**TITLE: AtlantEco [WP2] – Traditional and -omics dataset – Diazotroph qualitative (presence vs. absence) and quantitative (nifH counts, cell counts, biomass entries) records**

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

1. **INTRODUCTION**

After concatenating all diazotroph sources, duplicates have been removed using taxonomy, geospatial and temporal information. The final dataset includes more than 20 thousand records (Figure: 1A). The updated database increases existing database compilation efforts such as Luo et al. (2012) and Tang and Cassar (2019) by about 30%. About 84% of the records are presences, whereas about 16% correspond to potential absences.

The global sampling distribution covers all major ocean basins with the highest sampling effort in the Atltantic (45%), followed by the Pacific (39%), Indian Ocean (9%) and polar seas (Figure 1B). Within the Atlantic sampling events are highly uneven with more than 40% of observations falling into the North Atlantic, while only about 2% of the total observations have been sampled in the southern Atlantic. Within the Pacific Ocean, sampling effort between the northern and southern Pacific is more evenly distributed, when compared to the Atlantic Ocean. Roughly 23 % of total observations fall into the north Pacific and roughly 16% fall into the south Pacific. The Indian Ocean holds about 9% of the total observations, while other seas such as the Mediterranean and Baltic Sea account for roughly 7%. Both polar seas are highly underrepresented within our dataset with both polar seas accounting for less than 1% of global observations.

Most represented diazotrophic taxa is Trichodesmium, reaching a fraction of more than 50% of all observations (Figure 1C), followed by genus Richelia (18\%). All other diazotroph taxa contribute with less than 10% to the total fraction of observations. Non-cyanobacterial diazotrophs have only been retrieved from Tara Oceans survey and contribute less than 1% individually.

1. **METHODOLOGY USED**

We compiled diazotroph occurrence data collected from public sources or recent papers that are described hereafter. We focused on quantitative– (counts, or reads) and qualitative (presence vs. absence, non-detection) field records of diazotrophs. We compiled diazotrophic data from the Global Biodiversity Information Facility (GBIF; www.gbif.org/, last access: 20 October 2021), the Ocean Biogeographic Information System (OBIS; www.obis.org/, last access: 21 October 2021), Luo et al. (2012), Tang & Cassar (2019), Gradoville et al. (2020) , Amalia et al. (2022), Martínez- Pérez et al. (2016) Phytobase , Juan Karlusich et al. (2021) and Lucas Paoli et al. (2022). With the exception of data retrieved from public databases such as OBIS and GBIF, the accession of data was straightforward, as the georeferenced, post-processed, curated datasets were obtained from repositories or through personal communication.

1. **DATASET DESCRIPTION:**

**Data type:** Qualitative (Presence or absence) and quantitative (trichome counts, cell counts, nifH gene counts, biomass estimates).

**Latitudinal/Longitudinal format:** WGS 84 (-180°E/+180°E).

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

**Depth range covered by the dataset**: 0 – 3886 meters

**Time period covered by the dataset:** 1824 - 2020

**Dataset format:** .csv file with semicolon-delimited columns.

**Date of dataset creation:** 18/08/2022.

**Raw dataset repository:** ETHZ’s FTP server.

1. **DATA CONTRIBUTORS:**
* Dominic Eriksson (deriksson@ethz.ch), ETH Zürich, Switzerland
* Damiano Righetti (damiano.righetti@env.ethz.ch)
* Meike Vogt (meike.vogt@env.ethz.ch)
* Shinichi Sunagawa (ssunagawa@ethz.ch), ETH Zürich, Switzerland