**Data extraction “Bacillariophyceae” project**

**Extraction performed by Pierre Hélaouët**

**08/11/2024.**

**DOI: 10.1703/672deb7d70806**

**Note:** As always, this dataset has been carefully built and checked accordingly. However, it is the user’s responsibility to perform his own verifications.

**Quick description of the dataset**

**1 – The dataset contains 4 files:**

1. “CPR\_Data\_Bacillariophyceae\_08112024.docx”: This document
2. “CPR\_Bacillariophyceae\_ControlMap\_08112024.png”:

Map representing the selected samples from January 1958 to December 2021 (252890 samples)

1. “CPR\_Bacillariophyceae\_Data\_08112024.csv”: Abundance data for all selected phytoplankton belonging to the class “Bacillariophyceae“ (88 taxa, see CPR\_Bacillariophyceae\_List\_08112024.csv) and all selected samples in the selected area (30°N to 80°N, -80°E to 20°E).

Rows: All samples for the selected area (252890 samples).

Column 1: Unique sample id. For instance: “240B--27” corresponds to the 27th sample for the 240th transect on the B route.

Columns from 2 to 8: Spatio-temporal coordinates for each sample.

Columns from 9 to 95: Abundance data for all selected taxa (88).

Note 1: We may notice very small values (10^-10) or any other number with a very small fraction. Sometimes, our analysts can identify the presence of a specific taxa but are unable to quantify it. In that case, they report the taxa as “present”. This is hard-coded in our database as a very small value (10^-10) for statistical reasons.

Note 2: In a given sample, the abundance value of a specific taxon, is set to NaN (Not A Number) when the corresponding Data of Routine Identification (DRI) is posterior to the date of sample collection.

1. “CPR\_Bacillariophyceae\_List\_08112024.csv”: List of all selected phytoplankton belonging to the class “Bacillariophyceae“.

Rows: All selected taxa (88 taxa).

Column 1 “accepted\_id”: Unique identifier used by the CPR survey

Column 2 "Aphia\_id”: Identifier used by WoRMS

Column 3 "name-CPR”: Unique name used by the CPR survey.

Column 4 "Name\_worms”: Name used by WoRMS corresponding to the “aphia\_id”.

Column 5 "DRI”: Date of Routine Identification. Before that date, un taxon was not on our routine taxa list. For a given taxon, abundances associated with samples taken before the DRI are set to a NaN (Not A Number).