Paper review

"PHYTOBASE: A global synthesis of open ocean phytoplankton occurrences"

Righetti et al.

REVIEW SUMMARY

The authors present a compilation, named PhytoBase, of five data sources on phytoplankton occurrence records targeting open ocean, including two main data sources: the Global Biodiversity Information Facility (GBIF; www.gbif.org), and the Ocean Biogeographic Information System (OBIS; www.obis.org), complemented by three other sources: the Marine Ecosystem Data initiative (MareDat; Buitenhuis et al. 2013), a marine micro-phytoplankton dataset (Sal et al., 2013), and with a subset of the data collected during the TARA Oceans cruise (Villar et al., 2015). To my knowledge, this compilation leads to the largest dataset on open ocean phytoplankton. A huge effort of data harmonization is recognized, on several aspects of both data structure and taxonomy but also on data qualification (cleaning) required to ensure data quality. This database opens perspectives for phytoplankton research on niche modelling, species distributions, especially within the context of global changes. This database, if updated and maintained in time, will be a valuable bibliographic source for future phytoplankton studies.

I would suggest the acceptance of the paper with 'Minor changes', but I give some recommendations that if addressed will contribute to strengthen and improve the quality of the present paper but also PhytoBase, in particular on data structuring and processes to maintain the valuable product. In my vision, the potential of PhytoBase lies in the compilation of existing data sources but essentially lies in its reproducibility, sustainability and maintenance in time instead of one single snapshot, even more because PhytoBase relies on existing data sources that are maintained and grow in time. That is the reason why some comments insist on sustainability and maintenance aspects, with emphasis on the need of reproducible processes.

DETAILED COMMENTS

Abstract

No comments

1. Introduction

No comments

2. Compilation of occurrences

2.1. Data origin

- Line 100-104: The authors should argue further why they have chosen these three complementary data sources in particular: the MareDat, data from Sal et al. and TARA

data collection subset. Did the authors proceed to some extensive bibliographic work to search for potentially valuable datasets and were the sources used the only available open datasets? If yes, this deserves further statements on this bibliographic work, and eventual criteria (if applicable) to choose the data sources.

- Lines 120-121: R packages *RPostgreSQL* and *devtools* should be properly cited and referenced

2.2.Data selection

2.2.1. Data accessed through GBIF and OBIS

- Lines 146-149: please provide percentage of records excluded with filters on year and missing date.

2.3.Concatenation of source datasets

- Line 188, In table 1: The authors do not mention in the main document that the two main data sources GBIF and OBIS extensively rely on the Darwin Core standard (https://dwc.tdwg.org/terms/). This explains why most of column names are the same. In addition, the authors should precise that an attempt was done to match Darwin Core standard in the final column names, as working to comply with a standard is in general a best practice and an added value for the work. Due to the fact the two main sources are aligned on Darwin Core, not mentioning Darwin Core might be seen as regression. To understand that the Darwin Core standard has been exploited by the authors, we have to refer to the CSV table (http://hs.pangaea.de/Projects/PHYTOBASE/Column_definition_for_phytoplankton_harmonized_database.zip) available under section 'further details' of PhytoBase PANGEA record available at https://doi.pangaea.de/10.1594/PANGAEA.904397
- Lines 188, In table 1 : The table intends to harmonize the column names, and tries to use Darwin Core when possible. This is achieved with the following fields: scientificName, basisOfRecord, decimalLongitude, decimalLatitude, taxonRank, individualCount, year, month, day, but not for other fields. Indeed, for fields that are common to at least two data sources, such as Darwin Core column names for GBIF and OBIS, the table results in a some kind of de-harmonization and de-standardization of column names, such as for:
    - institutionCode (Darwin Core term), that is split in two separate columns specific to GBIF/OBIS, ie institutionCode_gbif / institutionCode_obis. It would have been preferable to keep the standard column name, and act at content level to keep source provenance, for example adding a prefix or URN such as *urn:gbif:<institutioncode>* or *urn:obis:<institutioncode>* as content of a single institutionCode
    - cellsPerLitre : This is a non standard term. It is recommended to keep aligning on the Darwin Core standard by relying on columns relative to measurements or facts, in particular to use standard terms measurementType ("number of cells"), measurementValue (value of cell number), and measurementUnit (number of cells per litre)

    - Depth: This is an non standard term and it deserves a reflexion whether the use of standard terms minimumDepthInMeters and maximumDepthInMeters could be relevant.

  In similar way, it is recommended that authors check in depth about existence of Darwin core terms that match the other column names: originDatabase, datasetKey, collectionCode, resname, resourceID, cruiseOrStationId, cruise, sampleId; while avoiding the use of source-specific column names, as illustrated above with the *institutionCode*.
- Lines 188, In table 1, row about cellsPerLitre: The authors should also revise the corresponding table row as it seems information has been wrongly copied-pasted ("taxonRank")
- Line 210. The authors make use of a column "group" to add either the Phylum or Class. It is recommended to keep using Darwin Core standard terms phylum and class as separate columns.
- Line 212: The authors make use of a column "sourceArchive" to refer to the data source from which the record comes from. It is recommended to look carefully at Darwin Core standard to find the appropriate standard term to use for referencing the data source.
- Beyond the harmonization of column names highlighted in Table 1, since I believe it is the core of paper describing the set-up of PhytoBase, it would be highly valuable to include in the main document the final data structure retained in the PhytoBase (as set in table http://hs.pangaea.de/Projects/PHYTOBASE/Column_definition_for_phytoplankton_harmonized_database.zip), including column definitions, and for the extra columns added by authors, to proceed with an in-depth check about existence of Darwin core terms to use instead of adhoc column names, as recommended with column names enumerated above. In fact, the high potential of PhytoBase and perspective to exploit it will be fostered by such Darwin Core standard compliance. By relying on Darwin Core, this will offer perspectives to facilitate growing of source global information systems such as GBIF or OBIS with datasets not yet available through it, while benefiting from data already harmonized and standardized through PhytoBase.

2.3.1. Extant species selection and taxonomic harmonization

- Lines 223-227: The authors refer to a screaning process performed by Algaebase founder and director, as personal communication. This screening led to exclude a relatively significant number of taxa and associated data. Hence, such process seems to appear as key harmonization task for PhytoBase. In my opinion, such process should be further described in the actual PhytoBase and paper materials & methods. In addition, there is no statement that make understand whether the screaning process was done manually or through a semi-automated procedure. If it is a manual process, this may be seen as a limitation referring to reproducibility, sustainability and maintenance of PhytoBase, even more because it has not been operated by PhytoBase creators/maintainers. It is then strongly recommended to describe further such

screening process within the main document (or through an appendix), and, if done manually, to suggest how this could be replaced or at least complemented by a semi-automated and reproducible process , thus leading to the possibility for future users to get an updated PhytoBase in time.

2.3.2. Data merger and synthesis

- Line 270: The *rgbif* R package should be properly cited and referenced. In addition, please note that there is a typo with the package name ('rgibf' instead of *rgbif*).

3. Results

3.1. Data

- This section is very welcome and acknowledged.

3.1.1. Spatiotemporal coverage

- Line 283: It is recommended to add the EPSG code of the World Geodetic System (WGS84). In addition, I recommend to include this as standard Darwin Core column in PhytoBase using the term geodeticDatum.

5. Data availability

- In principle, it is highly recommended, based on principles of open and reproducible science and sustainability, that authors make available already the R scripts together with the PhytoBase on PANGAEA, and avoid provision on demand through emails to the authors.