**TITLE: AtlantECO [WP2] – Traditional microscopy dataset –** **Copepoda abundance and biomass concentration data**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**1.- INTRODUCTION**

This dataset contains **9 080 989** georeferenced abundance and biomass concentration records of **861** accepted scientific names of Copepoda of various taxonomic levels. This dataset is a compilation of the following seven main global and regional datasets that reported abundances of marine planktonic copepods:

* The Coastal & Oceanic Plankton Ecology, Production & Observation Database (NMFS-COPEPOD, O’Brien, 2014) from the National Oceanic and Atmospheric Administration - <https://www.st.nmfs.noaa.gov/copepod/atlas/html/taxatlas_4212000.html>
* The North Atlantic and North Pacific Continuous Plankton recorder (NANP-CPR) survey (Johns D & Broughton D, 2019) - <https://doi.org/10.17031/1629>
* The Southern Ocean CPR (SO-CPR) survey (Hosie, 2021) - doi:10.26179/ksds-s610
* The Australian CPR (AusCPR) survey (AusCPR) - <https://catalogue-imos.aodn.org.au/geonetwork/srv/eng/catalog.search#/metadata/c1344e70-480e-0993-e044-00144f7bc0f4>
* The copepod observations compiled by Cornils et al. (2018) - <https://doi.org/10.5194/essd-2018-36>
* The copepod concentrations recorded during the plankton net samples of the Tara Oceans expeditions and that were analyzed through the ZooScan imaging system (Brandão, Benedetti et al., 2021) - <https://doi.org/10.1038/s41598-021-94615-5>
* The copepod observations compiled by Becker et al. (2021) - <https://doi.org/10.1016/j.pocean.2021.102710>

**2.- METHODOLOGY USED**

The datasets listed above were first downloaded and re-formatted to the common AtlantECO WP2 data format. To homogenize the taxonomic classification of all the AtlantECO WP2 datasets containing microbiome (i.e., plankton) diversity data, the species names of each dataset were automatically compared to the list of species names accepted in the World Register of Marine Species (WoRMS), using the ‘worms’ R package version 0.2.2 (Holstein, 2018). Typos, synonyms and deprecated species names were corrected and the unique numerical identifiers of each accepted scientific name (i.e., AphiaID) were provided as well. Then, the datasets were progressively combined in as follows.

First, the NMFS-COPEPOD dataset was used as a basis since it corresponds to the most widely used compilation of historic zooplankton records. 135 records were removed because they corresponded to a non-existing species (*Macrosetella* *sulcata*). 81 943 records were removed because they were associated with missing abundance values. Then, an additional 42 840 records were also discarded because they corresponded to older CPR observations which we wanted to remove to avoid potential duplicate records with the updated CPR surveys available (total n records at this stage = 463 977). Second, we added the 665 921 copepod abundance records from the NANP-CPR survey. 99 833 records were removed because they were associated with missing abundance values (total n records = 1 030 065). Third, we added the 6 568 059 copepod abundance records from the SO-CPR survey. 620 604 records were removed because they were associated with missing abundance values (total n records = 6 977 520). Fourth, we added the 5 390 784 copepod abundance records from the AusCPR survey. 620 604 records were removed because they were associated with missing abundance values (total n records = 12 029 616). Then, we added the smaller datasets of Cornils et al. (2018), Brandão, Benedetti et al. (2021) and Becker et al. (2021), which added 260 826, 10 088 and 8 695 records of copepod abundance, respectively (total n records = 12 309 225).

Finally, we checked for potential duplicates again through the definition of an ‘occurrence ID’ that was based on the following parameters: decimal longitude, decimal latitude, sampling date, maximum sampling depth, scientific name and measured abundance value (‘MeasurementValue’). No more than 5 records were removed prior to occurrence ID definition as they displayed missing maximum sampling depth (‘MaxDepth’) values. We found 9 080 989 unique records (26.2% of the data were duplicates) and the corresponding 3 228 236 were duplicates based on the parameters chosen to define the occurrence ID.

The cleaned abundance sheet were passed through the following stages of processing to reach the final data set files.

1. The taxonomy of the indivudal carbon values was checked against WoRMS. Individual carbon values are individual carbon content measurements and Mean/Median/Min/Max/Stdev estimates (in mgC.ind-1). For Copepoda, 290 carbon content measurements obtained from **COPEPEDIA** (all values accessible here: <https://www.st.nmfs.noaa.gov/copepedia/taxa/T4000000/html/biometricframe.html>). These values can be found on sheet 3 of “AtlantECO-BASEv1\_dataset\_Hexanauplia\_ind\_carbon\_values\_20220930.xlsx”. This is also done for the field observations in the abundance sheet.
2. The life form variable gives information about the organism recorded (e.g., "singular", "colonial", "symbiotic", "free living", etc.). As data is taken from different sources, the ‘LifeForm’ entries show a range of different information and formats. All the life form entries within the abundance sheet and individual carbon values are standardized by subdividing the information in more detailed variables; Life Form Phase, Stage Name, Stage Number, Life form Min Size and Sex.
3. The unit of which the field observation is measured in it standardized. All entries in #/m2 are converted to #/m3. This is done by dividing the MeasurementValue (numeric value of the measurement, i.e., organisms concentration) by the difference between maximum and minimum depth (MeasurementValue /(MaxDepth – MinDepth)).
4. To calculate biomass concentrations, estimates of individual carbon content at the most precise taxonomic resolution possible (i.e., species-level mean carbon content to convert a species-level abundance observations) were determined. We took the highest known taxonomic classification of the observation and match it to the individual carbon content values. When there was no immediate match, we worked up the taxonomic ranking until a match is found. For example, for an observation identified down to the species level, *Scolecithricella minor minor,* no associated individual carbon content values are found until the order level (Calanoida). Therefore, it is matched with all individual carbon content values that have order Calanoida. In this instance the mean carbon mass calculated for this observation is composed of 246 different individual carbon content values. The minimum (Min) and maximum (Max) carbon mass is the lowest and highest values that we used to compute the mean (Mean) and its standard deviation (Stdev). When additional information of life form attributes is known, of either minimum size, phase, larval stage name, larval stage number and/or sex, it is considered when looking for matches. In some instances, the highest classification is to group level, which is an arbitrary name we assigned to functional groups to organize data sheets. When this occurs all individual carbon content values are averaged.
5. Each abundance record was converted to minimum, maximum and mean carbon biomass concentration (‘MinBiomass’, ‘MaxBiomass’ and ‘MeanBiomass’, respectively) expressed in mgC.m-3 based on estimates minimum, maximum and mean individual carbon content (‘MinBiomassConversion’, ‘MaxBiomassConversion’ and ‘MeanBiomassConversion’, respectively) expressed in mgC.ind-1.
6. Then the biomass concentrations of sampling events were summed. A sample event is defined by the Latitude, Longitude, Event Date, Depth, Year, Month and Day. This produces the total biomass concentration of all organisms of that specific event. The average biomass value was taken of organisms that had multiple replications at a sample event to avoid increase in biomass due to sampling replication.
7. The summed biomasses are rasterized by latitude, longitude and depth levels (WOA Depth Levels). As the depth levels do not show even intervals, spatial cubes of different y lengths are produced. Then to produce 1 lat x 1 long x 122 depth x 12 month dimensions the mean biomass per month over all the years is taken per 1 x 1 x 122. The minimum biomass is (MinBiomass) therefore comes from the month with the lowest value and the maximum biomass (MaxBiomass) is the month with the highest value.

The main R packages used to implement ZooBase v2 were: ‘rlang’ version 1.0.6 (Henley and Wickham, 2022), ‘fst’ version 0.9.8 (Klik, 2022), ‘writexl version 1.4.0 (Ooms and McNamara, 2021), ‘reshape2’ version 1.4.4 (Wickham, 2020), ‘RNetCDF’ version 2.5-2 (Michna and Woods, 2022), ‘lubridate’ version 1.8.0, (Grolemund and Wickham, 2011), ‘progress’ version 1.2.2 (Csárdi, 2018), ‘units’ version 0.7-2 (Pebesma et al. 2016), ‘stringr’ version 1.4.1 (Wickham, 2022), ‘tidytable’ version 0.9.0 (Fairbanks et al., 2022), ‘readxl’ version 1.4.1 (Wickam and Bryan, 2022), ‘worms’ version 0.4.4 (Chamberlain and Bart, 2022), and ‘taxize’ version 0.9.100 (Chamberlain et al., 2022).

**3.- DATASET DESCRIPTION**

**Data type:** Abundances converted to biomass concentrations.

**Latitude/Longitude format:** WGS 84 (-180°E/+180°E).

**Geographic area covered by the dataset:** Global Ocean.

**Depth range covered by the dataset:** From 0m to 4911m.

**Time period covered by the dataset:** From 13-08-1929 to 31-01-2021.

**Dataset format:** .csv file withsemicolon-delimited columns.

**Date of dataset creation:** 20/12/2022.

**Raw dataset repository:** AtlantECO’s GeoNode (<https://atlanteco-geonode.eu/>).

**4.- MAIN VARIABLE DESCRIPTION**

MeasurementTypeID: Has not been defined within AtlantECO

MeasurementValue: Organisms concentration (i.e., abundance) in ind.m-3

MeasurementID: Has not been defined within AtlantECO

occurrenceID: Combination of decimalLatitude, decimalLongitude, Day, Month, Year, MaxDepth, ScientificName, MeasurementValue.

**5.- DATA OVERVIEW**

**

**6.- CONTRIBUTORS**

- Fabio Benedetti (fabio.benedetti@usys.ethz.ch), ETH Zürich, Switzerland.

- David Johns (djoh@mba.ac.uk), MBA, United Kingdom.

- Ruby Bader (ruby.bader@usys.ethz.ch), ETH Zürich, Switzerland.

- Meike Vogt (meike.vogt@usys.ethz.ch), ETH Zürich, Switzerland.

**Special thanks are given from the contributors to Todd O’Brien and his team for the efforts they made throughout the years to make plankton field observations and carbon content estimates available tot he community.**