Enhancing georeferenced biodiversity inventories: automated information extraction from literature records reveal the gaps (#73004)

First submission

Guidance from your Editor

Please submit by **26 May 2022** for the benefit of the authors (and your $200 publishing discount) .

✓ **Structure and Criteria**
Please read the 'Structure and Criteria' page for general guidance.

▦ **Raw data check**
Review the raw data.

▨ **Image check**
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files Download and review all files from the materials page.	1 Latex file(s) 2 Other file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

🗋 You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- ✓ Clear, unambiguous, professional English language used throughout.
- ✓ Intro & background to show context. Literature well referenced & relevant.
- ✓ Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- ✓ Figures are relevant, high quality, well labelled & described.
- ✓ Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- ✓ Original primary research within Scope of the journal.
- ✓ Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- ✓ Rigorous investigation performed to a high technical & ethical standard.
- ✓ Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- ℹ Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- ✓ All underlying data have been provided; they are robust, statistically sound, & controlled.
- ✓ Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Enhancing georeferenced biodiversity inventories: automated information extraction from literature records reveal the gaps

Bjørn Tore Kopperud [1, 2, 3], **Scott Lidgard** [4], **Lee Hsiang Liow** [Corresp. 1, 5]

[1] Natural History Museum, University of Oslo, Oslo, Norway
[2] GeoBio-Center, Ludwig-Maximilians-Universität München, Müchen, Germany
[3] Department of Earth and Environmental Sciences, Ludwig-Maximilians-Universität München, Müchen, Germany
[4] Negaunee Integrative Research Center, Field Museum of Natural History, Chicago, Illinois, U.S.A.
[5] Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway

Corresponding Author: Lee Hsiang Liow
Email address: l.h.liow@ibv.uio.no

We use natural language processing (NLP) to retrieve location data for cheilostome bryozoan species (text-mined occurrences [TMO]) in an automated procedure. We compare these results with data combined from two major public databases (DB): the Ocean Biogeographic Information System (OBIS), and the Global Biodiversity Information Facility (GBIF). Using DB and TMO data separately and in combination, we present latitudinal species richness curves using standard estimators (Chao2 and the Jackknife) and range-through approaches. Our combined DB and TMO species richness curves quantitatively document a bimodal global latitudinal diversity gradient for extant cheilostomes for the first time, with peaks in the temperate zones. 79% of the georeferenced species we retrieved from TMO (N = 1408) and DB (N = 4549) are non-overlapping. Despite clear indications that global location data compiled for cheilostomes should be improved with concerted effort, our study supports the view that many marine latitudinal species richness patterns deviate from the canonical latitudinal diversity gradient (LDG). Moreover, combining online biodiversity databases with automated information retrieval from the published literature is a promising avenue for expanding taxon-location datasets.

Enhancing georeferenced biodiversity inventories: automated information extraction from literature records reveal the gaps

Bjørn Tore Kopperud[1,2,3], **Scott Lidgard**[4], **and Lee Hsiang Liow**[3,5]

[1]GeoBio-Center LMU, Ludwig-Maximilians-Universität München, 80333 Munich, Germany
[2]Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, 80333 Munich, Germany
[3]Natural History Museum, University of Oslo, PO Box 1172 Blindern, 0318 Oslo, Norway
[4]Negaunee Integrative Research Center, Field Museum, 1400 South Lake Shore Drive, Chicago IL, 60605, U.S.A
[5]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, PO Box 1066 Blindern, 0316 Oslo, Norway

Corresponding author:
Lee Hsiang Liow[3,5]

Email address: l.h.liow@nhm.uio.no

ABSTRACT

We use natural language processing (NLP) to retrieve location data for cheilostome bryozoan species (text-mined occurrences [TMO]) in an automated procedure. We compare these results with data combined from two major public databases (DB): the Ocean ~~Biogeographic~~ Information System (OBIS), and the Global Biodiversity Information Facility (GBIF). Using DB and TMO data separately and in combination, we present latitudinal species richness curves using standard estimators (Chao2 and the Jackknife) and range-through approaches. Our combined DB and TMO species richness curves quantitatively document a bimodal global latitudinal diversity gradient for extant cheilostomes for the first time, with peaks in the temperate zones. 79% of the georeferenced species we retrieved from TMO (N = 1408) and DB (N = 4549) are non-overlapping. Despite clear indications that global location data compiled for cheilostomes should be improved with concerted effort, our study supports the view that many marine latitudinal species richness patterns deviate from the canonical latitudinal diversity gradient (LDG). Moreover, combining online biodiversity databases with automated information retrieval from the published literature is a promising avenue for expanding taxon-location datasets.

INTRODUCTION

Biogeography

Global biogeographical and macroecological studies require data on aggregate entities, such as location-specific occurrences of taxa and regional species assemblages, in order to understand emergent patterns at global and/or temporal scales (McGill, 2019). Assembly of such detailed yet broad-scale data is highly labor-intensive; the sampling effort required for a specific research question can be daunting for any one researcher or single research team. This is one reason why collaborative and often public databases have gained traction (Klein et al., 2019; Heberling et al., 2021). For instance, empirical global biogeographic analyses (Costello et al., 2017; Rivadeneira and Poore, 2020; Chaudhary et al., 2021; Hughes et al., 2021) are increasingly based on public databases of georeferenced taxonomic occurrences, such as the Ocean ~~Biogeographic~~ Information System (OBIS, ~~www.iobis.org)~~ and the Global Biodiversity Information Facility (GBIF, www.gbif.org). Analyzing such georeferenced databases with tools that partially alleviate incomplete or biased sampling (Edgar et al., 2017; Kusumoto et al., 2020; Zizka et al., 2020; Grenié et al., 2022), allows us to address questions on large-scale distributions of clades, especially those that are well-represented in such databases. For less well-studied clades, however, prospects for obtaining large amounts of such data are lower. Answering pattern-based questions such as ‘how many species of clade z are found in location y’ and more process-oriented questions such as ‘how did the current latitudinal diversity gradient form’ both require location-specific taxonomic data in substantial volume. In addition, generalized biogeographic hypotheses have the potential to be supported more robustly if they include a greater diversity of clades.

Cheilostomes

Cheilostome bryozoans, though less well-studied than several metazoan clades of similar size, are ubiquitous in benthic marine habitats. They are the most diverse order of Bryozoa with a conservatively estimated 4921 extant described species (Bock and Gordon, 2013), or 83% of all Bryozoa. Bryozoans are ecologically important habitat builders (Wood et al., 2013) and are vital components of the marine food chain (Lidgard, 2008). Despite important analyses of regional species distributions (Clarke and Lidgard, 2000; Gappa, 2000; Barnes and Griffiths, 2008; Hirose, 2017; Boonzaaier-Davids et al., 2020; Denisenko, 2020), their global species richness distribution has never been quantified. We argue that even with concerns about the incompleteness of public records for the purpose of inferring regional to global diversity patterns (e.g. Klein et al. 2019; Chollett and Robertson 2020; Moudrý and Devillers 2020; Hughes et al. 2021), it is worth exploring cheilostome data in such public databases. We do so in order to identify spatial gaps in sampling but also to ask if automated information retrieval can enhance the species occurrence data available in public databases, specifically OBIS and GBIF (henceforth shortened as DB).

Automated information retrieval

Automated information retrieval (Hirschberg and Manning, 2015) is one recent approach to complement the time-consuming manual activity of data compilation from the scientific literature. Automated text-mining is well-established in the biomedical realm (Percha et al., 2012; Christopoulou et al., 2020), but has only recently been adopted for biodiversity studies (Peters et al., 2017; Kopperud et al., 2019). As far as we are aware, automated text-mining has never been applied to the literature for extraction of taxon occurrences in given locations for the purpose of understanding biogeography (but see Page 2019). We use natural language processing tools (De Marneffe et al., 2014; Bojanowski et al., 2017), to compile cheilostome text-mined occurrence data (TMO) for comparisons with data from DB.

Taxon occurrence

Taxon occurrence data from DB and TMO are not expected to be the same. We ask if they could, separately or in combination, shed light on a long-standing biogeographic hypothesis in the bryozoological literature. Many different groups of organisms show the canonical LDG, a species richness peak in tropical regions and decreasing species richness towards the temperate and polar zones (Hillebrand, 2004; Menegotto et al., 2019). Despite being common across marine and terrestrial realms, and among diverse eukaryote clades, the LDG is not universal (Chaudhary et al., 2021). Marine extratropical bimodal species richness peaks have been observed, for example in brittle stars (Woolley et al., 2016), polychaetes (Pamungkas et al., 2021), crustaceans (Rivadeneira and Poore, 2020), fishes (Lin et al., 2021), and brown macroalgae (Fragkopoulou et al., 2022), among other groups. Bimodality has also been suggested for cheilostome bryozoans (Schopf, 1970; Clarke and Lidgard, 2000; Barnes and Griffiths, 2008).

The TMO and DB data in combination support the view that the latitudinal diversity pattern of living cheilostomes is bimodal. These data reveal highest levels of estimated species richness in temperate latitudes, but TMO species richness has a peak in the northern hemisphere while DB has a peak in the temperate south. Moreover, two datasets differ significantly in the geographic richness patterns in Atlantic versus Pacific ocean basins (Schopf, 1970; Barnes and Griffiths, 2008). We discuss the pros and cons of TMO and public databases such as OBIS and GBIF and how their differences can help us understand the uncertainties of the retrieved spatial diversity patterns, beyond what is estimated within the confines of each dataset.

METHODS

DB Data Retrieval

We use the R-package robis (Provoost and Bosch, 2021) to access OBIS, and the web interface of GBIF to retrieve latitude/longitude occurrence records of cheilostomes (both accessed on 21.02.2022). We remove records without species epithets. For taxonomic ambiguities such as cf., aff., we disregard the uncertainty; for instance, *Microporella* cf. *ciliata* becomes *Microporella ciliata*. Records with genus names that are not accepted according to either the Working List of Genera and Subgenera for the Treatise on Invertebrate Paleontology (pers comm. Dennis P. Gordon, 2019), World Register of Marine Species (WoRMS Editorial Board, 2022) or www.bryozoa.net (Bock, 2022) are also removed. For all unaccepted species names that are found in WoRMS, we translate the species name to the accepted species name according to WoRMS. We also drop all Linnean binomials that are not found in WoRMS. The result is 831 unique genus names and 4549 unique genus-species combinations (henceforth simply species) in 149042 retained OBIS and GBIF records.

TMO (Text-Mined Occurrence) Data Retrieval

We follow a previously detailed text-mining procedure (Kopperud et al., 2019) with modifications. We extract text from two collections of published works, our own corpus (3233 pdf documents) and the GeoDeepDive archive (GDD, https://geodeepdive.org/), which contains full-text contents of journal articles. Only English language publications and those likely to feature extant bryozoans were used for information extraction (see Appendix S1 in Supporting Information).

We use CoreNLP (Manning et al., 2014) for an initial natural language analysis prior to information extraction, including tokenization, named-entity recognition, and dependency grammar annotation (Hirschberg and Manning, 2015). We use a pre-trained machine-learning model to recognize location names in the text (Finkel et al., 2005). To facilitate extraction of species, we compile names from the Working List of Genera and Subgenera for the Treatise on Invertebrate Paleontology (pers comm. Dennis P. Gordon, 2019), World Register of Marine Species (WoRMS Editorial Board, 2022) and www.bryozoa.net (Bock, 2022) that we then use in rule-based recognition (Chang and Manning, 2014). For example, consider a sentence from Tilbrook et al. (2001, p. 50): “The avicularia resemble those seen in *B. intermedia* (Hincks, 1881b), from Tasmania and New Zealand, but this species is only just over half the size of *B. cookae*.”

This sentence contains two species names (“*B. intermedia*” and “*B. cookae*”) and two location names (“Tasmania” and “New Zealand”). Each species-location pair is a candidate relation. The sentence implies that *B. intermedia* is found in New Zealand (a positive relation), but does not say anything about where *B. cookae* is found (a negative relation). We automate this distinction using a machine-learning classifier that we trained using a dataset of 4938 unique candidates labelled as positive or negative by two persons. Part of our procedure resolves the genus name referred to as ‘*B.*’ above (see Appendix S1).

We use a test data set comprising 10% of the labelled candidates to evaluate several aspects of our machine-classifier: (i) accuracy, the ratio of correct predictions to all predictions; (ii) precision, the ratio of true positive predictions to all positive predictions; (iii) recall, the ratio of true positive predictions to all positive labels; (iv) false positive rate (FPR), the ratio of false positive predictions to all negative labels; and (v) F1, the harmonic mean of precision and recall. Each of these metrics yields different information on the reliability of the extracted data. We treat taxonomic ambiguities within TMO data in the same manner as OBIS and GBIF records (see previous section).

From TMO location names to spatial data

Location names (e.g., New Zealand, Tasmania) are submitted to the Google geocoding service (`https://developers.google.com/maps/documentation/geocoding/`) to acquire a bounding box with four latitude-longitude coordinates and a centroid, based on Google’s defaults (Fig. S1). We remove species occurrences in locations represented by bounding boxes that are larger than about 2% of the Earth’s surface using area calculations assuming a spherical globe. See Fig. S2 for how the bounding box sizes are distributed, and Fig. S3 for how alternative thresholds impact the results.

Estimating latitudinal species richness

We initially evaluate species richness in thirty-six 5° latitudinal bands using two standard richness estimators that perform relatively well under a suite of conditions (Walther and Moore, 2005): Chao2 and Jackknife using the function specpool in the R package vegan (Oksanen et al., 2015). We treat these latitudinal bands as independent. We then repeat the procedure using thirty-six equal area bands, since areas represented within equal angle bands decrease poleward. To apply these estimators, we divide each (equal angle or area) latitudinal band into 5° longitudinal sampling units. We use the bias-corrected form of Chao2 $= S_{\text{obs}} + Q_1^2(N-1)/(2NQ_2)$, and incidence-based Jackknife = $S_{\text{obs}} + Q_1(N-1)/N$. Here, S_{obs}

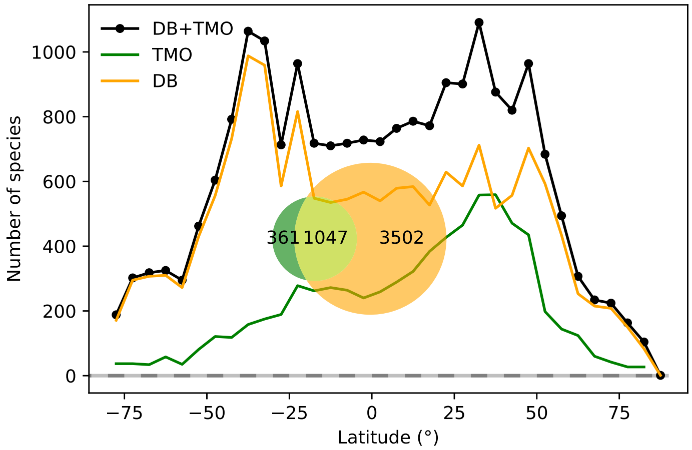

Figure 1. Global range-through latitudinal species richness for cheilostome bryozoans. The black line shows combined database (DB = OBIS and GBIF) and text-mined occurrence (TMO) richness, and orange and green curves show range-through richness for DB and TMO separately. The inset is a Venn diagram showing the global overlap in species between DB and TMO.

is the number of observed species in each band, N is the number of (longitudinal) sampling units, Q_1 is the number of species observed in only one sampling unit, and Q2 is the number observed in two sampling units. Note that we measure incidence as whether a species is either observed or not observed, for each geographical unit. Hence, any duplicate records in OBIS and GBIF do not inflate richness. Because terrestrial regions are not suitable habitats for marine cheilostomes, we mapped all landlocked longitudinal sampling bins (Fig. S4) based on a 1:10 m map of global coastlines (Patterson, 2019). We removed the landlocked bins prior to richness estimation. For DB data where spatial coordinates are points, it is trivial to assign data to sampling units. For TMO, we assume that a species occurs in all of the sampling units that intersect the bounding box associated with the location. TMO bounding boxes vary in size, but most are smaller in area than our sampling units (Fig. S2).

In addition to Chao2 and Jackknife estimators, we also determined range-through species richness. Here, we assume that a species spans its southernmost and northernmost occurrence record, regardless of whether it is observed in any intermediate latitudinal band. We acknowledge that all richness estimators, including the ones we chose, have different limitations (Gwinn et al., 2016). Confidence that the inferred patterns are ”real” is improved by the extent that different estimators making different assumptions yield consistent results.

RESULTS

Capturing species diversity: comparing DB and TMO

Applying the text-mining procedure to our corpora, we retrieved 1408 species in 343 genera, and 1400 unique location names among 7204 TMO records. Only 23% of the species in the DB data that we retained were also in TMO. On the other hand, 68% of species in the TMO occurred in DB. 21% of the species richness is common to both (Fig. 1). In combination with DB data, we have species-location information from 4910 species, almost tallying with the 4921 described cheilostome species (Bock and Gordon, 2013).

Our machine-classifier achieved an accuracy of 73.1%, F1 of 76.8%, recall of 78.9%, FPR of 34.3% and precision of 74.8% as estimated with our test set (Fig. S6b). These results are substantially better than a random classifier baseline, but not as good as the human annotator repeatability. Specifically, the FPR among annotators is about 15% (n = 200). A random classifier that is as unbalanced as our training data (60% positive labels) would yield 60% false positives, but a random classifier equaling our classifier’s recall of 78.9% would have the same false positive rate of 78.9% (see Appendix S1).

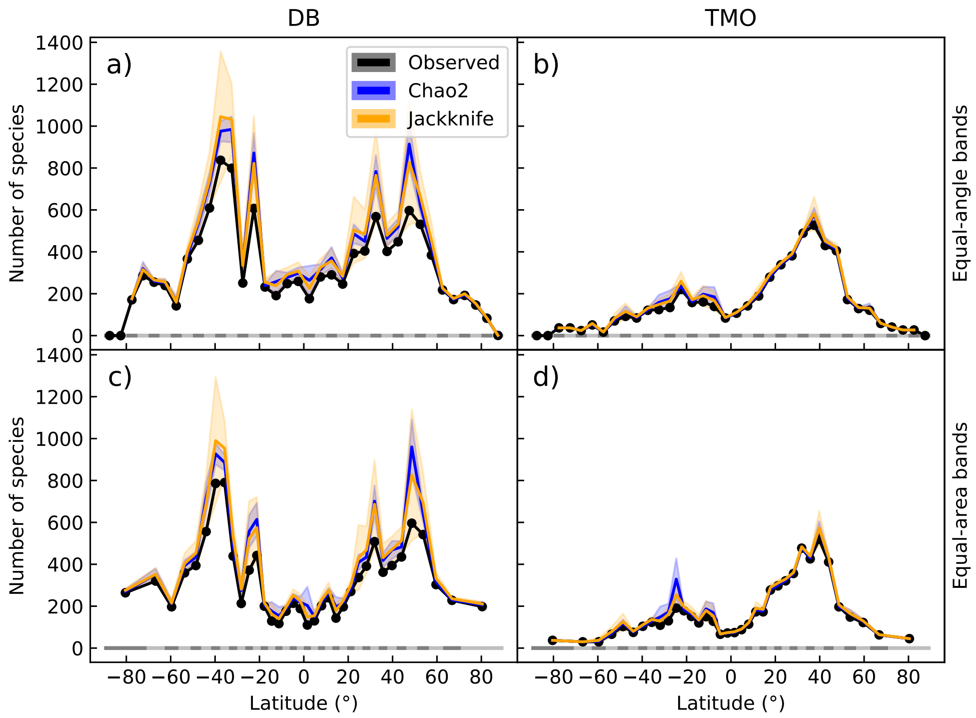

Figure 2. Global latitudinal species richness for cheilostome bryozoans, estimated using Chao2 and Jackknife. The top panels show richness for database (DB = OBIS and GBIF) and text-mined occurrences (TMO) data in 5° equal-angle latitudinal bands. The lower panels show the equivalent in 5° equal-area latitudinal bands. Black lines show the observed richness, while blue and orange lines show the Chao2 and Jackknife estimates, respectively. The shaded areas are 95% confidence intervals. See Figs. S7 and S8 for alternative band and bin sizes.

Latitudinal species richness patterns

Combined TMO and DB data in plots of range-through species richness show a bimodal pattern with species richness peaks in both hemispheres surrounding 40° and -40° (Fig. 1). Inferred species richness in both of these peaks is about double that in the tropics (Fig. 2). The two data sources contribute different latitudinal constituents, as suggested by the limited overlap in their species composition (Fig. 1 inset).

Chao2 and Jackknife estimated species richness from DB shows two peaks between -20° and -45° that are more than double the next highest peak between 25° and 50° (Fig. 2a). In contrast, TMO estimated richness shows a highest peak between 30° and 45° (Fig. 2b). With minor exceptions in the Antarctic where spatial distortion is largest, equiangular and equi-areal bands yield nearly identical inferences (compare Fig. 2a,c). The latitudinal pattern appears smoother when using larger latitudinal band sizes (Fig. S7), while retaining a qualitatively similar picture. Longitudinal sampling bins of varied sizes did not lead to notable variation for the Jackknife and Chao2 estimators (Fig. S8).

The northern hemisphere peak in richness (Fig. 1) reflects TMO records from the Mediterranean and Japan, but also from the Atlantic Ocean (Fig. 3a,e), including the British Isles. Note that we did not include the Mediterranean as part of the Atlantic basin for Fig. 3. A portion of the TMO data are spatially imprecise, for example the location names "France", "Spain" or "Morocco" may be associated with Mediterranean endemics, yet these records could contribute to the Atlantic richness counts in Fig. 3. The spatially precise DB data show a much lower peak in the Eastern Atlantic (Fig. 3e, orange line shifted slightly northward), reflecting data from the British Isles and northern Europe. Conversely, DB data mainly from Australia and New Zealand contribute disproportionately to the huge southern hemisphere peak. The richness captured by DB in Australia and New Zealand is not reflected by TMO species

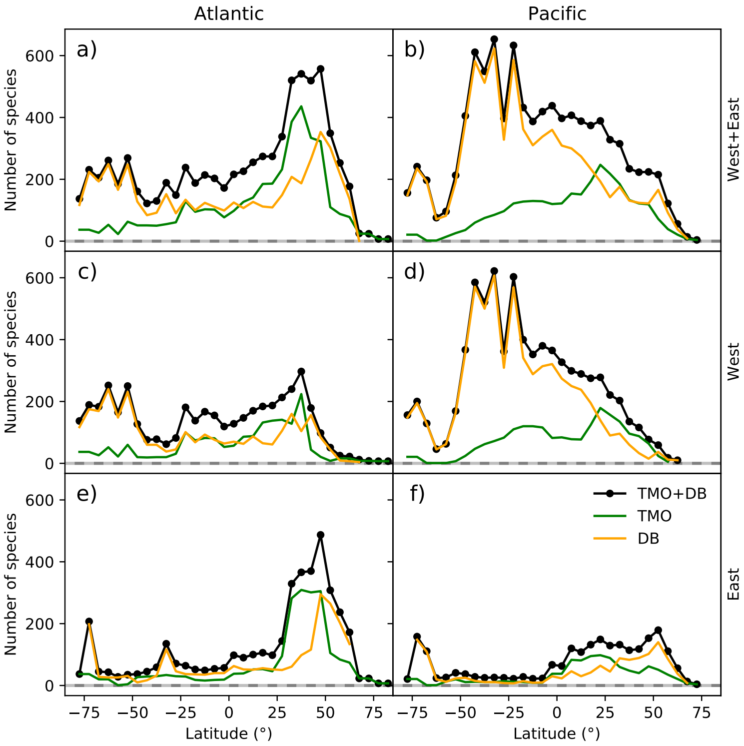

Figure 3. Range-through latitudinal species richness for cheilostome bryozoans in the Atlantic and Pacific Oceans. The left column shows species richness in the Atlantic, and the right column shows that in the Pacific. The panel rows represent the eastern, western or the entire ocean basins. Orange and green lines represent database (DB = OBIS and GBIF) and text-mined occurrences (TMO), respectively, and black lines are the joint data. Note that in this figure, the Atlantic borders Greenland and Iceland in the north, and the Antarctic in the south, but does not include the Gulf of Mexico, the Caribbean, the Baltic Sea or the Mediterranean. The Pacific borders the Bering Strait in the north, and includes the South China Sea, the Java Sea, north and east Australia, Tasmania as well as the Antarctic border.

richness (Fig. 3b,d). The western Atlantic and eastern Pacific do not display such pronounced temperate zone peaks (Fig. 3c,f). Looking at individual ocean basins, TMO and DB are sometimes congruent and other times incongruent. For example, there is an absence of DB records in Japanese waters, and there are similarly few TMO and DB records in the Indian Ocean (Fig. 4).

Such varied regional species richness patterns are in part influenced by the geographic occurrence of samples. Figure 4 summarizes the relative distribution of species-location records for TMO and DB data as global heatmaps. For DB data, there are about one order of magnitude fewer records in tropical regions than for subtropical and temperate ones (Fig. S9a). While there are also fewer TMO records in tropical regions, the effect is not as pronounced (Fig. S9b). Northern and southern hemisphere species richness peaks in the two data sets (Fig. 1) correspond with high regional densities of TMO and DB records, respectively (Fig. 3e,d).

DISCUSSION

Causal hypotheses for a LDG and contrarian patterns are plentiful (Rivadeneira and Poore, 2020; García Molinos and Alabia, 2021). Such hypotheses can sometimes be tested in groups with rich and relatively unbiased spatial data from both extant and extinct taxa (Jablonski et al., 2006; Krug et al., 2007; Jablonski et al., 2013) or those with independent molecular phylogenetic evidence (Rabosky et al., 2018). We believe ours is the first study to quantify global cheilostome species biogeographic patterns. Using a combined TMO and DB perspective, and a bimodal latitudinal diversity gradient in cheilostome species richness is quite apparent. Yet, at present, we can merely speculate about what processes that may

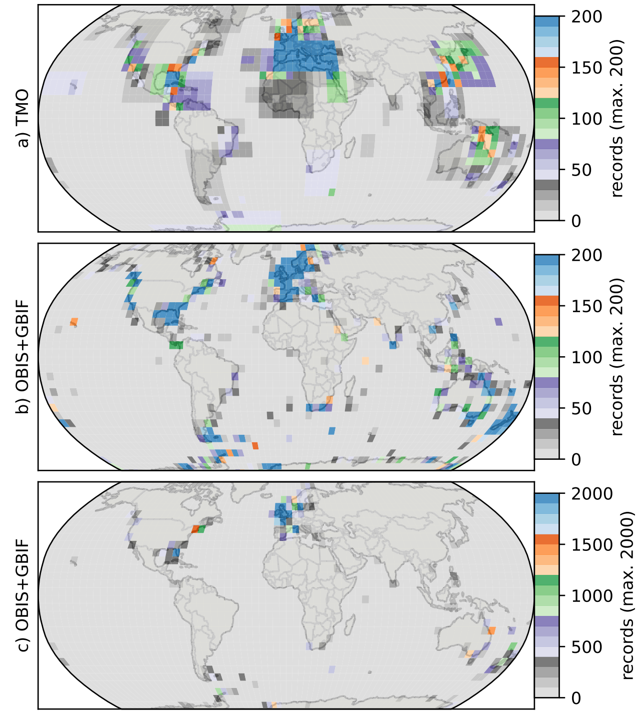

Figure 4. Heatmaps for cheilostome bryozoan occurrence records per 5° latitude by 5° longitude bins. The color axes are truncated for visualization purposes, to a maximum of 200, 200 and 2000 in a), b), c), respectively. There are about 900 maximum records per bin in the Mediterranean for the text-mined occurrences (TMO), and about 66000 maximum records in the British Isles for the Ocean Biogeography Information System (OBIS) and Global Biodiversity Information Facility (GBIF) data combined. The globe is plotted using the Robinson projection. See Fig. S11 for the same figure plotted using the plate carrée projection.

have led to their latitudinal pattern. Given the biases and heterogeneity of the data we explored which are striking when comparing our two data sources, we also need to consider (i) how this pattern coincides with previous observations, and (ii) methodological, sampling, and taxonomic concerns.

Two patterns in our analyses are similar to Schopf's (1970) findings from then-scarce available data: higher species richness on the eastern margin of the Atlantic and the western margin of the Pacific compared to their opposite margins, and increasing richness with latitude away from the equator. Our combined data conforms with the first finding, but still does not capture the richness of the severely-understudied Philippine-Indonesian region and its many archipelagoes (Okada and Mawatari, 1953; Gordon, 1999; Tilbrook and De Grave, 2005). Changes to the second finding are more nuanced, and may partly reflect relatively lower equatorial sampling density (Menegotto and Rangel, 2018) apparent in both the datasets (Fig. S9). However, our observed peaks of species richness are at significantly higher latitudes than those reported for bryozoans in Chaudhary et al. (2016).

Fossil and modern patterns of bryozoan skeletal abundance in cool-water carbonate sediments suggest that the lower tropical species richness is not merely a sampling artifact. Modern bryozoan-dominated carbonate platforms are far more common on cool-water temperate shelves than on tropical ones (Schlanger and Konishi, 1975; James and Clarke, 1997). Cenozoic tropical bryozoan faunas are both less abundantly preserved and less diverse than those from temperate latitudes, possibly reflecting biotic interactions, preservational biases, and cryptic existence in shallower-water habitats dominated by corals, calcareous algae, and other photobiont organisms (Winston, 1986; Taylor and Di Martino, 2014). A far-reaching study by Taylor and Allison (1998) showed that 94% of bryozoan-rich post-Paleozoic sedimentary deposits formed outside of the paleotropics, which may be especially significant if regional species richness and skeletal abundance are linked. About a third of all described bryozoan species occur south of -30°, and 87% of these are cheilostomes (Barnes and Griffiths, 2008).

We chose to discretize the data in latitudinal bands and longitudinal bins that are larger than those used previously (e.g. Rabosky et al., 2018). The choice of band- and bin sizes for species richness estimation is somewhat arbitrary. Differing choices suggest quantitatively dissimilar inferences, although the bimodality is still apparent in the cases we have explored (Figs. S7 and S8). A range-through latitudinal diversity approach (Fig. 3) assumes that any species that is not observed in a gap between two adjacent latitudinal bands should contribute to species richness in that gap, but this assumption is quite easily broken (Menegotto and Rangel, 2018). The bounding boxes used for TMO locations may also tend to bleed range margins as opposed to DB point location data. Richness estimates may be inflated via range-through estimates, particularly in the tropics, compared to estimating richness independently in each latitudinal band which yields lower estimates (Fig. 2). Regardless, both methods for estimating species richness give a picture of bimodality.

Global biogeographic studies such as ours are more prone to the issues of sampling and taxonomic concerns than local or regional ones, simply due to their scope. Large sampling gaps are apparent in both TMO and DB datasets. The development and application of richness estimation models that distinguish true absences from non-observations (Iknayan et al., 2014) may help improve inferences, but are likely insufficient to fully overcome acute sampling gaps. Overall, there are relatively few records in the Indian Ocean, most of the South Atlantic, and eastern margin of the Pacific. TMO records for the Arctic are sparse, as are OBIS records for the northwest Pacific. Aside from a few extreme outliers from DB British Isles locations, species richness and number of records per 5° latitudinal band have a strong positive relationship (Fig. S10). Independent taxonomic surveys of underrepresented regions in one or both datasets corroborate the existence of significant gaps (Gappa, 2000; Barnes and Griffiths, 2008; Liu, 2008; Vieira et al., 2008; Hirose, 2017; Boonzaaier-Davids et al., 2020; Denisenko, 2020; Sanjay et al., 2020). The DB records may partly reflect recent histories of active bryozoan research programs in the Antarctic (Barnes and Griffiths, 2008) and Australia and New Zealand (Wood et al., 2013) as well as contributions to OBIS and GBIF that differ substantially among research institutions. On the other hand, TMO extracted extensive species-location information from the Mediterranean (27° to 50°) that are severely wanting in OBIS, demonstrating that combining disparate data sources can help bridge gaps in global biogeographic studies.

Taxonomic errors inevitably exist in large databases. Taxonomy is continuously subject to revisions (Bock and Gordon, 2013), not all of which are accounted for in our datasets. Many species await description; Gordon et al. (2019) suggest that there are over 6,400 'known' cheilostome species without commenting on nomenclatural status, suggesting that there are up to 600 'known' species that need naming. Yet, a recent study, based on bryozoans, comparing datasets with taxonomic synonyms and without, found that synonymization does not contribute to qualitative changes in broad scale inferences Lidgard et al. (2021). Our machine-classifier is currently unable to extract location information for 21% of the species that were detected in our corpus of published works (Fig. S5). Our conversion of taxonomic ambiguities into certainties likely deflated species richness estimates, while mistaken inclusion of fossil species names may have inflated richness estimates. We have assumed these do not necessarily introduce spatial bias. Additionally, many bryozoan species determined by traditional morphological methods may actually consist of unrecognized species complexes (Lidgard and Buckley, 1994; Fehlauer-Ale et al., 2014), although cheilostome bryozoan species are perhaps unusually delimitable using morphological information preserved in the skeleton (Jackson and Cheetham, 1990).

While the portion of TMO data that is derived from the taxonomic literature may be less plagued by taxonomic misidentifications, the same cannot be easily argued for faunal lists or ecological surveys, much of which DB data is based on. However, in our experience, broad inferences based on synoptic, large-scaled databases tend to change significantly with different models, more so than data updates (Sepkoski, 1993; Liow et al., 2015; Lidgard et al., 2021).

In terms of our text-mining task, we found that generating and classifying species-location candidates here is more challenging than classifying species-age candidates (Kopperud et al., 2019). An F1 result of about 77.5% is not uncommon for relation extraction studies (Kim et al., 2019; Henry et al., 2020), especially for datasets with low label assignment repeatability. Nonetheless, while the accuracy of the machine-classifier is less sensitive than human evaluation, its FPR is substantially lower than a null model. Note that the classifier merely provides a probabilistic measure of whether the sentence provides evidence that a species is present at a geographic location. In the event of a false positive, it is still possible that the species is actually present in that particular location. On the other hand, there is a wealth of species mentions for which we were not able retrieve any species-location candidates (Fig. S5). It is possible to extend our approach by considering cross-sentence candidates (Gupta et al., 2019), although these methods are usually less accurate. Alternatively, we could go beyond standard NLP tools, which are relatively flexible and easy to adopt, and use non-linguistic features (such as tables and spatial layout common in primary diversity publications e.g. (Rosso and Sanfilippo, 2000; Gordon, 2016)) for information extraction, as has been suggested in the knowledge base creation literature (Schlichtkrull et al., 2018). However, such methods for information extraction that combine linguistic and non-linguistic features are still at an early stage of development.

The main advantage of automatic information retrieval over collaborative data-entry is that of reduced time and resource investment. The information retrieval procedure is largely independent of the size of the literature, or the taxonomic scope, say for cheilostomes versus all metazoans. Public biodiversity inventories such as GBIF and OBIS require large consortia and networks of research factions to contribute their data. Conversely, there is a wealth of biodiversity knowledge available in the published literature, and it is feasible for one person or a small team to extract substantial amounts of data quickly using automated information retrieval. We have used some supervised classification methods, which require us to generate training data. However as NLP is adopted in the biodiversity literature, it will become easier to use distantly supervised relation extraction (Hirschberg and Manning, 2015).

Biodiversity inventories such as OBIS and GBIF are vital for supplying data for inferences of global biogeographic patterns. While we strongly support the continued development of these databases, we demonstrated that our automated information retrieval approach can enhance such inventories when answering global-scale questions, especially for under-studied taxa. To understand how the spatial diversity of cheilostomes has come to be will require continued and concerted efforts in taxonomic investigations (Bock and Gordon, 2013), compilation of more spatial data especially in areas currently devoid of deposited information (Klein et al., 2019), tool-development in automated data retrieval (Kopperud et al., 2019), and continued research in molecular phylogenetics (Orr et al., 2021).

SUPPORTING INFORMATION

- Appendix S1: Extended methods and supplementary figures.
- Appendix S2: Bibliographic references for TMO data.

The code and data required to reproduce the analyses and figures and will be available at zenodo.org.

ACKNOWLEDGEMENTS

We thank Mali H. Ramfjell for compiling part of our training dataset, the GeoDeepDive group, especially Ian Ross and Shanan Peters, for providing access to articles, and OBIS and GBIF and their contributors for their georeferenced taxonomic data. We thank Phil Bock for maintaining bryozoa.net, Dennis Gordon for an updated version of the Working List of Genera and Subgenera for the Treatise on Invertebrate Paleontology, and both for their contributions to WoRMS. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 724324 to L.H. Liow).

REFERENCES

Barnes, D. K. and Griffiths, H. J. (2008). Biodiversity and biogeography of southern temperate and polar bryozoans. *Global Ecology and Biogeography*, 17(1):84–99.

Bock, P. (2022). Recent and fossil bryozoa. `http://www.bryozoa.net`. Accessed: 2022-03-04.

Bock, P. E. and Gordon, D. P. (2013). Phylum Bryozoa Ehrenberg, 1831. *Zootaxa*, 3703(1):67–74.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boonzaaier-Davids, M., Florence, W., and Gibbons, M. (2020). Zoogeography of marine Bryozoa around South Africa. *African Journal of Marine Science*, 42(2):185–198.

Chang, A. X. and Manning, C. D. (2014). Tokensregex: Defining cascaded regular expressions over tokens. *Stanford University Computer Science Technical Reports. CSTR*, 2:2014.

Chaudhary, C., Richardson, A. J., Schoeman, D. S., and Costello, M. J. (2021). Global warming is causing a more pronounced dip in marine species richness around the equator. *Proceedings of the National Academy of Sciences*, 118(15):e2015094118.

Chaudhary, C., Saeedi, H., and Costello, M. J. (2016). Bimodality of latitudinal gradients in marine species richness. *Trends in Ecology & Evolution*, 31(9):670–676.

Chollett, I. and Robertson, D. R. (2020). Comparing biodiversity databases: Greater Caribbean reef fishes as a case study. *Fish and Fisheries*, 21(6):1195–1212.

Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M., and Ananiadou, S. (2020). Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46.

Clarke, A. and Lidgard, S. (2000). Spatial patterns of diversity in the sea: bryozoan species richness in the North Atlantic. *Journal of Animal Ecology*, 69(5):799–814.

Costello, M. J., Tsai, P., Wong, P. S., Cheung, A. K. L., Basher, Z., and Chaudhary, C. (2017). Marine biogeographic realms and species endemicity. *Nature communications*, 8(1):1–10.

De Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

Denisenko, N. (2020). Species richness and the level of knowledge of the bryozoan fauna of the Arctic region. *Proceedings of the Zoological Institute RAS*, 324:353–363.

Edgar, G. J., Alexander, T. J., Lefcheck, J. S., Bates, A. E., Kininmonth, S. J., Thomson, R. J., Duffy, J. E., Costello, M. J., and Stuart-Smith, R. D. (2017). Abundance and local-scale processes contribute to multi-phyla gradients in global marine diversity. *Science Advances*, 3(10):e1700419.

Fehlauer-Ale, K. H., Mackie, J. A., Lim-Fong, G. E., Ale, E., Pie, M. R., and Waeschenbach, A. (2014). Cryptic species in the cosmopolitan *Bugula neritina* complex (Bryozoa, Cheilostomata). *Zoologica Scripta*, 43(2):193–205.

Finkel, J. R., Grenager, T., and Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370.

Fragkopoulou, E., Serrão, E. A., De Clerck, O., Costello, M. J., Araújo, M. B., Duarte, C. M., Krause-Jensen, D., and Assis, J. (2022). Global biodiversity patterns of marine forests of brown macroalgae. *Global Ecology and Biogeography*, n/a(n/a).

Gappa, J. L. (2000). Species richness of marine Bryozoa in the continental shelf and slope off Argentina (South-West Atlantic). *Diversity and distributions*, 6(1):15–27.

García Molinos, J. and Alabia, I. (2021). *Marine Biogeography*, page 245–273. John Wiley & Sons, Ltd.

Gordon, D. (2016). Bryozoa of the South China Sea - an overview. *Raffles Bulletin of Zoology*, 34:604–618.

Gordon, D. P. (1999). Bryozoan diversity in New Zealand and Australia. *The other 99%. The Conservation and Biodiversity of Invertebrates*, 99:199–204.

Gordon, D. P., Bock, P. E., Souto-Derungs, J., and Reverter-Gil, O. (2019). A bryozoan tale of two continents: faunistic data for the Recent Bryozoa of Greater Australia (Sahul) and Zealandia, with European comparisons. *Australasian Palaeontological Memoirs*, pages 13–22.

Grenié, M., Berti, E., Carvajal-Quintero, J., Dädlow, G. M. L., Sagouis, A., and Winter, M. (2022). Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. *Methods in Ecology and Evolution*, n/a(n/a).

Gupta, P., Rajaram, S., Schütze, H., and Runkler, T. (2019). Neural relation extraction within and across sentence boundaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6513–6520.

Gwinn, D. C., Allen, M. S., Bonvechio, K. I., V. Hoyer, M., and Beesley, L. S. (2016). Evaluating estimators of species richness: the importance of considering statistical error rates. *Methods in Ecology and Evolution*, 7(3):294–302.

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., and Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, 118(6):e2018093118.

Henry, S., Buchan, K., Filannino, M., Stubbs, A., and Uzuner, O. (2020). 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Hillebrand, H. (2004). On the generality of the latitudinal diversity gradient. *The American Naturalist*, 163(2):192–211.

Hirose, M. (2017). Diversity of freshwater and marine bryozoans in japan. *Species diversity of animals in Japan*, pages 629–649.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., Yang, Q., Zhu, C., and Qiao, H. (2021). Sampling biases shape our view of the natural world. *Ecography*, 44(9):1259–1269.

Iknayan, K. J., Tingley, M. W., Furnas, B. J., and Beissinger, S. R. (2014). Detecting diversity: emerging methods to estimate species diversity. *Trends in Ecology & Evolution*, 29(2):97–106.

Jablonski, D., Belanger, C. L., Berke, S. K., Huang, S., Krug, A. Z., Roy, K., Tomasovych, A., and Valentine, J. W. (2013). Out of the tropics, but how? Fossils, bridge species, and thermal ranges in the dynamics of the marine latitudinal diversity gradient. *Proceedings of the National Academy of Sciences*, 110(26):10487–10494.

Jablonski, D., Roy, K., and Valentine, J. W. (2006). Out of the tropics: evolutionary dynamics of the latitudinal diversity gradient. *Science*, 314(5796):102–106.

Jackson, J. B. and Cheetham, A. H. (1990). Evolutionary significance of morphospecies: a test with cheilostome Bryozoa. *Science*, 248(4955):579–583.

James, N. and Clarke, J., editors (1997). *Cool-Water Carbonates*. Society for Sedimentary Geology, Tulsa, Oklahoma.

Kim, J., Kim, J.-J., and Lee, H. (2019). Digchem: Identification of disease-gene-chemical relationships from medline abstracts. *PLoS Computational Biology*, 15(5):e1007022.

Klein, E., Appeltans, W., Provoost, P., Saeedi, H., Benson, A., Bajona, L., Peralta, A. C., and Bristol, R. (2019). OBIS infrastructure, lessons learned, and vision for the future. *Frontiers in Marine Science*, 6:588.

Kopperud, B. T., Lidgard, S., and Liow, L. H. (2019). Text-mined fossil biodiversity dynamics using machine learning. *Proceedings of the Royal Society B*, 286(1901):20190022.

Krug, A. Z., Jablonski, D., and Valentine, J. W. (2007). Contrarian clade confirms the ubiquity of spatial origination patterns in the production of latitudinal diversity gradients. *Proceedings of the National Academy of Sciences*, 104(46):18129–18134.

Kusumoto, B., Costello, M. J., Kubota, Y., Shiono, T., Wei, C.-L., Yasuhara, M., and Chao, A. (2020). Global distribution of coral diversity: Biodiversity knowledge gradients related to spatial resolution. *Ecological Research*, 35(2):315–326.

Lidgard, S. (2008). Predation on marine bryozoan colonies: taxa, traits and trophic groups. *Marine*

Ecology Progress Series, 359:117–131.

Lidgard, S. and Buckley, G. A. (1994). 101 toward a morphological species concept in cheilostomates: phenotypic variation in *Adeonellopsis yarraensis* (waters). In *Biology and Palaeobiology of Bryozoans: Proceedings of the 9th International Bryozoology Conference, School of Biological Sciences, University of Wales, Swansea, 1992*, volume 9, page 101. Olsen & Olsen.

Lidgard, S., Di Martino, E., Zágoršek, K., and Liow, L. H. (2021). When fossil clades 'compete': local dominance, global diversification dynamics and causation. *Proceedings of the Royal Society B: Biological Sciences*, 288(1959):20211632.

Lin, H.-Y., Corkrey, R., Kaschner, K., Garilao, C., and Costello, M. J. (2021). Latitudinal diversity gradients for five taxonomic levels of marine fish in depth zones. *Ecological Research*, 36(2):266–280.

Liow, L. H., Reitan, T., and Harnik, P. G. (2015). Ecological interactions on macroevolutionary time scales: clams and brachiopods are more than ships that pass in the night. *Ecology Letters*, 18(10):1030–1039.

Liu, L. (2008). *Checklist of marine biota of China seas*. China: Institute of Oceanography, Chinese Academy of Science.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

McGill, B. J. (2019). The what, how and why of doing macroecology. *Global Ecology and Biogeography*, 28(1):6–17.

Menegotto, A., Kurtz, M. N., and Lana, P. d. C. (2019). Benthic habitats do show a significant latitudinal diversity gradient: A comment on Kinlock et al.(2018). *Global Ecology and Biogeography*, 28(11):1712–1717.

Menegotto, A. and Rangel, T. F. (2018). Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness. *Nature communications*, 9(1):1–6.

Moudrý, V. and Devillers, R. (2020). Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. *Ecological Informatics*, 56:101051.

Okada, Y. and Mawatari, S. (1953). Distributional provinces of marine Bryozoa in the Indo-Pacific region. *Int. Proc. 8th Pac. Sci. Congr*, 3:391–402.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H. (2015). Vegan: community ecology package. 2015. *R package version*, 2(10).

Orr, R. J. S., Di Martino, E., Gordon, D. P., Ramsfjell, M. H., Mello, H. L., Smith, A. M., and Liow, L. H. (2021). A broadly resolved molecular phylogeny of new zealand cheilostome bryozoans as a framework for hypotheses of morphological evolution. *Molecular Phylogenetics and Evolution*, 161:107172.

Page, R. D. (2019). Ozymandias: a biodiversity knowledge graph. *PeerJ*, 7:e6739.

Pamungkas, J., Glasby, C., and Costello, M. (2021). Biogeography of polychaete worms (Annelida) of the world. *Marine Ecology Progress Series*, 657:147–159.

Patterson, T. (2019). Free vector and raster map data. `http://www.naturalearthdata.com/`. Accessed: 2019-11-27.

Percha, B., Garten, Y., and Altman, R. B. (2012). Discovery and explanation of drug-drug interactions via text mining. In *Biocomputing 2012*, pages 410–421. World Scientific.

Peters, S. E., Husson, J. M., and Wilcots, J. (2017). The rise and fall of stromatolites in shallow marine environments. *Geology*, 45(6):487–490.

Provoost, P. and Bosch, S. (2021). *robis: Ocean Biodiversity Information System (OBIS) Client*. R package version 2.8.2.

Rabosky, D. L., Chang, J., Cowman, P. F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T. J., Coll, M., Alfaro, M. E., et al. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, 559(7714):392–395.

Rivadeneira, M. and Poore, G. (2020). *Latitudinal gradient of diversity of marine crustaceans: towards a synthesis*, volume 8, page 389–412. Oxford University Press.

Rosso, A. and Sanfilippo, R. (2000). *Shallow-water bryozoans and serpuloideans from the Ross Sea (Terra Nova Bay, Antarctica)*, page 515–525. Springer-Verlag, Berlin.

Sanjay, M., Venkatraman, C., Louis, S., and Shrinivaasu, S. (2020). *Cheilostomatous Bryozoa from West Coast of India*, page 73–88. Directorate of Public Relations and Publications, Kochi, India.

Schlanger, S. and Konishi, K. (1975). The geographic boundary between the coral-algal and the bryozoanalgal limestone facies: a paleolatitude indicator. In *9th international geological congress of sedimentology, nice, theme*, volume 1, pages 187–190.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. (2018). Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Schopf, T. J. M. (1970). Taxonomic diversity gradients of ectoprocts and bivalves and their geologic implications. *Geological Society of America Bulletin*, 81(12):3765–3768.

Sepkoski, J. J. (1993). Ten years in the library: new data confirm paleontological patterns. *Paleobiology*, 19(1):43–51.

Taylor, P. D. and Allison, P. A. (1998). Bryozoan carbonates through time and space. *Geology*, 26(5):459–462.

Taylor, P. D. and Di Martino, E. (2014). Why is the tropical cenozoic fossil record so poor for bryozoans. *Studi Trentini di Scienze Naturali*, 94:249–257.

Tilbrook, K. J. and De Grave, S. (2005). A biogeographical analysis of Indo-West Pacific cheilostome bryozoan faunas. In *Bryozoan Studies 2004*, pages 351–360. CRC Press.

Tilbrook, K. J., Hayward, P., and Gordon, D. (2001). Cheilostomatous Bryozoa from Vanuatu. *Zoological Journal of the Linnean Society*, 131(1):35–109.

Vieira, L. M., Migotto, A. E., and Winston, J. E. (2008). Synopsis and annotated checklist of Recent marine Bryozoa from Brazil. *Zootaxa*, 1810(1):1–39.

Walther, B. A. and Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, 28(6):815–829.

Winston, J. E. (1986). An annotated checklist of coral-associated bryozoans. *American Museum Novitates*, pages 1–39.

Wood, A. C., Rowden, A. A., Compton, T. J., Gordon, D. P., and Probert, P. K. (2013). Habitat-forming bryozoans in new zealand: their known and predicted distribution in relation to broad-scale environmental variables and fishing effort. *PloS One*, 8(9):e75160.

Woolley, S. N., Tittensor, D. P., Dunstan, P. K., Guillera-Arroita, G., Lahoz-Monfort, J. J., Wintle, B. A., Worm, B., and O'Hara, T. D. (2016). Deep-sea diversity patterns are shaped by energy availability. *Nature*, 533(7603):393–396.

WoRMS Editorial Board (2022). World register of marine species (worms). `https://www.marinespecies.org`. Accessed: 2022-03-14.

Zizka, A., Carvalho, F., Calvente, A., Baez-Lizarazo, M., Cabral, A., Coelho, J., Colli-Silva, M., Fantinati, M., Fernandes, M., Ferreira-Araújo, T., Moreira, F., Santos, N., Santos, T., Santos-Costa, R. d., Serrano, F., Silva, A. d., Soares, A. d. S., Souza, P. d., Tomaz, E., Vale, V., Vieira, T., and Antonelli, A. (2020). No one-size-fits-all solution to clean gbif. *PeerJ*, 8:e9916.