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

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

**1.- INTRODUCTION**

This dataset contains **5 167 282** georeferenced occurrences of **1 691** accepted phytoplankton species names. The data come from a previous phytoplankton species occurrence compilation (PhytoBase, Righetti et al., 2020; accessed from: <https://doi.pangaea.de/10.1594/PANGAEA.904397> on the 22-02-2021) that was complemented with the observations (presence-absence) data from the MALASPINA expedition (Villarino et al., 2018) and 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).

**2.- METHODOLOGY USED**

The species occurrence data of PhytoBase (Righetti et al., 2020) compiled observations from various original sources: the Global Biodiversity Information Facility (GBIF; https://www.gbif.org), the Ocean Biogeographic Information System (OBIS; https://www.obis.org), the data from Villar et al. (2015), and the MAREDAT initiative (Buitenhuis et al., 2013). Righetti et al. gathered >106 presences from nearly 1 300 species sampled through various methodologies within the monthly climatological mixed-layer depth, at an average depth of 5.41 ± 6.95 m (mean ± sd), between 1800 and 2015. The species names were corrected and harmonized following the reference list of Algaebase (http://www.algaebase.org/) and were further validated by expert opinion. The original PhytoBase dataset is accessible at: <https://doi.pangaea.de/10.1594/PANGAEA.904397>.

PhytoBase was 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. The quantitative and semi-quantitative cell counts recorded in PhytoBase were all converted to binary presence-absence data (1/0) based on the ‘individualCount’ and ’organismQuantity’ columns (i.e., all values > 0 were converted to presences). Data points that showed missing values in those two columns came from GBIF and OBIS and were thus considered as presences only. Data points associated with positive values of ‘organismQuantity’ correspond to cell counts from the Sal et al. (2013) compilation and the MAREDAT datasets, so they could be confidently transformed into presence-absence data. This reformatted dataset contained 1 360 621 occurrences of 1 710 different species names. To homogenize the taxonomic classification of all the AtlantECO WP2 datasets containing microbiome (i.e., plankton) diversity data, the list of species names was 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. After visually inspecting the list of scientific names, we chose to discard the 27 537 occurrences of ‘Picoeukaryotes’ (total n occurrences = 1 333 084) as this functional category has no real taxonomic meaning. Nonetheless, we kept the 128 names that had no match found in WoRMS or AlgaeBase and therefore no AphiaID could be associated to those names (7.5% of the 1 703 unique names, corresponding to 0.86% of 1 333 084 occurrences).

Furthermore, when examining the institution codes of the various data sources, we noted that some of the PhytoBase occurrences corresponded to fossil records or sediment samples from drilling programs such as the Ocean Drilling Program (ODP). Therefore, the ‘robis’ (Provoost & Bosch, 2021) and ‘rgbif’ (Chamberlain et al., 2022) R packages (versions 2.8.2 and 3.7.1, respectively) were used to extract the original datasets’ names based on the OBIS.key and GBIF.key and a list of keywords was used to identify those fossil and sediment records (*full list*: "ODP Hole","DSDP Hole","Hole 81","Hole 84","surface sediment samples","ODP Site","Site 175-1075","Hole 302","palynological","nannofossils","sediment core"," Hole","fossil","site GIK23414","surface sediments of the Arabian Sea","site GIK2325","surface sediments","NU2\_trap","DSDP Site","site GIK23259","Core SO82","Sediment components","deep-sea sediments","site GIK23414","Trap data","sediment trap","palynomorpha","surface sediments of cores","Site 177-1092","mooring MST-8","mooring MST-9","palinológica","dinoflagellate cysts of multicorer","cysts of surface sediments","multicorer surface","cysts from surface sediments","sediment trap MST-9","Terrestrial and Limnetic","trap WR2","Dinoflagellate cyst species composition of recent surface samples","sediment profile PS1730","grain fraction > 20 µm","grain fraction < 20 µm"). The phytoplankton species occurrences coming from such datasets were not removed but were flagged as ‘Records collected from sediment core or trap’ in the ‘Flag’ column.

Then, this reformatted PhytoBase was updated based on 4 datasets containing phytoplankton species occurrence data: 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; *Noctiluca scintillans* occurrences only; Hosie et al., 2021; doi:10.26179/ksds-s610), and iv) the diatoms, dinoflagellates and coccolithophores species occurrences (presence-absence) from the MALASPINA expedition (Villarino et al., 2018; datasets available on PANGAEA as follows: <https://doi.org/10.1594/PANGAEA.874647>, <https://doi.org/10.1594/PANGAEA.874650> and <https://doi.org/10.1594/PANGAEA.874646>, respectively). Prior to being added to PhytoBase, those four datasets were reformatted to the AtlantECO WP2 format and the scientific names of the phytoplankton species were also corrected/updated based on the backbone classification of WoRMS. PhytoBase and those four complementary datasets constitute the present PhytoBase v2 dataset.

First, because the original PhytoBase already contained CPR occurrences, the later had to be identified in order to be replaced by the new and updated CPR data to avoid retaining duplicate occurrences in PhytoBase v2. Therefore, like above, the occurrences linked to any CPR survey were identified and removed thanks to the occurrences’ institution codes and the ‘robis’ and ‘rgbif’ R packages. Any occurrence affiliated to one of the following key words was thus discarded: ‘Continuous Plankton Recorder’, ‘AusCPR’, ‘SAHFOS’, and ‘CSIRO Marine and Atmospheric Research, Australia (CMAR)’.

While doing so, it appeared that the original PhytoBase dataset available contained duplicate phytoplankton occurrences from the WODC Plankton Database (a dataset recorded by both OBIS and GBIF; GBIF.key == "838e2626-f762-11e1-a439-00145eb45e9a" & OBIS.key == "2546"). Yet, the OBIS version contained 91 more scientific species names and 220 701 more occurrences than the GBIF version. Consequently, the GBIF version of the WODC Plankton Database was discarded except the 17 340 occurrences corresponding to scientific names not present in the OBIS version (i.e., we assumed these corresponded to occurrences not recorded in the OBIS version of the WODC Plankton Database for a reason we cannot identify based on the metadata made available). This removed 149 882 occurrences from the original PhytoBase dataset and lead to a corrected total number of occurrences of 1 183 202.

Then, all the occurrences associated with any CPR survey were removed. This discarded no less than 511 988 occurrences (43.2% of the 1 183 202 occurrences from the corrected PhytoBase; 671 214 occurrences remaining).

Second, the phytoplankton species occurrences from the four additional datasets mentioned above were progressively added. The NA-NP CPR (Johns & Broughton 2019; <https://doi.org/10.17031/1629>) added 755 168 presences (absences corresponding to null cell counts were not provided) to PhytoBase (total occurrences = 1 215 445). The AusCPR survey (<https://catalogue-imos.aodn.org.au/geonetwork/srv/eng/catalog.search#/metadata/c1344e70-480e-0993-e044-00144f7bc0f4>) added 3 863 540 new occurrences (presences and absences). To complete the contribution of the CPR surveys to the updated PhytoBase dataset, the 51 717 occurrences (presences and absences) of *Noctiluca scintillans* stemming from the SO CPR (Hosie et al., 2021; doi:10.26179/ksds-s610) were also added (total occurrences = 5 130 702).

Finally, the 36 922 occurrences (presences and absences of: diatom, dinoflagellate and coccolithophore taxa) stemming from the MALASPINA expedition (Villarino et al., 2018) were added (total occurrences = 5 167 624).

After a final check on all scientific names and occurrence metadata, 342 occurrences were removed because they had an incomplete or an impossible sampling year. This lead to the final version of the present PhytoBase v2 dataset (5 167 282 georeferenced occurrences of 1 691 different accepted phytoplankton species names).

On top of the libraries already mentioned above, the main R packages used to implement PhytoBase 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 4647m.

**Time period covered by the dataset:** From 02-07-1817 to 16-04-2021.

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

**Date of dataset creation:** 05/09/2022.

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

**4.- MAIN VARIABLE DESCRIPTION**

MeasurementTypeID: Not defined within AtlantECO yet

MeasurementValue: Presence or absence of species

MeasurementID: Not defined within AtlantECO yet

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

**5.- DATA COVERAGE**

**

**6.- CONTRIBUTORS**

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

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

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