abundances/biomass for Appendicularia (no obs.: 218,159)	AtlantECO-BASE-v1_microbiome_traditional_Appendicularia_abund+biomass_20221220.csv **Carbon conversion tables per taxonomic unit:** AtlantECO-BASE-v1_microbiome_traditional_Appendicularia_ind_carbon_values_20220930.xlsx **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_traditional_Appendicularia_abund+biomass_20220930.nc	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	n	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch)		u
Bacillariophyta	MAREDAT diatoms (LeBlanc et al. 2012) (reformatted; no obs: 91,704)	AtlantECO-BASE-v1_microbiome_traditional_Diatoms_abund+biomass_MAREDAT_Leblanc2012_20220722.csv **Gridded biomass data (AtlantECO-GRID; as originally provided in MAREDAT on pangaea.de):**	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	o	n	n	Contact: Meike Vogt (meike.vogt@env.ethz.ch) Leblanc et al., 2012; doi:10.5194/essd-4-149-2012	https://doi.pangaea.de/10.1594/PANGAEA.777384	p

Taxon/ Type	Description (name; no. obs)	Filename(s)	Vars	R	G	E	M	N	Data contact & Related reference/s	Data doi	Diss L
		AtlantECO-GRID-v1_microbiome_traditional_Diatoms_MarEDat20120716Diatoms.nc **Readme:** see corresponding journal article									
Chaetognatha	Global compilation of field observations of abundances/biomass for Chaetognatha (no obs.: 383,151)	AtlantECO-BASE-v1_microbiome_traditional_Chaetognatha_abund+biomass_20221220.csv **Carbon conversion tables per taxonomic unit:** AtlantECO-BASE-v1_microbiome_traditional_Chaetognatha_ind_carbon_values_20220930.xlsx **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_traditional_Chaetognatha_abund+biomass_20220930.nc	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	n	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch)		u
Coccolithophores	MAREDAT coccolithophores (O'Brien/Peloquin et al. 2012) (reformatted; no obs. 57,321)	AtlantECO-BASE-v1_microbiome_traditional_Coccolithophores_abund+biomass_MAREDAT_OBrien2013_20220722.csv **Gridded biomass data (AtlantECO-GRID; as originally provided in MAREDAT on pangaea.de):** AtlantECO-GRID-v1_microbiome_traditional_Coccolithophores_MarEDat20120620Coccolithophores.nc **Readme:** see corresponding journal article	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	o	n	n	Contact: Meike Vogt (meike.vogt@env.ethz.ch) O'Brien et al., 2013; doi:10.5194/essd-5-259-2013	https://doi.pangaea.de/10.1594/PANGAEA.785092	p
Copepoda	Global compilation of field observations of abundances/biomass for Copepoda (no obs.: 9,080,989)	AtlantECO-BASE-v1_microbiome_traditional_Copepoda_abund+biomass_20221220.csv **Carbon conversion per taxonomic unit:**	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	n	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch)		u

Taxon/ Type	Description (name; no. obs)	Filename(s)	Vars	R	G	E	M	N	Data contact & Related reference/s	Data doi	Diss L
		AtlantECO-BASE-v1_microbiome_traditional_Copepoda_ind_carbon_values_20220930.xlsx **Gridded data (AtlantECO/GRID):** AtlantECO-MAPS-v1_microbiome_traditional_Copepoda_abund+biomass_20220930.nc									
Diazotrophs	Global compilation of field observations of abundances/biomass for Diazotrophs (no obs.: 294,060)	AtlantECO-BASE-v1_microbiome_omics+imaging+traditional_Diazotrophs_pres-abs+abund+biomass_20230215.csv **Gridded data (AtlantECO/GRID):** AtlantECO/GRID/Diazotroph_quantitative_fields_gridded.2022/ Gridded fields contain Mean/Median/Stdev/Min and Max of the diazotroph abundances (trichome_counts_perVolume, cell_counts_perVolume and nifH_gene_counts_perVolume). netCDF file also contains one variable indicating the number of observations.	Abundances (trichomes.m-3 & m-2, cells..L-1 & cm-2 & m-2 & m-3, nifH.m-3 & L-1 & m-2) converted to biomass concentrations (mg.C.m-3 & mg.C.m-2)	y	y	o	y	n	Contact: Dominic Eriksson (dominic.eriksson@usys.ethz.ch)		u
Dinoflagellata	Global compilation of field observations of abundances/biomass for Dinoflagellata (no obs.: 4,339,695)	AtlantECO-BASE-v1_microbiome_traditional_Dinoflagellates_abund+biomass_20230203.csv **Gridded data (AtlantECO/GRID):** Work in progress **Readme:** None so far	Abundances (cells.m-3 & cells.L-1) converted to biomass concentrations (mgC.m-3 and µgC.L-1)	y	o	o	n	n	Contact: Meike Vogt (meike.vogt@env.ethz.ch) N. Chénier, in prep.		u
Euphausiacea	Global compilation of field observations of abundances/biomass for Euphausicea (no obs.: 1,071,450)	AtlantECO-BASE-v1_microbiome_traditional_Euphausiacea_abund+biomass_20221220.csv **Carbon conversion per taxonomic unit:** AtlantECO-BASE-v1_microbiome_traditional_Euphausiacea_ind_carbon_values_20220930.xlsx	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	n	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch)		u

Taxon/ Type	Description (name; no. obs)	Filename(s)	Vars	R	G	E	M	N	Data contact & Related reference/s	Data doi	Diss L
		Gridded data (AtlantECO/GRID): AtlantECO-GRID-v1_microbiome_traditional_Euphausiacea_abund+biomass_20220930.nc									
Foraminifera	Global compilation of field observations of abundances/biomass for Foraminifera (no obs.: 1,026,933)	AtlantECO-BASE-v1_microbiome_traditional_Foraminifera_abund+biomass_20221126.csv **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_traditional_Foraminifera_abund+biomass_20221126.nc **Extrapolated data:** https://data.up.ethz.ch/shared/AtlantECO/MAPS/Zooplankton_Calcifier_biomass_Knechtetal.2023/ AtlantECO-MAPS-v1_microbiome_traditional_Foraminifera_extrapolated_model_outputs_20221126.nc	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	y o	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch) Knecht et al., submitted; doi:10.2254/essoar.167283650.05543210/v1	https://doi.pangaea.de/10.1594/PANGAEA.957258	p
Jellyfish	Global compilation of field observations of abundances/biomass for Jellyfish (Cnidaria+Ctenophora; no obs.: 590,371)	AtlantECO-BASE-v1_microbiome_traditional_Jellyfish_abund+biomass_20221220.csv **Carbon conversion table per taxonomic group:** AtlantECO-BASE-v1_microbiome_traditional_Jellyfish_ind_carbon_values_20220930.xlsx **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_traditional_Jellyfish_abund+biomass_20220930.nc	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	n	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch)		u
Ostracoda	Global compilation of field observations of abundances/biomass	AtlantECO-BASE-v1_microbiome_traditional_Ostracoda_abund+biomass_20221220.csv	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	n	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch)		u

Taxon/ Type	Description (name; no. obs)	Filename(s)	Vars	R	G	E	M	N	Data contact & Related reference/s	Data doi	Diss L
	for Ostracoda (no obs.: 133,533)	**Carbon conversion tables per taxonomic group:** AtlantECO-BASE-v1_microbiome_traditional_Ostracoda_ind_carbon_values_20220930.xlsx **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_traditional_Ostracoda_abund+biomass_20220930.nc									
Pteropoda	Global compilation of field observations of abundances/biomass for Pteropoda (no obs.: 841,239)	AtlantECO-BASE-v1_microbiome_traditional_Pteropoda_abund+biomass_20220714.csv **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_traditional_Pteropoda_abund+biomass_20220714.nc **Extrapolated data:** https://data.up.ethz.ch/shared/AtlantECO/MAPS/Zooplankton_Calcifier_biomass_Knechtetal.2023/ AtlantECO-MAPS-v1_microbiome_traditional_Pteropoda_extrapolated_model_outputs_20220714.nc	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	y o	y	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch) Knecht et al., submitted; doi:10.2254/essoar.167283650.05543210/v1	https://doi.pangaea.de/10.1594/PANGAEA.957258	p
Thaliacea	Global compilation of field observations of abundances/biomass for Thaliacea (no obs.: 447,920)	AtlantECO-BASE-v1_microbiome_traditional_Thaliacea_abund+biomass_Clerc&al._20220922.csv **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_traditional_Thaliacea_abund+biomass_20220930.nc	Abundances (ind.m-3) converted to biomass concentrations (mgC.m-3)	y	y	o	y	n	Contact: Corentin Clerc (corentin.clerc@usys.ethz.ch; cclerc@lmd.ipsl.fr) Paper: Clerc et al., 2023; 10.5194/bg-20-869-2023	Clerc et al. 2023 (https://doi.org/10.5194/bg-20-869-2023) Data supplement: https://zenodo.org/record/7573432#.Y9ebhuzMJZ0	p

Table A 4.4 Optical imaging data sets gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields; M: readme; N: published on AtlantECO's GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request); y: yes; n: no; o: ongoing work/preliminary results available; d: derived fields only

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
Synthesis products:											
Vertical profiles of particles size and biovolume distribution	Particle size distribution data collected during multiple cruises globally with several regularly intercalibrated Underwater Vision Profilers, Version 5 (UVP5) (no. obs: 2,083,352 across 29 particles size classes)	AtlantECO-BASE-v1_microbiome_particles_imaging_abund+biovol_Kiko&al_2008_2010.csv AtlantECO-BASE-v1_microbiome_particles_imaging_abund+biovol_Kiko&al_2011_2013.csv AtlantECO-BASE-v1_microbiome_particles_imaging_abund+biovol_Kiko&al_2014_2016.csv AtlantECO-BASE-v1_microbiome_particles_imaging_abund+biovol_Kiko&al_2017_2019.csv AtlantECO-BASE-v1_microbiome_particles_imaging_abund+biovol_Kiko&al_2020_2022.csv	Particles concentrations and biovolume (mm3.l-1) across 29 size classes (ranging from 0.0403 mm to > 26 mm)	y	y	n	y	n	Contact person: Fabien Lombard (fabien.lombard@imev-mer.fr) ; Lionel Guidi (lionel.guidi@imev-mer.fr) Kiko et al., 2022; doi:10.5194/essd-14-4315-2022	Kiko et al. (2021) https://doi.pangaea.de/10.1594/PANGAEA.924375	p
Broad zooplankton clades and Copepod families	Zooscan Tara Oceans data: WP2, Bongo and Régent nets (no. obs = 15,908)	AtlantECO-BASE-v1_microbiome_imaging_zooplankton_abundance_Brandao.et.al.2021_20210726.csv	Abundances (concentrations) of large zooplankton groups and copepod families	y	n	n	n	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch) Brandao et al., 2021; doi:10.1038/s41598-021-94615-5	Brandãoet al. (2021) https://www.nature.com/articles/s41598-021-94615-5 doi:10.17632/nwvjwccgvh.1	p
Broad zooplankton clades and Copepod families	Zooscan Tara Oceans data: WP2, Multinet (no. obs = 5,415)	AtlantECO-BASE-v1_microbiome_imaging_protists+zooplankton_abundance+biomass_Soviadan.et.al.2022_20220509.csv	Abundances (concentrations) and biomass of zooplankton groups	y	n	n	n	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch) Soviadan et al., 2022;https://doi.org/10.1016/j.pocean.2021.102717.	Soviadan et al. (2022) https://doi.org/10.1016/j.pocean.2021.102717	p

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
Broad zooplankton clades (most precise taxonomic resolution allowed by the UVP5)	Brazilian UVP: casts performed at 37 sampling stations around the Vitória Trindade Seamount Chain (Brazil), during "Ilhas" cruise.	AtlantECO-BASE-v1_microbiome_imaging_UVP5_Brazil_zooplankton_abundance+biomass_20230102.xlsx AtlantECO-BASE-v1_microbiome_imaging_UVP5_Brazil_zooplankton_biovolume+biomass_20230102.xlsx	Abundance, biovolume and zooplankton biomass, images obtained in the casts can be accessed in EcoTaxa platform (Project "Uvp5_sn200_ilhas_2017_filtered_vignettes").	y	y	n	y	n	Contact: Gleice de Souza Santos (gleicesantos@usp.br), Rubens Mendes Lopes (rubens@usp.br)		u
Images (ECOTAXA):											
Tara Oceans	Tara confocal microscopy (e-HFCM), Flowcam, Zooscan and UVP data; Ecotaxa	images recorded in Ecotaxa: https://ecotaxa.obs-vlfr.fr/prj/3365 (eHFCM >5µm 336,655 images) https://ecotaxa.obs-vlfr.fr/prj/2274 (eHFCM >20µm 1,235,640 images) https://ecotaxa.obs-vlfr.fr/prj/326 https://ecotaxa.obs-vlfr.fr/prj/327 https://ecotaxa.obs-vlfr.fr/prj/328 https://ecotaxa.obs-vlfr.fr/prj/329 https://ecotaxa.obs-vlfr.fr/prj/330 https://ecotaxa.obs-vlfr.fr/prj/331 (Flowcam 744,409 images) https://ecotaxa.obs-vlfr.fr/prj/378 https://ecotaxa.obs-vlfr.fr/prj/377 (Zooscan - WP2 net 399,204 images) https://ecotaxa.obs-vlfr.fr/prj/395 https://ecotaxa.obs-vlfr.fr/prj/398 https://ecotaxa.obs-vlfr.fr/prj/397 (Zooscan - Bongo net 326,896 images) https://ecotaxa.obs-vlfr.fr/prj/412 (Zooscan - Regent net 126,389 images) https://ecotaxa.obs-vlfr.fr/prj/579	images with taxonomic identification and morphological measurements	y	n	n	n	n	Contact person: Fabien Lombard (fabien.lombard@imev-mer.fr)		

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
		(UVP 776,497 images)									
Tara Pacific	Tara Pacific Flowcam and Zooscan data; Ecotaxa; Ongoing work of analysis (displayed here by legs, only images validated by taxonomic expert visible)	images recorded in Ecotaxa: https://ecotaxa.obs-vlfr.fr/prj/211 https://ecotaxa.obs-vlfr.fr/prj/1736 https://ecotaxa.obs-vlfr.fr/prj/1743 https://ecotaxa.obs-vlfr.fr/prj/1744 https://ecotaxa.obs-vlfr.fr/prj/1737 https://ecotaxa.obs-vlfr.fr/prj/1745 https://ecotaxa.obs-vlfr.fr/prj/1738 https://ecotaxa.obs-vlfr.fr/prj/5795 https://ecotaxa.obs-vlfr.fr/prj/5350 https://ecotaxa.obs-vlfr.fr/prj/4612 https://ecotaxa.obs-vlfr.fr/prj/4611 https://ecotaxa.obs-vlfr.fr/prj/4896 https://ecotaxa.obs-vlfr.fr/prj/4897 https://ecotaxa.obs-vlfr.fr/prj/4159 https://ecotaxa.obs-vlfr.fr/prj/4160 https://ecotaxa.obs-vlfr.fr/prj/2857 https://ecotaxa.obs-vlfr.fr/prj/2892 https://ecotaxa.obs-vlfr.fr/prj/4889 https://ecotaxa.obs-vlfr.fr/prj/4894 https://ecotaxa.obs-vlfr.fr/prj/4892 https://ecotaxa.obs-vlfr.fr/prj/4895 (Flowcam 2,528,719 images) https://ecotaxa.obs-vlfr.fr/prj/1345 https://ecotaxa.obs-vlfr.fr/prj/1344 https://ecotaxa.obs-vlfr.fr/prj/6825 https://ecotaxa.obs-vlfr.fr/prj/6824 https://ecotaxa.obs-vlfr.fr/prj/7051 https://ecotaxa.obs-vlfr.fr/prj/3274 https://ecotaxa.obs-vlfr.fr/prj/4299 https://ecotaxa.obs-vlfr.fr/prj/2141 https://ecotaxa.obs-vlfr.fr/prj/7046 https://ecotaxa.obs-vlfr.fr/prj/6044 https://ecotaxa.obs-vlfr.fr/prj/5930 (Zooscan 375,732 images)	images with taxonomic identification and morphological measurements	r	n	n	n	n	Contact person: Fabien Lombard (fabien.lombard@imev-mer.fr)		
Mission Microbiomes	Tara Microbiome Flowcam/Planktoscop	images recorded on Ecotaxa:	images with taxonomic identification and	r	n	n	n	n	Contact person: Fabien Lombard		

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
	e/P2/net/UVP data; Ecotaxa Ongoing work of analysis (displayed here by legs, only images validated by taxonomic expert visible)	https://ecotaxa.obs-vlfr.fr/prj/3891 (Flowcam 88,465 images) https://ecotaxa.obs-vlfr.fr/prj/3892 (Flowcam 66,243 images) https://ecotaxa.obs-vlfr.fr/prj/4343 (Planktoscope 115,119 images) https://ecotaxa.obs-vlfr.fr/prj/4356 (P2/miniHSN 98,260 images) https://ecotaxa.obs-vlfr.fr/prj/4362 (P2/miniHSN 13,176 images) https://ecotaxa.obs-vlfr.fr/prj/4363 (P2/deckNet 17,203 images) https://ecotaxa.obs-vlfr.fr/prj/5254 (UVP 298,478 images) https://ecotaxa.obs-vlfr.fr/prj/4559 (Flowcam Net 20 518,671 images) https://ecotaxa.obs-vlfr.fr/prj/6058 (Flowcam Net 20 43,053 images)	morphological measurements						(fabien.lombard@imev-mer.fr)		
JERICO	UVP data	images recorded on Ecotaxa https://ecotaxa.obs-vlfr.fr/prj/578 (74,769 images)		y	n	n	n	n	Contact person: Lars Stemmann (lars.stemmann@imev-mer.fr)		

Table A 4.5 Genetic data sets gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields; M: readme; N: published on AtlantECO's GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request); y: yes; n: no; o: ongoing work/preliminary results available, d: derived fields only.

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
MGnify Superstudy	AtlantECO Superstudy, Metagenomes (no. obs = 835)	Online source: https://www.ebi.ac.uk/metagenomics/super-studies/atlanteco	Metagenomes globally distributed, covering different plankton size fractions and water column depths	y	n	n	y	n	Contact person: Hugo Sarmento (hsarmento@ufscar.br) Paula Huber (mariapaulahuber@gmail.com)		p
Metagenomes: Functional diversity - planktonic communities 0.2-3 µm size fraction	AtlantECO Metagenomic database (no. obs = 451)	"AtlantECO-BASE-v1_microbiome_omics_MetaG-Pk_FunctionalDiversity_02-3um_202211121.csv" "[README]_AtlantECO-BASE-v1_microbiome_genomic_MetaG-Pk_FunctionalDiversity_02-3um_202211121.txt" **Gridded data (AtlantECO-GRID):** AtlantECO-GRID-v1_microbiome_omics_MetaG_FunctionalDiversity_02-3um_202211121.nc	Functional diversity based on KEGG orthologs (KOs)	y	y	n	y	o	Contact person: Hugo Sarmento (hsarmento@ufscar.br) Paula Huber (mariapaulahuber@gmail.com)		u
Metagenomes:Nitrogen metabolisms - planktonic communities 0.2-3 µm size fraction	AtlantECO Metagenomic database (no. obs = 451)	AtlantECO-BASE-v1_microbiome_omics_MetaG-Pk_NitrogenPath_KO_Abundance_20230220.txt [README]AtlantECO-BASE-v1_microbiome_omics_MetaG-Pk_NitrogenPath_KO_Abundance_20230220.txt	Counts of 18 genes based on KEGG orthologs (KOs) representative of Nitrogen metabolism	y	o	n	y	n	Contact person: Hugo Sarmento (hsarmento@ufscar.br) Paula Huber (mariapaulahuber@gmail.com)		u
Amplicon 16S rRNA Database	AtlantECO Amplicon database (no. obs = 428)	AtlantECO-BASE-v1_microbiome_omics_AmpSeq16S-Pk_UFSCar_20221201.csv	Tara Oceans and Malaspina 16S Amplicon database (counts of ASVs per sampling stations).	y	n	n	y	n	Contact person: Hugo Sarmento (hsarmento@ufscar.br)		u

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
		[README] AtlantECO-BASE-v1_microbiome_omics_AmpSeq16S-Pk_UFSCar_20221201.txt							Clara Arboleda-Baena (claraarboledab@gmail.com)		
Amplicon 18S rRNA V9 region Database	AtlantECO Amplicon database (no. obs = 1405)	AtlantECO-BASE-v1_microbiome_omics_AmpSeq18S_UFSCar_20230201.csv [README] AtlantECO-BASE-v1_microbiome_omics_AmpSeq18S_UFSCar_20230201.txt	Tara Oceans and Malaspina 18S V9 region Amplicon databases (counts if ASVs per sampling station).	y	n	n	y	n	Contact person: Hugo Sarmento (hsarmento@ufscar.br) Clara Arboleda-Baena (claraarboledab@gmail.com)		u
Ocean Prokaryotic Diversity	AtlantECO Amplicon 16S database (no. obs = 428)	AtlantECO-BASE-v1_microbiome_omics_AmpSeq16S-Pk_ShannonDiversityProkaryotes_20230208.csv [README]AtlantECO-BASE-v1_microbiome_omics_AmpSeq16S-Pk_ShannonDiversityProkaryotes_20230208.txt **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_omics_AmpSeq16S-Pk_ShannonDiversityProkaryotes_20230208.nc	Shannon diversity index of Prokaryotes based on 16S rRNA gene amplicon sequencing	y	y	n	y	n	Contact person: Hugo Sarmento (hsarmento@ufscar.br) Clara Arboleda-Baena (claraarboledab@gmail.com)		u
Ocean Autotrophic Prokaryotes Abundance	AtlantECO Amplicon 16S database (no. obs = 428)	AtlantECO-BASE-v1_microbiome_omics_AmpSeq16S-Pk_AutotrophicProkaryotesAbundance_20230208.csv [README]AtlantECO-BASE-v1_microbiome_omics_AmpSeq16S-Pk_AutotrophicProkaryotesAbundance_20230208.txt **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_omics_AmpSeq16S-Pk_AutotrophicProkaryotesAbundance_20230208.nc	Autotrophic Prokaryotes Abundance based on 16S rRNA gene amplicon sequencing	y	y	n	y	n	Contact person: Hugo Sarmento (hsarmento@ufscar.br) Clara Arboleda-Baena (claraarboledab@gmail.com)		u

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
Ocean Heterotrophic Prokaryotes Abundance	AtlantECO Amplicon 16S database (no. obs = 428)	AtlantECO-BASE-v1_microbiome_omics_AmpSeq16S-Pk_HeterotrophicProkaryotesAbundance_20230208.csv [README]AtlantECO-BASE-v1_microbiome_omics_AmpSeq16S-Pk_HeterotrophicProkaryotesAbundance_20230208.txt **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_omics_AmpSeq16S-Pk_HeterotrophicProkaryotesAbundance_20230208.nc	Heterotrophic Prokaryotes Abundance based on 16S rRNA gene amplicon sequencing	y	y	n	y	n	Contact person: Hugo Sarmento (hsarmento@ufscar.br) Clara Arboleda-Baena (claraarboledab@gmail.com)		u
Ocean Eukaryotic Diversity	AtlantECO Amplicon 18S V9 database (no. obs = 1405)	AtlantECO-BASE-v1_microbiome_omics_AmpSeq18S-Pk_ShannonDiversityEukaryotes_20230208.csv [README]AtlantECO-BASE-v1_microbiome_omics_AmpSeq18S-Pk_ShannonDiversityEukaryotes_20230208.txt **Gridded data (AtlantECO/GRID):** AtlantECO-GRID-v1_microbiome_omics_AmpSeq18S-Pk_ShannonDiversityEukaryotes_20230208.nc	Shannon diversity index of Eukaryotes based on 18S rRNA gene V9 region amplicon sequencing	y	y	n	y	n	Contact person: Hugo Sarmento (hsarmento@ufscar.br) Clara Arboleda-Baena (claraarboledab@gmail.com)		u
Ocean Eukaryotic Abundance	AtlantECO Amplicon 18S V9 database (no. obs = 1405)	AtlantECO-BASE-v1_microbiome_omics_AmpSeq18S-Pk_EukaryoticAbundance_20230208.csv [README]AtlantECO-BASE-v1_microbiome_omics_AmpSeq18S-Pk_EukaryoticAbundance_20230208.txt **Gridded data (AtlantECO-GRID):** AtlantECO-GRID-v1_microbiome_omics_AmpSeq18S-Pk_EukaryoticAbundance_20230208.nc	Abundance of Eukaryotes based on 18S rRNA gene V9 region amplicon sequencing	y	y	n	y	n	Contact person: Hugo Sarmento (hsarmento@ufscar.br) Clara Arboleda-Baena (claraarboledab@gmail.com)		u

Table A 4.5.1 Plastisphere data sets with data-type specific formatting gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields; M: readme; N: published on AtlantECO's GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request); y: yes; n: no; o: ongoing work/preliminary results available, d: derived fields only.

Taxon/ Type	Description (no. obs) and links to public data repository	Filename(s) and formats	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
Plastisphere Amplicon 16S rRNA Database	AtlantECO Amplicon database (no. obs = 357)	AtlantECO-Base-v1_microbiome_genomic_GenDB-AmpSeq16S-Plastisphere-database_V1.0.csv	Plastisphere Amplicon 16S rRNA Database; georeference database + sapling metadata & env. variables	y	o	n	o	n	Contac person: Linda Amaral–Zettler (linda.amaral-zettler@nioz.nl)	https://doi.org/10.1890/150017	p

Table A 4.6 Carbon flux data sets gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields; M: readme; N: published on AtlantECO,s GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request); y: yes; n: no; o: ongoing work/preliminary results available

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
POC fluxes based on Thorium-234 deficits and sediment traps	Carbon flux rate measured by sediment traps and Thorium (Th-234) disequilibria. Data depth range: 5 - 5847 m. 15,425 obs.	AtlantECO-BASE-v1_carbon-flux_Sediment_traps_Thorium234_20220912.csv	MgC.m-2.day-1	y	y	n	y	n	Contact person: Florian Ricour (Florian.Ricour@uliege.be) Ricour et al, PhD thesis		u
Particle size distribution based on UVP imaging data	2'083'352 obs. taken from depth range of 2.5 to 6017.5 m.	AtlantECO-BASE-v1_UVP5_size_spectra_20220912.csv	27 size ranges (from ESD: 0.0508-0.064mm to 20.6-26 mm) of particle concentration fractions	y	n	n	y	n	Contact person: Florian Ricour (Florian.Ricour@uliege.be) Ricour et al, PhD thesis	https://doi.pangaea.de/10.1594/PANGAEA.924375	u
Carbon fluxes based on particle size distribution based on UVP imaging data derived from Guidi et al., 2008	Information (number of observations per grid cell) are provided in the NetCDF	AtlantECO-BASE-v1_carbon-flux_UVP5_20230301.nc	MgC.m-2.day-1	n	y	n	n	n	Contact person: Lionel Guidi and Florian Ricour (lionel.guidi@imev-mer.fr, florian.ricour@uliege.be)		u

Table A 4.7 Microplastics data sets gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields; M: readme; N: published on AtlantECO,s GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request); y: yes; n: no; o: ongoing work/preliminary results available, d: derived fields only.

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	DissL
Microplastics	13,159 obs.	AtlantECO_BASE-v1_microplastic_VanSebille&al_20221124.csv	Concentration of floating plastics	y	y	n	y	n	Contact: Erik Van Sebille (e.vansebille@uu.nl) Schmiz et al. MSc thesis		u

Table A 4.8.1 AtlantECO-ELSE: Connectivity data sets with data-type specific formatting gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields; M: readme; N: published on AtlantECO's GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request); y: yes; n: no; o: ongoing work/preliminary results available, d: derived fields only.

Taxon/ Type	Description (no. obs) and links to public data repository	Filename(s) and formats	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
Connectivity matrices	This data submission contains Lagrangian connectivity analysis of the Atlantic Ocean. Numerical 2D simulations with 375,570 particles across the Atlantic Ocean at different depths (surface, 50, 100, 250, 500 m) were performed using GLOB16 data from CMCC to compute connectivity matrices between all the grid cells. The main results are obtained using Resolution 3 of Uber H3 library for spatial gridding. Different statistics of temperatures corresponding to the connected grid cells were also exported in matrices (.npz) similar to the connectivity matrices: • Minimum of minimum/	Example file: Annual_Binary_DomainAdjacency_z0_csr.npz Example file for temperature: Annual_min_MinTemperature_z0_csr.npz in *.npz format (compressed Python NumPy array archive) The source and destination grid cells (rows-columns in the matrices) can be geo-referenced using the H3_Res3_MasterHexList.npz file corresponding to Resolution 3 hexagon-IDs of Uber H3 library. https://science.public.data.uu.nl/vault-oceanparcels/atlantic_plankton_connectivity_thermal_constraints%5B1673872194%5D/original/H3_Res3_MasterHexList.npz	Connectivities (0/1) Minimum/Maximum Temperatures (°C)	y	y	n	y	n	Contact: Darshika Manral (d.manral@uu.nl) cc: Erik van Sebille (e.vansebille@uu.nl) Related papers: Manral et al. 2022 https://www.frontiersin.org/articles/10.3389/fmars.2023.1066050/full https://doi.org/10.3389/fmars.2023.1066050	https://doi.org/10.24416/UU01-HVXLB0	p

Taxon/ Type	Description (no. obs) and links to public data repository	Filename(s) and formats	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
	maximum temperatures • Maximum of minimum/ maximum temperatures • Average of minimum/ maximum temperatures Link to the data repository: https://public.yoda.uu.nl/science/UU01/HVXLB0.html										

Table A 4.8.2 AtlantECO-ELSE: Omics data sets with data-type specific formatting gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields; M: readme; N: published on AtlantECO's GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request); y: yes; n: no; o: ongoing work/preliminary results available, d: derived fields only.

Taxon/ Type	Description (no. obs) and links to public data repository	Filename(s) and formats	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
Eukaryotic MAGs from Delmont et al (2022a)	This data submission consists of the analysis of eukaryotic Tara Oceans (meta)genomes, including: 11raw metagenomic co-assemblies, the FASTA files for 713 MAGs and SAGs, the $10 million protein-coding sequences (nucleotides, amino acids and gff format), and the curated DNA-dependent RNA polymerase genes (MAGs and SAGs and METdb transcrip-tomes) Link to resources: https://www.genoscope.cns.fr/tara/ Data subheader: Tara Oceans Eukaryotic Genomes (the "SMAGs")	multiple zip archives containing files with multiple file formats	**SMAGs:** curated SMAGs and single cell genomes metagenomic co-assemblies curated DNA-dependent RNA polymerase	n	n	y	y	n	Contact: Olivier Jaillon (ojaillon@genoscope.fr) Related papers: Delmont et al. 2022a https://www.cell.com/cell-genomics/pdf/S2666-979X(22)00047-7.pdf	n	p

Taxon/ Type	Description (no. obs) and links to public data repository	Filename(s) and formats	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
Bacterial MAGs including 48 Diazotroph MAGs from Delmont et al., 2022b	This data submission consists of resources related to the analysis of Tara Oceans (meta)genomic information, including metagenomic co-assemblies, FASTA files, and a diazotroph genomic database. Link to resources: https://www.genoscope.cns.fr/tara/ Data subheader: Tara Oceans Bacterial and Archaeal Genomes (the "BAC_ARC_MAGs") Diazotrophs: https://figshare.com/articles/dataset/Marine_diazotrophs/14248283	multiple zip archives containing files with multiple file formats	**BAC_ARC_MAGs:** 1888 curated MAGs including 48 curated diazotroph MAGs	n	n	n	y	n	Contact: Olivier Jaillon (ojaillon@genoscope.fr) Related papers: Delmont et al. 2022b https://www.nature.com/articles/s41396-021-01135-1 https://figshare.com/articles/dataset/Marine_diazotrophs/14248283	n	P
Further collection of Genoscope Tara resources	A large collection of genomic resources based on the analysis of Tara Oceans data: https://www.genoscope.cns.fr/tara/	multiple zip archives containing files with multiple file formats	**MGT: Metagenome-based Transcriptome** **MATOU:** unigene sequences	n	n	n	n	n	Contact: Olivier Jaillon (ojaillon@genoscope.fr) Related papers:	n	P

Taxon/ Type	Description (no. obs) and links to public data repository	Filename(s) and formats	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
	Data subheaders: MetaGenomic Transcriptomes (MGT) Tara Oceans Eukaryote Gene Catalog (the "MATOU") Single-cell genomes		taxonomic affiliations protein domain identification metagenomic occurrences metatranscriptomic occurrences unigenes cluster compositions and properties **Single-cell genomics:** MAST-3 and MAST-4 clades Chrysophyte clades						MGT: Vorobev et al. 2020 MATOU: Carradec et al. 2020 Single-cell genomes: Seeleuthner et al. 2018		

Table A 4.8.3 AtlantECO-ELSE: Interactome data sets with data-type specific formatting gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields; M: readme; N: published on AtlantECO's GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request); y: yes; n: no; o: ongoing work/preliminary results available, d: derived fields only.

Taxon/ Type	Description (no. obs) and links to public data repository	Filename(s) and formats	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
Plankton interactome from Chaffron et al., 2021	This data submission contains: - OTUs abundance matrices used for inferring the GPI: https://uncloud.univ-nantes.fr/index.php/s/yePQAsTtxeFW7Sw - The GPI interactome as a graphML file: https://uncloud.univ-nantes.fr/index.php/s/G6ybiG8pHsikYiJ - The GPI interactome, merged at the OTU level, as a graphML file: https://uncloud.univ-nantes.fr/index.php/s/tSL3Xks6yRPZ5o4	zip archives with abundance matrices and graphML files Example file name: GPI.graphml.zip	OTU abundance matrices, GPI interactome graphs (vertices and edges)	n	n	n	n	n	Contact: Samuel Chaffron (samuel.chaffron@univ-nantes.fr) Related papers: Chaffron et al. 2021 https://www.science.org/doi/10.1126/sciadv.abg1921 (see Supplementary materials)	https://doi.org/10.1126/sciadv.abg1921_SOM	p

Table A 4.9 AtlantECO-MAPS: Extrapolated data sets gathered:

Vars: variables included; R: raw data; G: gridded data; E: extrapolated fields; M: readme; N: published on AtlantECO's GeoNode (status: 01-23); DissL: Dissemination Level (p: public, u: upon request); y: yes; n: no; o: ongoing work/preliminary results available, d: derived fields only.

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
Phyto- and zooplankton species	Monthly Median/Min/Max/ Stdev of habitat suitability indices (HSI) for 341 phytoplankton species and 504 zooplankton species; time period covered: 2012-2031	Files in https://data.up.ethz.ch/shared/AtlantECO/MAPS/Species_habitat_suitability_Benedettietal.2021/	Habitat suitability indices (HSI, aka presence probability) - Contemporary conditions only	n	n	y	n	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch) Benedetti et al., 2021; doi:10.1038/s41467-021-25385-x	Benedetti et al. (2021) https://www.nature.com/articles/s41467-021-25385-x	p
Phyto- and zooplankton taxonomic groups	Monthly Median/Min/Max/ Stdev of Species Richness (SR) based on habitat suitability indices (HSI) for 3 phytoplankton groups and 11 zooplankton groups; time periods covered: contemporary (2012-2031) and future (2081-2100)	Files in https://data.up.ethz.ch/shared/AtlantECO/MAPS/Groups_species_richness_Benedettietal.2021/ (28 files) One netCDF per plankton group and time period (contemporary and future), for 14 different groups: Amphipoda, Appendicularia, Calanoida, Chaetognatha, Coccolithophores, Diatoms, Dinoflagellates, Euphausiids, Foraminifera, Jellyfish, Oithonida, Poecilostomatoida, Pteropoda and Thaliacea	Habitat suitability indices (HSI, aka presence probability) - Contemporary and future (RCP8.5) conditions	n	n	y	n	n	Contact: Fabio Benedetti (fabio.benedetti@usys.ethz.ch) Benedetti et al., 2021; doi:10.1038/s41467-021-25385-x	Benedetti et al. (2021) https://www.nature.com/articles/s41467-021-25385-x	u
Phyto- & Zooplankton (monthly) from the MAPMAKER project	Monthly Mean/Min/Max/ Stdev of habitat suitability indices (HSI) and presence absence data for 3 phytoplankton groups and 11 zooplankton functional groups; derived biodiversity patterns for time span of 2012 - 2100	Files in https://data.up.ethz.ch/shared/AtlantECO/MAPS/Species_habitat_suitability_MAPMAKER_Eriksonetal_2022/	Dimensions: RCPs (rcp26, rcp45, rcp85), stat(mean, std, min, max), time, lat and lon. Data variables:859 (species_names)	n	y	y	n	n	Contact: Dominic Eriksson (dominic.eriksson@usys.ethz.ch) X. Li, MSc thesis, 2023		u
Zooplankton calcifier biomass	Monthly median/mean/min/m	Files in	Extrapolated biomass estimatesfor pteropods and	y	y	y	y	n	Contact: Fabio Benedetti	Knecht et al.,	u

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
	ax/stdev pteropod and foraminifera biomass	https://data.up.ethz.ch/shared/AtlantECO/MAPS/Zooplankton_Calcifier_biomass_Knechtetal.2023/	foraminifera, monthly climatology, 1x1degree						(fabio.benedetti@usys.ethz.ch) Knecht et al., submitted; doi:10.2254/essoar.167283650.05543210/v1		
Eukaryotic Metagenomes Assembled Genomes (MAGs) from Delmont et al., 2022	Modelled annual probability of presence for individual 366 MAGs and 8 SAGs for present (2006-15) and end of the century (2090-99) following RCP8.5 greenhouse gas emission scenario	Files in https://data.up.ethz.ch/shared/AtlantECO/MAPS/Biogeographic_maps_Fremontetal.2022/MAGs_Delmontetal.2022/ AtlantECO-MAPS-v1_*_Delmontetal.2022_20221117.nc	presence probabilities MAGs (0-1), see readme	n	y	y	y	n	Contact person: Paul Frémont/Olivier Jaillon (pfremont@genoscope.cns.fr; ojaillon@genoscope.cns.fr) Frémont et al., 2022; doi:10.1038/s41558-022-01314-8 Delmont et al., 2022; doi:10.1016/j.xgen.2022.100123	Frémont et al. (2022) https://www.nature.com/articles/s41558-022-01314-8	u
Bray-Curtis Dissimilarity from Frémont et al., 2022	Modelled annual Bray curtis dissimilarity index between present day and end of the century plankton biogeographies following RCP8.5 greenhouse gas emission scenario	Files in https://data.up.ethz.ch/shared/AtlantECO/MAPS/Biogeographic_maps_Fremontetal.2022/Bray-curtis_dissimilarity_index_Fremontetal.2022/ AtlantECO-MAPS-v1_bray_curtis_dissimilarity_plankton_community_composition_*_Fremontetal.2022_20221117.nc	Bray-curtis dissimilarity index (0-1), see readme	n	y	y	y	n	Contact person: Paul Frémont/Olivier Jaillon (pfremont@genoscope.cns.fr; ojaillon@genoscope.cns.fr) Frémont et al., 2022; doi:10.1038/s41558-022-01314-8	Frémont et al. (2022) https://www.nature.com/articles/s41558-022-01314-8	p
Climato-genomic provinces from Frémont et al., 2022	Annual modelled probabilities of presence of each "climato-genomic" provinces (Figure 3 and extended data Figure 4 of the article) for	Files in https://data.up.ethz.ch/shared/AtlantECO/MAPS/Biogeographic_maps_Fremontetal.2022/Climato-genomic_provinces_Fremontetal.2022/	Size fractions and provinces probabilities of presence (0-1), see readme	n	y	y	y	n	Contact person: Paul Frémont/Olivier Jaillon (pfremont@genoscope.cns.fr; ojaillon@genoscope.cns.fr)	Frémont et al. (2022) https://www.nature.com/articles/s41558-022-01314-8	p

Taxon/ Type	Description (no. obs)	Filename(s)	Vars	R	G	E	M	N	Contact person & Related reference/s	Data doi	Diss L
	modelled present day (2006-15) and end of the century (2090-99) following RCP8.5 greenhouse gas emission scenario.	AtlantECO-MAPS-v1_climato-genomic_province_*_Fremontetal.2022_20221117.nc							Frémont et al., 2022; doi:10.1038/s41558-022-01314-8		
Mean annual biomass of large zooplankton groups derived from UVP5 data (Drago et al., 2022)	Mean predictions of annual biomass derived from Boosted Regression Trees (BRT) trained on mean annual biomass concentrations of zooplankton groups derived from global UVP5 profiles, integrated over two depth ranges: 0-200m and 200-500m.	20 files in https://data.up.ethz.ch/shared/AtlantECO/MAPS/Zooplankton_groups_biomass_UVP5_Drago_et_al._2022/ One netCDF file per large zooplanton group: Acantharea, Annelida, Appendicularia, Chaetognatha, Cnidaria (others), Collodaria, Collodaria (colonial), Copepoda, Crustacea (others), Ctenophora, Doliolida, Eumalacostraca, Foraminifera, Hydrozoa (others), Limacinidae, Mollusca (others), Ostracoda, Phaeodaria, Rhizaria (others), Siphonophorae.	Mean predicted biomass (mgC/m3) + standard deviation + standard error of the predictions	n	y	y	n	n	Contact person: Laetitia Drago (laetitia.drago@imev-mer.fr) Drago et al., 2022; doi:10.3389/fmars.2022.894372 Raw data: Stemmann et al., in prep.	Drago et al., (2022) https://doi.org/10.3389/fmars.2022.894372	p