**TITLE: AtlantECO [WP2] – Traditional microscopy dataset –** **Zooplankton species occurrences (presence-absence data; ZooBase v2)**

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

**1.- INTRODUCTION**

This dataset contains **16 662 867** georeferenced occurrences of **4 072** accepted scientific names of zooplankton taxa. The data come from a previous zooplankton species occurrence compilation (ZooBase, Benedetti et al., 2021) that was complemented with: i) the most recent observations from the various Continuous Plankton Recorder (CPR) surveys (presence-absence for the Australian CPR and the Southern Ocean CPR, and presence-only data for the North Atlantic and North Pacific CPR); ii) the presence-absence data of gelatinous zooplankton from the MALASPINA expedition (Villarino et al., 2018); iii) the presence-absence data from the Jellyfish Database Initiative (JeDI; Lucas et al., 2014); iv) the presence-absence data of Antarctic salps and krill from KRILLBASE (Atkinson et al., 2017); v) the pteropod presence-absence data from the Atlantic Meridional Transect (AMT) #24 (Burridge et al., 2016), vi) the copepod presence-absence data from Becker et al. (2021), and vii) a suite of individual cruise-level datasets communicated by Ralf Schiebel (MPI, Germany) and which contained counts of calcifying zooplankton (pteropods and foraminifera; see references below).

**2.- METHODOLOGY USED**

An early version ZooBase species occurrence compilation of Benedetti et al. (2021) was used as a basis for the present AtlantECO dataset. The full description of how the final ZooBase was implemented can be found in Benedetti et al. (2021) - <https://doi.org/10.1038/s41467-021-25385-x>. We here first provide an early version of ZooBase, meaning that not all occurrences that were discarded for the specific purposes of the Benedetti et al. (2021) study were discarded here. This way we let potential future users decide which occurences they would like to retain, as a function of their scientific question(s) and the associated specific methodological requirements.

First, prior to retrieving the occurrence data online, we first identified the phyla (Order/Class/Family) that comprise the bulk of extant oceanic zooplankton communities: Copelata, Ctenophora, Cubozoa, Euphausiidae, Foraminifera, Gymnosomata, Hydrozoa, Hyperiidea, Myodocopina, Mysidae, Copepoda, Podonidae and *Penilia* *avirostris*, Chaetognatha, Scyphozoa, Thaliacea, Thecosomata, and pelagic Polychaeta (Tomopteridae, Alciopidae, Lopadorrhynchidae, Typhloscolecidae). The presence data associated with species belonging to these groups were retrieved from OBIS and GBIF between the 12/04/2018 and the 18/04/2018. In addition, the copepod occurrences from Cornils et al. (2018) were added as well to increase the coverage of observations in the Southern Ocean. These three main datasets were first re-formatted to the AtlantECO WP2 data format, with the exception that the dataset keys from GBIF and OBIS were kept as separate columns (GBIF.key and OBIS.key, respectively) for the sake of data traceability. We discarded the records that: (i) presented at least one missing spatial coordinate, (ii) were associated with an incomplete sampling date (d/m/y), (iii) were associated with a year of collection older than 1800, (iv) were not associated with any sampling depth, (v) and that were not identified down to the species level. At this stage, the OBIS zooplankton occurrences amounted to 2 288 615, the GBIF occurrences amounted to 922 964 and the copepod occurrences from Cornils et al. (2018) amounted to 260 826 (total number of occurrences = 3 472 405). 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. This reformatted compilation contained 3 472 405 occurrences of 3 512 different species names. We kept the 25 names (0.71% of the 3 512 unique scientific names, corresponding to 0.02% of 3 472 405 occurrences) that had no match found in WoRMS and therefore no AphiaID could be associated to those names.

Because this early ZooBase version already contained occurrences from the CPR surveys, the later had to be identified to be replaced by the new and updated CPR data to avoid retaining duplicate occurrences in ZooBase v2. The qualitative ‘Institution Code’ values and data mapping enabled to identify those early CPR occurrences. Occurrences whose Institution Codes were equal to ‘SAHFOS’, ‘AADC’ or ‘CSIRO Oceans and Atmosphere, Australia (CMAR)’ were discarded as these corresponded to observations stemming from the North Atlantic and North Pacific CPR (NA-NP CPR), the Southern Ocean CPR (SO CPR) and the Australian CPR (AusCPR), respectively. Further checking of the sampling metadata and mapping of sampling events confirmed that removing those occurrences did remove all CPR-based zooplankton occurrences. This way, 1 763 689 zooplankton species occurrences were removed. Then, because OBIS and GBIF are two online data atlases that are known to have highly overlapping data sources, duplicate occurrences were identified and removed within this early ZooBase version based on combinations of: decimal longitude, decimal latitude, sampling date, sampling depth and scientific name. This identified 784 839 unique zooplankton species occurrences, which translates to a rate of duplicate of nearly 47%. At this stage, this early ZooBase version contained occurrences across 3 496 accepted species names (25 scientific names had no accepted entry in WoRMS).

Then, this early ZooBase version was updated by adding nine datasets containing zooplankton species occurrences: i) North Atlantic and North Pacific CPR (NA-NP CPR; Johns & Broughton 2019; <https://doi.org/10.17031/1629>), ii) the Australian CPR (AusCPR; <https://catalogue-imos.aodn.org.au/geonetwork/srv/eng/catalog.search#/metadata/c1344e70-480e-0993-e044-00144f7bc0f4>), iii) the Southern Ocean CPR (SO CPR; Hosie et al., 2021; doi:10.26179/ksds-s610), iv) the neustonic zooplankton species occurrences (presence-absence) from the MALASPINA expedition (Villarino et al., 2018; <https://doi.org/10.1594/PANGAEA.874651>), v) the gelatinous zooplankton species occurrences (presence-absence) from JeDI (Lucas et al., 2014; accessed via <https://www.bco-dmo.org/dataset/526852>), vi) the Antarctic salp and krill occurrences (presence-absence) from KILLBASE (Atkinson et al., 2017; accessed via <https://data.bas.ac.uk/full-record.php?id=GB/NERC/BAS/PDC/00915>), vii) the pteropod occurrences (presence-absence) from the AMT24 (Burridge et al., 2016), viii) the copepod occurrences (presence-absence) from the two cruises compiled in supplementary material of Becker et al. (2021), and ix) the following suites of cruise-level datasets communicated by Ralf Schiebel (MPI, Germany) and which contained presence-absence data for both planktic foraminifers and pteropods:

* Schiebel et al. (1995) - <https://doi.org/10.1016/0377-8398(95)00035-6>
* Schiebel (2002) - <https://doi.org/10.1029/2001GB001459>
* Schiebel et al. (2001) - <https://doi.org/10.1016/S0967-0637(00)00065-0>
* Jentzen et al. (2018) - <https://doi.org/10.2113/gsjfr.48.3.251>
* Schiebel et al. (2000) - <https://doi.org/10.1016/S0967-0645(00)00008-4>
* Schiebel et al. (2002) - <https://doi.org/10.1016/S0967-0645(02)o00141-8>
* Schiebel et al. (2004) - <https://doi.org/10.1016/j.marmicro.2004.02.001>

*Note: the associated pteropod counts were actually never used in the seven references above although they come from the exact same zooplankton sampling events. This is why the bibliographic citation of these pteropod occurrences was labelled as ‘Unpublished data from R. Schiebel (pers. comm.)’. Details about the sampling methodology are provided in the references cited above though.*

Prior to being added to ZooBase, those nine groups of datasets were reformatted to the AtlantECO WP2 format and the scientific names of the zooplankton species were also corrected/updated based on the backbone classification of WoRMS. ZooBase and those nine complementary datasets constitute the present ZooBase v2 dataset.

The NA-NP CPR only contained positive counts of zooplankton organisms (in numbers per m3) therefore missing count values were removed (n = 105 587) and all counts were converted to presences. This updated NA-NP CPR survey added 607 133 zooplankton presences. The AusCPR survey provided positive and null counts of zooplankton organisms (i.e., presences and absences) therefore missing count values were removed (n = 3) and all counts were converted to presence-absence. The AusCPR survey added 7 373 517 zooplankton occurrences. After a rapid examination of the taxonomic content of the dataset, it appeared that 183 456 occurrences at this stage actually belonged to the Ochrophyta and the Myzozoa phyla therefore they were removed (total n occurrences = 8 582 033). The SO CPR survey provided 15 204 798 positive and null counts of zooplankton organisms (i.e., presences and absences) therefore all count data (no missing values) were converted to presence-absence. Out of those, 362 019 data points corresponded to non zooplankton taxa (e.g., ‘fish eggs’ or ‘Nauplius’) so they were removed. Therefore the SO CPR survey added 14 842 779 zooplankton occurrences. At this stage, the new ZooBase dataset amounted to 23 424 812 occurrences across 3 747 accepted zooplankton species names.

Then, we added the 671 occurrences (presence-absence) of neustonic zooplankton observed during the MALASPINA expedition and the 2 983 occurrences (presence-absence) of Antarctic salp and krill (*Salpa* spp. and *Euphausia* *superba*) coming from KRILLBASE (total occurrences = 23 446 433).

Another previously published global zooplankton observations compilation is the JeDI (Lucas et al., 2014) which provides both occurrence (presence-absence; n = 298 969) and abundance data (in number of individuals per m3 or m2; n = 238 083). Both were added to the present ZooBase update after abundance were converted to presence-absence data (n = 537 051 occurrences). One record had to be discarded because it corresponded to a negative ‘Hydrozoa’ abundance value which is unrealistic. JeDI also being a data compilation, it could provide occurrences overlapping with those provided by the OBIS, GBIF or other datasets mentioned above. Therefore, at this stage, we decided to identify and duplicates. Those were identified through an “occurrence ID” defined by the: decimal longitude, decimal latitude, sampling date, sampling depth, scientific name, and measured occurrence value (‘presence’ or ‘absence’). Prior to this, 296 040 occurrences had to be removed because they were not associated with any sampling depth. Then, 16 604 195 unique occurrences were retained (7 083 249 duplicate occurrences were discarded).

Finally, we added the 8 732 copepod occurrences (presence-absence) from Becker et al. (2021); the 1 953 pteropod occurrences (presence-absence) from Burridge et al. (2016); the 259 900 foraminifera occurrences and the 48 378 pteropod occurrences coming from the seven datasets provided by R. Schiebel (total occurrences = 16 923 121; total number of accepted scientific names = 4 104).

Finally, we checked for potential duplicates again through the definition of an ‘occurrence ID’ that was based on the same set of metadata mentioned above. We found 16 670 316 unique zooplankton occurrences (corresponding to 98.5% of the data) and 252 805 were duplicates based on the parameters chosen to define the occurrence ID. An additional 7 449 occurrences were removed because they were clearly located on land based on the associated bathymetry value and position on a global map (i.e., map below). This lead to the final version of the present ZooBase v2 dataset (16 662 867 georeferenced occurrences of 4 072 different accepted scientific names of zooplankton taxa).

On top of the libraries already mentioned above, the main R packages used to implement ZooBase v2 were: ‘tidyverse’ (Wickham et al., 2019), ‘reshape2’ version 1.4.4 (Wickham, 2007), ‘marmap’ version 1.0.6 (Pante & Simon-Bouhet, 2013), ‘lubridate’ version 1.8.0 (Grolemund, 2011) and ‘raster’ version 3.5-15 (Hijmans, 2022).

**3.- DATASET DESCRIPTION**

**Data type:** Presence (1) or absence (0) of species.

**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 12 500m.

**Time period covered by the dataset:** From 06-10-1853 to 31-01-2021.

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

**Date of dataset creation:** 09/09/2022.

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

**4.- MAIN VARIABLE DESCRIPTION**

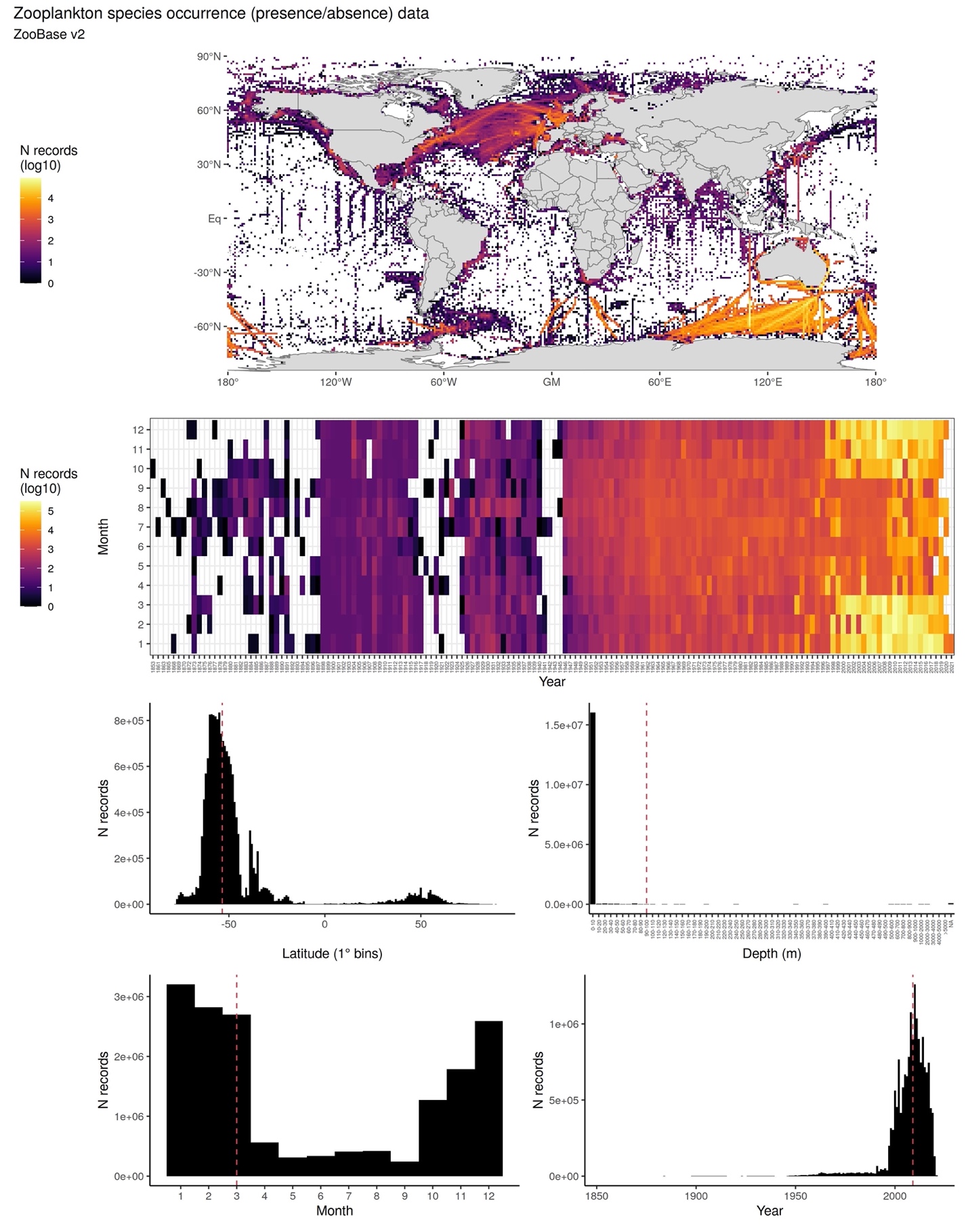
MeasurementTypeID: Has not been defined within AtlantECO

MeasurementValue: Presence or absence of species

MeasurementID: Has not been defined within AtlantECO

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

**5.- LINKS**

**

**6.- CONTRIBUTORS**

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

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

- Nielja Knecht (nielja.knecht@gess.ethz.ch), ETH Zürich, Switzerland.

- Katja Peijnenburg (katja.peijnenburg@naturalis.nl), Naturalis Biodiversity Center, The Netherlands.

- Ralf Schiebel (ralf.schiebel@mpic.de), MPI, Germany.

- Sonia Chaabane (sonia.chaabane@fondationbiodiversite.fr), FRB-CESAB, France.

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