PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Kopperud BT, Lidgard S, Liow LH. 2019 Text-mined fossil biodiversity dynamics using machine learning. *Proc. R. Soc. B* **286**: 20190022. http://dx.doi.org/10.1098/rspb.2019.0022

Received: 4 January 2019
Accepted: 2 April 2019

Subject Category:
Palaeobiology

Subject Areas:
palaeontology, evolution, bioinformatics

Keywords:
cheilostome bryozoans, fossil occurrences, palaeobiodiversity, natural language processing, information extraction, literature compilation

Authors for correspondence:
Bjørn Tore Kopperud
e-mail: b.t.kopperud@nhm.uio.no
Lee Hsiang Liow
e-mail: l.h.liow@ibv.uio.no

Electronic supplementary material is available online at http://dx.doi.org/10.6084/m9.figshare.c.4464296.

THE ROYAL SOCIETY
PUBLISHING

Text-mined fossil biodiversity dynamics using machine learning

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

[1]Natural History Museum, University of Oslo, PO Box 1172, Blindern, 0318 Oslo, Norway
[2]Integrative Research Center, Field Museum, 1400 South Lake Shore Drive, Chicago IL, 60605, USA
[3]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, PO Box 1066, Blindern, 0316 Oslo, Norway

BTK, 0000-0002-7360-7087; SL, 0000-0002-0446-4705; LHL, 0000-0002-3732-6069

Documented occurrences of fossil taxa are the empirical foundation for understanding large-scale biodiversity changes and evolutionary dynamics in deep time. The fossil record contains vast amounts of understudied taxa. Yet the compilation of huge volumes of data remains a labour-intensive impediment to a more complete understanding of Earth's biodiversity history. Even so, many occurrence records of species and genera in these taxa can be uncovered in the palaeontological literature. Here, we extract observations of fossils and their inferred ages from unstructured text in books and scientific articles using machine-learning approaches. We use Bryozoa, a group of marine invertebrates with a rich fossil record, as a case study. Building on recent advances in computational linguistics, we develop a pipeline to recognize taxonomic names and geologic time intervals in published literature and use supervised learning to machine-read whether the species in question occurred in a given age interval. Intermediate machine error rates appear comparable to human error rates in a simple trial, and resulting genus richness curves capture the main features of published fossil diversity studies of bryozoans. We believe our automated pipeline, that greatly reduced the time required to compile our dataset, can help others compile similar data for other taxa.

1. Introduction

How have scientists determined the history of biodiversity on our planet? The radiations of unicellular organisms, plants and animals, rates of diversification and extinction, correlation of past biodiversity levels with environmental forcing factors, mass extinctions and recoveries—all of these and more are ultimately reliant on, or at least calibrated by, published occurrences of fossil taxa and their geologic ages. For more than 200 years, palaeontologists have used their own work and mined the work of others to document taxonomic richness in successive geologic time intervals [1]. This effort has historically involved searching out and manually compiling published occurrences of lower-level taxa (e.g. species, genera, families) within a given higher-level group (e.g. order, class, phylum) and then inferring the changing patterns of richness [2,3].

Despite considerable progress in statistical methods that aim to compensate for occurrence gaps and known biases of the fossil record [4–8], we are still some ways away from a comprehensive understanding of the history of global biodiversity. One of the foremost impediments remaining is the costly, labour-intensive business of extracting fossil occurrence data from a continually expanding, scattered literature, even for palaeontological experts analysing one or another moderately taxon-rich group. Depending upon the extent and availability of previous fossil diversity studies on that group, this manual process could take months or even years to complete. Moreover, the methods and sources that underpin such compilations may or may not have been recorded and made available [9,10], potentially compromising scientific repeatability.

Hence, community efforts have built large public data compilations of taxonomic nomenclature or taxon occurrences, for instance, the Global Biodiversity Information Facility [11], World Register of Marine Species [12] and the Paleobiology Database (https://paleobiodb.org/). While one million records entered over 20 years and 300 credited publications attest to the impact of the Paleobiology Database, key data gaps persist and require sustained data gathering and input.

A promising approach to bridging data gaps is to partially automate the process of populating such data compilations, using information extraction techniques. Information extraction using natural language processing has become a mainstay in the biomedical field [13]. For instance, drug–drug interactions and mental illness symptoms are presented in new publications every month and such data extracted [14,15] from disparate journals can be used to further our understanding of human diseases and treatments. Text-mining approaches have also begun to impact knowledge discovery in biodiversity studies of living organisms [16,17]. With respect to palaeontological biodiversity studies, Peters *et al*. [18] pioneered automated information retrieval of fossil observations and their geologic time intervals. Here, we improve the pipeline proposed by Peters *et al*. [18]. We apply natural language processing tools [19,20] that aid in information retrieval by supervised learning [21–23]. Advantages of such a text-mining machine-learning approach include (i) enhanced efficiency in terms of time expended to find and retrieve data from the primary literature and the costs thereof, (ii) the ease of error estimation, and (iii) repeatability and transparency in terms of the source and nature of the data.

We show how our pipeline can be applied to organismal groups, especially those in which data are wanting in public databases, yet abundantly available in scattered publications. One such group is Bryozoa, a phylum of colonial animals in freshwater and more commonly in benthic marine environments. They are habitat constructors that enhance the biodiversity of other organisms [24], are instrumental in oceanic organic filtering and are major biogeochemical carbonate engines in the global carbon cycle [25]. Their rich fossil record and complex morphologies have also been critically important in testing macroevolutionary theory [26]. Despite several reviews of their fossil biodiversity history [27–29], publicly available data are lacking in a form that is amenable to relevant statistical analyses [5,8,30–32]. Our aim here is to build an occurrence database of observations of fossil cheilostomes (order Cheilostomatida [33]). The cheilostomes are the most species-rich group of bryozoans for which there are currently about 4800 known extant members [29]. We explicitly quantify both human and machine error in retrieving taxon names and their time intervals of occurrence. We then compare the inferred history of cheilostome bryozoans using our machine-read data to their established palaeobiodiversity pattern [29] and find that major features of their diversity changes are recovered across their 150-million-year history. Finally, we discuss the pros and cons of semi-automated text-mining techniques and suggest avenues for future improvements.

2. Material and methods

(a) Data sources

Public compendia such as Web of Science, Zoological Record and Google Scholar are primary sources for scientific publications from major publishers, especially those with a modern Web-presence. However, many key taxonomic publications appear in more obscure platforms such as museum publications, conference volumes, theses and governments reports, where accessibility through online sources is often limited. For this reason, we obtained bibliographic collections from experienced scientists working on bryozoans for much of their careers, including one of us (SL). PDFs (Portable Document Format) of publications that were likely to contain relevant species occurrence data from the Jurassic through the Holocene (including those in [27,28]) were then combined. This corpus consisted of over 10 000 bibliographic references, only some of which were likely to contain fossil species occurrence data. We filtered those with accompanying PDFs, yielding more than 2000 documents in the English language; more than 800 of these contained fossil occurrences that were useful for our text-mining pipeline (see below).

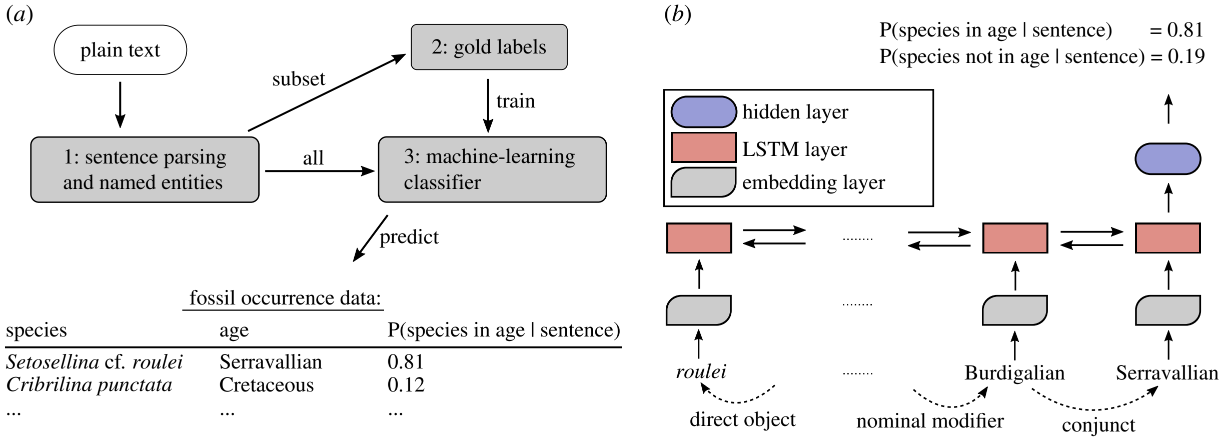
Recognizing words as bryozoan taxa or geologic time intervals entails reference lists of names. For bryozoan taxon names, we used the World Register of Marine Species [12] and the online taxonomic compendium of an experienced bryozoologist [34]. For geologic time-interval names, we used Macrostrat [35] and GeoWhen [36]. Publications documenting fossil occurrences include taxonomic monographs, systematic treatments and faunal lists. These publications contain formal taxonomic descriptions of fossils, discussions of the described or previously described taxa, taxon lists, summary tables or illustrations of taxa. For our text-mining procedure, we focused on mentions of taxonomic names in full sentences. In other words, we ignored lists, tables and figures, as they are more specialized data formats for which methods in information extraction are not yet as well-developed. We discuss the consequences of not compiling these latter data for taxonomic richness in later sections.

(b) Overview of information retrieval

The end-product that we seek is a dataset of observations of cheilostome fossil taxa and their geologic ages and corresponding bibliographic references. To do so, we used Poppler (https://poppler.freedesktop.org/) to extract plain text from the PDFs and then subject this text to information retrieval, as summarized in figure 1. There are three main steps in our information retrieval procedure. First, we used automated approaches to isolate sentences in which taxa and age names co-occur and produced data consisting of 'candidate' taxon-age pairs and the sentences in which they occur. Second, we manually labelled a subset of these candidates for constructing a set of 'gold labels' which are used in training a machine-learning classifier. Third, we applied the machine-learning classifier to the entire set of the candidates and generated our fossil occurrence data. We briefly detail each of these steps in the next sections.

(c) Sentence parsing and named entities

We used Stanford CoreNLP [37] to automate the annotation of linguistic information, such that we can recognize 'tokens' (i.e. words and punctuation), starts and ends of sentences, and named entities, specifically bryozoan fossil taxa and geologic time intervals. Taxon names are Latinized and can be identified in part by word morphology, but also via in-sentence context and grammar. In other words, we could have set up named-entity recognition using a machine-learning classifier [38]. However, it is next to impossible to distinguish among names from different taxonomic groups based on grammatical context and word morphology alone. Hence, we chose to use a rule-based approach, not least because a nearly exhaustive list of post-Palaeozoic bryozoan Linnaean binomials (including all cheilostome bryozoans, our target group) is already available. We used this list and our compiled list of geologic age interval



species	age	P(species in age | sentence)
Setosellina cf. *roulei*	Serravallian	0.81
Cribrilina punctata	Cretaceous	0.12
...	...	...

Figure 1. (*a*) The general workflow for automatic information extraction of fossil occurrence data. (*b*) The machine-learning classifier. A bidirectional long short-term memory (LSTM) recurrent neural network, with the first example candidate as input. The numbers given are illustratory. Dashed arrows indicate dependency grammar links. See electronic supplementary material for details on the classifier. The figure style is inspired by Miwa & Bansal [23]. (Online version in colour.)

names to generate a set of TokensRegex expressions [39] that were matched with occurrences within sentences (i.e. named-entity recognition, see electronic supplementary information for details). For one or more consecutive tokens that indicate one entity (e.g. '*Bugula*', '*Setosellina roulei*', 'lower Miocene'), we use the term 'span'. A candidate refers to a pair of spans, one of which is a full Linnaean binomial or genus name, and the other a geologic age. We used the Snorkel framework [40] to locate candidates, treating each candidate as independent information. The automated workflow thus far resulted in a product consisting of candidate pairs and the sentences in which they co-occur.

(d) Gold labels: manually labelled candidates

The set of candidates from the previous section consists of co-occurrences of a name and age, but these can be either positive or negative candidates. An example of a positive candidate (from [41], p. 54), where it is explicitly stated or strongly implied that the species occurred in the given geologic time interval, is:

> Remarks—A few, small, infertile colonies of ***Setosellina* cf. *roulei*** have been found encrusting the undersides of very thin platy corals from the [late Burdigalian] and the **Serravallian**.

Bold font indicates the relevant spans, and square brackets indicate spans that are not currently under consideration. An example of a negative candidate (from [42], p. 419) is:

> ... fix the identity of ***Cribrilina punctata***, the type species of the genus [*Cribrilina*], itself the type genus of the cosmopolitan **Cretaceous** to [Recent] family Cribrilinidae.

The candidates could be manually labelled as positive or negative for downstream use as part of a fossil occurrence dataset. By automating this task with a machine-learning classifier, we can more easily annotate large volumes of data, tackle new sources of data (e.g. newly published articles or old ones made available) or apply the classifier to other groups of taxa. The machine-learning classifier learns from examples (a supervised learning approach, see the next section). Two human annotators manually labelled 10 416 candidates, 1000 of which were labelled by both persons. These annotators were not taxonomists or bryozoologists, but had basic degrees in biology, understood the Linnaean classification system and had geological knowledge sufficient to recognize stratigraphic age units. For candidates where the annotators disagreed, we assigned a positive or negative label at random (see electronic supplementary material). These resulting human-annotated data are our labelled candidates, each with a corresponding gold label.

This labelling of candidates in the context of retrieved sentences takes much less time than an approach that starts with reading through numerous papers or monographs to find the relevant data wherever it might occur in the text and subsequently entering each separate taxon name and age.

(e) Machine-learning classifier

In order for a machine-learning classifier to retrieve the semantic relationship between two spans in a sentence, we characterized features of the sentence's elements. We described the links among words [19] using pre-trained machine-learning models for English dependency grammar in CoreNLP [43]. We used the dependency grammar tree to compute the shortest dependency path (SDP) between the two spans [22,44], as illustrated in electronic supplementary material, figure S4. The resulting SDPs from the labelled candidates and their gold labels were then used to train a supervised machine-learning classifier (figure 1*b*), specifically a Long Short-Term Memory recurrent neural network or LSTM [45] implemented similarly to Xu *et al.* [22].

To summarize, our neural network consists of three layers: a word embedding layer, a bidirectional LSTM layer and a third hidden layer (figure 1*b*). The output of the third and final layer is a vector of length two, which represents the probability mass for our relation classification task. We used Keras [46] to implement this neural network (see electronic supplementary information for details).

We split the labelled candidates into training, validation and test sets (80, 10, 10% of the labelled candidates, respectively). The training data were used to fit the parameters of the classifier. The validation data were used for two purposes, to decide when to stop training, and to hand-tune the hyperparameters (e.g. the layer dimensions and learning rates, see electronic supplementary information for details). We used the test set to evaluate classifier performance. We employed metrics that are standard in binary classification problems [47] to evaluate model performance: accuracy (the ratio of correct predictions to all predictions), precision (the ratio of true positive predictions to all positive predictions), recall (the ratio of true positive predictions to all positive labels), false positive rate (FPR, the ratio of false positive predictions to all negative labels) and F1 (the harmonic mean of precision and recall). In classification problems such as ours, a decision boundary divides the decision space into categories. Here, the outputs of our classifier are in the interval [0,1], and we set our binary decision boundary (b) to 0.5 unless otherwise stated. If we assume that the test set is representative of the unlabelled candidates, the performance of the classifier on the test set will approximate how well it classifies the unlabelled data. As an indicator of the estimation error of the above-mentioned metrics, we fitted the neural network 100 times with the same hyperparameters and report their mean ± s.d. evaluated on the test set. To avoid confounding effects of similar sentences in a given document, we ensured that all candidates found in a single document were contained within only one of the sets used for training, validation, and testing. We used the relation classifier to automatically label remaining candidates. As there were some duplicates among these remaining candidates, we estimate that there were about 10 000 unique unlabelled ones.

(f) Cheilostome palaeodiversity

In order to compare text-mined taxonomic richness through geologic time with previous publications, we considered only genera, even though we also text-mined full binomials. After the relation extractions of genera or binomials and geologic ages, we performed several steps of post-processing. We used Macrostrat [35], supplemented in a few cases by online lookups, to obtain the minimum and maximum ages of each named time interval. As genus names are commonly abbreviated (e.g. '*C. punctata*') after their first mention in a text section, we performed a de-abbreviation step by doing a lookup of genus names in the previous 15 sentences, that begin with the capitalized letter in the abbreviation. The abbreviated genus names were successfully de-abbreviated in about 80% of the candidates without a full genus name (electronic supplementary material, figure S2). Biodiversity compilations must account for invalid names, synonymizations and related historical artefacts of the literature, reflecting taxonomic revisions. Our cheilostome genera were standardized and revised to valid ones as far as possible using WoRMS [12] and data from two of its primary contributors and editors—Bock [34] and Gordon (Dennis P. Gordon, personal communication, 2018), who is maintaining the most recent working document of the Treatise on Invertebrate Paleontology, Part G: Bryozoa [48]. Species-level synonymies are not treated here.

3. Results and discussion

(a) Gold label annotation repeatability

We estimated an accuracy of 84.1% between the decisions made by the two human annotators for the mutually examined gold labels. This inter-annotator accuracy is a joint measure of noisy data, language ambiguity, human error and degree of conservatism by the annotators. Examples of noisy data include tables parsed as paragraphs, incorrect sentence boundaries, errors in optical character recognition and incomplete sentences such as those found in figure captions. We interpret the annotator repeatability as a workable baseline when judging the performance of the machine-classifier. Similar inter-person accuracy estimates have been reported in analogous human error assessments made by Peters *et al*. [18]. We also note that human error rates in data compilations are rarely explored quantitatively and never modelled in palaeobiological analyses on large datasets compiled from the literature (e.g. [49,50], but see [51] for an exceptional investigation of such errors).

(b) Classifier performance

The relation classifier applied to our test set achieved a recall of 88.0 ± 1.5%, accuracy of 82.2 ± 0.8%, precision of 82.8 ± 1.3% and an F1 of 85.3 ± 0.6%. Figure 2 illustrates the trade-off between recall and the occurrence of false positives. The classifier accuracy at 82.2% is comparable to our inter-annotator labelling accuracy at 84.1%. Since human and machine accuracies are comparable, yet both far from perfect, we suggest that machine-reading performance is more limited by the ambiguousness of the data and annotator scoring ability than by the classification algorithms. In other words, repeatable candidate labelling is probably a major bottleneck for accurate machine-based relation classification. Whether the types of machine-classification errors overlap substantially with human-annotator errors remains an open question. While we do not know if our inter-annotator accuracy is representative of the human accuracy involved in manually populating similar knowledge databases (e.g. the Paleobiology Database), we have no reason to believe that repeatability in manual data-entry is substantially better or worse.

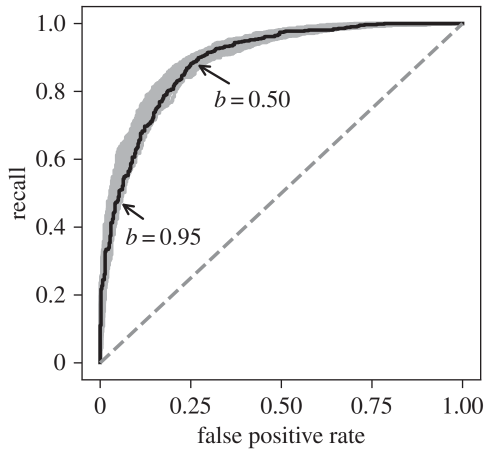


Figure 2. Receiver operating characteristic curve. The rates for the relation classifier are evaluated on the test set. Ninety-nine iterations are plotted in grey, and the one in black is chosen at random. $b = 0.50$ is the standard decision boundary, and $b = 0.95$ represents false positive rate of 5%. The area under the black curve is 0.90. The dashed line represents the expected rates given a random classifier.

Incomplete fossil preservation, biased sampling and selective reporting of data in publications can all contribute to gaps in compiled databases. However, there are diverse approaches that model sampling probabilities or account for sampling incompleteness in the estimation of species richness [52,53] or diversification dynamics [4,5,7,31]. False negatives from machine-read texts are analogous to the above-mentioned gaps in compiled databases in terms of our ability to alleviate their effect in inferring past richness patterns and dynamics. False positives, however, falsely inflate estimates of taxonomic richness and we do not currently have modelling approaches to deal with this bias. To minimize the impact of false positives in our text-mined data, we constrained the predictions to have a relatively low FPR while maintaining a relatively high recall, as shown in the next section.

(c) Genus richness counts

We used our text-mined predictions as well as the list of extant genera from the World Register of Marine Species [12] to produce a joint dataset from which range-through cheilostome genus richness is determined for successive time intervals over the past 160 Myr (figure 3). We set the decision boundary (b) to obtain a balance between relatively

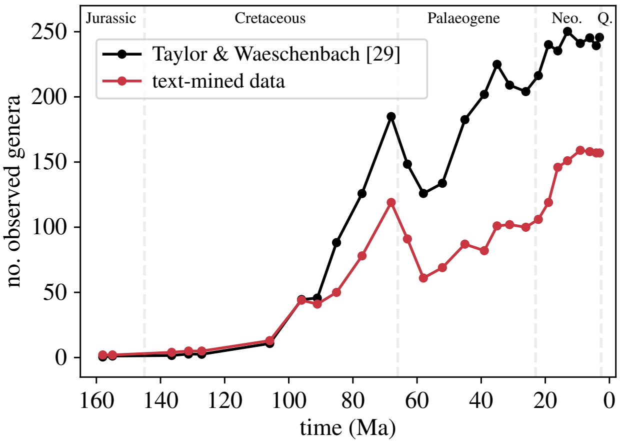


Figure 3. Range-through genus richness for cheilostomes. The curve from Taylor & Waeschenbach ([29], fig. 12) was obtained using a plot digitizer [54]. Our richness counts are supplemented with extant observations from WoRMS [12]. We used bins that are comparable with the bins used by Taylor & Waeschenbach [29]. The false positive rate evaluated on the test set is 27%. The geologic ranges for all genera are detailed in the electronic supplementary material, table S2. (Online version in colour.)

low FPR (25.9%) and high recall (88.1%). The FPRs are estimated for the test dataset. The net FPR is smaller than what is reported in figure 3. This is because some portion of candidates overlap with the training data, for which the classifier is comparably overfitted (FPR = 8.2 ± 4.2%, recall = 97.8 ± 1.0% for training data). However, the range-through genus counts are similar for subsets of data with FPR between 0.25 and 1 (electronic supplementary material, figure S6), indicating that the false positives are mostly adding to 'internal occurrences' rather than contributing to geologic range extensions (see electronic supplementary information, error inspections). In a previous, more comprehensive error-correction experiment using trilobites [51], the authors found that even after extensive corrections to geologic ranges and taxonomic nomenclature, the impact on genus richness counts through time was negligible.

The relative changes in our text-mined cheilostome genus richness curve (figure 3) are similar to those in a recently published review of bryozoan history [29], fig. 12), although our range-through genus counts are currently underestimates of that work. Genus richness increased in the Cenomanian (100.5–93.9 Ma) and steepened through the late Cretaceous (100–66 Ma), with a second diversification beginning in the Eocene (56–33.9 Ma). These trends are evident in both curves, as is the sharp decline in genus richness at the K-Pg extinction event and the end of the Danian around 62 Ma [55]. An advantage of our newly acquired occurrence-based dataset is that we now have multiple observations for most genera, including explicit literature sources. Such an occurrence-based dataset is amenable to taxonomic updates by systematists and consequently, revisions to inferred histories regardless of the approach applied. More beneficially, it facilitates the use of modern approaches to estimate richness patterns and diversification rates while accounting for incomplete sampling and sampling heterogeneity. These extensive analyses are outside the scope of the current work. However, the pipeline presented here will aid in paving the way for a more robust and nuanced inference of bryozoan palaeodiversity.

(d) Common problems

While there are substantial advantages to text-mining genus/species age-observation data, we recognize several avenues for improvement in this study. Other problems and avenues that we do not discuss may become relevant as developments in machine-reading approaches unfold, as this field is in its infancy concerning applications to biodiversity studies. First, there are substantial amounts of historical and recent literature in languages other than English, notably French, German, Spanish, Italian, Japanese and Russian. While the text-mining task is conceptually similar regardless of the language, natural language processing tools are more advanced and accurate for high-resource languages such as English and Chinese [13]. Specialized tools (e.g. named-entity recognition, dependency grammar and relation classifiers) are typically tailor-made for a specific language, meaning that adding other languages effectively multiplies the effort required for automatic information extraction.

Second, our text-mining pipeline is entirely reliant on grammatically structured, isolated, single (complete or incomplete) sentences that describe the taxa and their ages. We do not take into account contextual information, e.g. linguistic coreference [56] among sentences. Similarly, taxonomic treatments are an example of extreme context dependence, where genera or species typically are given in headlines followed by concise descriptions in the following paragraphs. Thus, many taxonomic descriptions are not based on grammatical information but rather the spatial layout of paragraphs and typefaces of words in the article—both of which are ignored by our approach. More specialized tools are required to process taxonomic treatments more effectively.

Third, we neglect information in tables and figures apart from their captions. Govindaraju *et al.* [57] demonstrated how combining natural language and tables could improve information extraction. However, this is conditional on a machine-readable data structure for tables, and table extraction from PDFs remains difficult to achieve with accuracy at this time [58].

A fourth problem is one of optical character recognition errors (e.g. incorrectly parsing a letter, or parsing words as letters separated by whitespace). Text normalization has been employed to handle similar problems in social media corpora [59,60]; however, normalization may also remove or distort information, and it remains to be tested on taxonomic literature. Correct parsing of words and paragraphs is inherently limited by design choices in the PDF standard. Publishers are becoming increasingly aware of this problem; for instance, Elsevier provides researchers with machine-readable plain text or XML (Extensible Markup Language) formats of their articles through their text and data mining initiative.

Any of these above-mentioned problems may contribute to our underestimations in comparison with Taylor & Waeschenbach [29], even though any data compilation of fossil occurrences will always be incomplete due to both the nature of the fossil record and the process of fossil recovery.

(e) Outlier inspection

In addition to false negatives that may contribute to underestimations, we also have false positives, some of which are temporal outliers. While most genera have reasonable estimated time spans (73% under 25 Myr), there are some genera that exhibit suspiciously long time spans (electronic supplementary material, figure S3). False positives that contribute to internal occurrences of taxon stratigraphic ranges are simultaneously less problematic but more difficult to remove from machine-read data. We inspect and discuss some of the stratigraphic outliers in the electronic supplementary material, figure S3. Some of the errors we found could have been avoided by setting a slightly more conservative FPR when sub-setting the data before analysis. Another type of false positive arises from taxonomic ambiguities. For instance, the genus *Callopora* is an accepted name for a Cenozoic cheilostome named by Gray [61]. However, we have mined data from Ernst & Nakrem [62] who referred to a historical mention of *Callopora* Hall [63] in their discussion of a Palaeozoic trepostome. Such an issue could in principle have been avoided by using author names for genera or species. However, author names are not always consistently supplied in-sentence and are often given in a different font or typeface, which is problematic for text parsing. In other words, knowledge of taxonomic practice in general and the taxonomy of the group in question is still crucial for an informed usage of machine-compiled data. We have in fact not dealt with species-level synonyms in this contribution because of the immensity of the task. However, past studies of fossil biodiversity have shown that taxonomic revisions do not necessarily change the nature of global scale patterns when datasets are sufficiently large [51,64].

(f) Concluding remarks

During the course of this work, it was clear that publications containing useful information on relatively understudied organisms such as Bryozoa remain difficult to locate without expert knowledge. In order to understand both past and present biodiversity, we need continued training and support of taxonomic experts [65,66]. Automating some information retrieval tasks will save time and allow taxonomists to focus on other tasks where their expertise is essential, such as describing new species and revising old names.

We believe that the pipeline presented here can be easily adapted for other groups of organisms, likely without taxon-specific gold labels, to build and augment large knowledge-bases of fossil occurrences. While human-vetted databases may be superior taxonomically, they are expensive to curate and maintain, and their data sources are sometimes difficult to trace, unlike our text-mined data. Faced with a growing literature and diminishing person-power and funding, machine-reading approaches can complement vetted databases by targeting knowledge gaps and hence contribute to understanding large-scale biodiversity changes and evolutionary dynamics. The machine-read fossil occurrence data presented here captured major palaeodiversity patterns in cheilostome bryozoans, despite varied sources of error that are generally inherent in large databases.

Data accessibility. The labelled candidates, word embeddings, cheilostome genus ranges, genera synonyms, code for named-entity recognition, candidate extraction and the classifier are available on the Open Science Framework at https://osf.io/dn4gy.

Authors' contributions. B.T.K. and L.H.L. designed the study, B.T.K. did the analyses and all authors wrote the manuscript.

Competing Interests. We declare we have no competing interests.

Funding. This project was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 724324).

Acknowledgements. We thank Olja Toljagić for help with labelling the candidates, Björn Berning, Dennis P. Gordon, Paul D. Taylor and Phil Bock for assistance with taxon lists, Caroline Schuette, Björn Berning and Kamil Zágoršek for assistance with the bibliography, and Farhad Nooralahzadeh, Shanan E. Peters, Ian Ross and Trond Reitan for discussions. We also thank the two anonymous reviewers whose criticism helped improve the narrative of our manuscript.

References

1. Sepkoski DJ. 2013 Towards 'A natural history of data': evolving practices and epistemologies of data in paleontology, 1800–2000. *J. Hist. Biol.* **46**, 401–444. (doi:10.1007/s10739-012-9336-6)
2. Valentine JW. 1985 *Phanerozoic diversity patterns: profiles in macroevolution*. Princeton, NJ: Princeton University Press.
3. Sepkoski JJ. 2002 A compendium of fossil marine animal genera. *Bull. Am. Paleontol.* **363**, 1–560.
4. Nichols JD, Pollock KH. 1983 Estimating taxonomic diversity, extinction rates, and speciation rates from fossil data using capture-recapture models. *Paleobiology* **9**, 150–163. (doi:10.1017/S0094837300007533)
5. Foote M. 2000 Origination and extinction components of taxonomic diversity: general problems. *Paleobiology* **26**, 74–102. (doi:10.1666/0094-8373(2000)26[74:oaecot]2.0.co;2)
6. Wang SC, Marshall CR. 2004 Improved confidence intervals for estimating the position of a mass extinction boundary. *Paleobiology* **30**, 5–18. (doi:10.1666/0094-8373(2004)030<0005:ICIFET>2.0.CO;2)
7. Alroy J *et al.* 2008 Phanerozoic trends in the global diversity of marine invertebrates. *Science* **321**, 97–100. (doi:10.1126/science.1156963)
8. Silvestro D, Schnitzler J, Liow LH, Antonelli A, Salamin N. 2014 Bayesian estimation of speciation and extinction from incomplete fossil occurrence data. *Syst. Biol.* **63**, 349–367. (doi:10.1093/sysbio/syu006)
9. Meyer C, Kreft H, Guralnick R, Jetz W. 2015 Global priorities for an effective information basis of biodiversity distributions. *Nat. Commun.* **6**, 8221. (doi:10.1038/ncomms9221)
10. Franz NM, Sterner BW. 2018 To increase trust, change the social design behind aggregated biodiversity data. *Database* **2018**, bax100. (doi:10.1093/database/bax100)
11. GBIF. 2018 The global biodiversity information facility. See https://www.gbif.org/.
12. WoRMS. 2018 World register of marine species: Bryozoa. See http://www.marinespecies.org/aphia.php?p=taxdetails&id=146142.
13. Hirschberg J, Manning CD. 2015 Advances in natural language processing. *Science* **349**, 261–266. (doi:10.1126/science.aaa8685)
14. Percha B, Garten Y, Altman RB. 2012 Discovery and explanation of drug-drug interactions via text mining. In *Proc. Pacific Symposium on Biocomputing* (eds RB Altman, KA Dunker, L Hunter, TA Murray), pp. 410–421. Toh Tuck Link, Singapore: World Scientific. (doi:10.1142/8254)
15. Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, Roberts A, Dobson RJ, Stewart R. 2017 Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* **7**, e012012. (doi:10.1136/bmjopen-2016-012012)
16. Batista-Navarro R, Zerva C, Nguyen NTH, Ananiadou S. 2017 A text mining-based framework for constructing an RDF-compliant biodiversity knowledge repository. In *Information management and Big data* (eds J Lossio-Ventura, H Alatrista-Salas), pp. 30–42. Berlin, Germany: Springer.
17. Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris RA, Penev L. 2018 OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system.

J. Biomed. Semant. **9**, 5. (doi:10.1186/s13326-017-0174-5)

18. Peters SE, Zhang C, Livny M, Ré C. 2014 A machine-compiled macroevolutionary history of Phanerozoic life. *PLoS ONE* **9**, e113523. (doi:10.1371/journal.pone.0113523)
19. de Marneffe M-C, Dozat T, Silveira N, Haverinen K, Ginter F, Nivre J, Manning CD. 2014 Universal Stanford dependencies: a cross-linguistic typology. In *Proc. 9th Int. Conf. Language Resources and Evaluation* (eds N Calzolari, K Choukri, T Declerck, H Loftsson, B Maegaard, J Mariani, A Moreno, J Odijk, S Piperidis), pp. 4585–4592. Paris: European Language Resources Association.
20. Bojanowski P, Grave E, Joulin A, Mikolov T. 2017 Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **5**, 135–146. (doi:10.1162/tacl_a_00051)
21. Zeng D, Liu K, Lai S, Zhou G, Zhao J. 2014 Relation classification via convolutional deep neural network. In *Proc. COLING 2014, the 25th Int. Conf. Computational Linguistics: technical papers* (eds L Tounsi, R Rak), pp. 2335–2344. Stroudsburg, PA: Association for Computational Linguistics.
22. Xu Y, Mou L, Li G, Chen Y, Peng H, Jin Z. 2015 Classifying relations via long short term memory networks along shortest dependency paths. In *Proc. 2015 Conf. Empirical Methods in Natural Language Processing* (eds L Màrquez, C Callison-Burch, J Su), pp. 1785–1794. Stroudsburg, PA: Association for Computational Linguistics. (doi:10.18653/v1/D15-1)
23. Miwa M, Bansal M. 2016 End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics* (eds K Erk, NA Smith), pp. 1105–1116. Stroudsburg, PA: Association for Computational Linguistics. (doi:10.18653/v1/P16-1105)
24. Wood ACL, Probert PK, Rowden AA, Smith AM. 2012 Complex habitat generated by marine bryozoans: a review of its distribution, structure, diversity, threats and conservation. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **22**, 547–563. (doi:10.1002/aqc.2236)
25. James NP, Clarke JAD (eds) 1997 Cool-water carbonates. SEPM Society for Sedimentary Geology. (doi:10.2110/pec.97.56.0001)
26. Jackson JBC, Cheetham AH. 1990 Evolutionary significance of morphospecies: a test with cheilostome Bryozoa. *Science* **248**, 579–583. (doi:10.1126/science.248.4955.579)
27. McKinney FK, Lidgard S, Sepkoski JJ, Taylor PDT. 1998 Decoupled temporal patterns of evolution and ecology in two post-Paleozoic clades. *Science* **281**, 807–809. (doi:10.1126/science.281.5378.807)
28. Sepkoski JJ, McKinney FK, Lidgard S. 2000 Competitive displacement among post-Paleozoic cyclostome and cheilostome bryozoans. *Paleobiology* **26**, 7–18. (doi:10.1666/0094-8373(2000)026<0007:CDAPPC>2.0.CO;2)
29. Taylor PD, Waeschenbach A. 2015 Phylogeny and diversification of bryozoans. *Palaeontology* **58**, 85–599. (doi:10.1111/pala.12170)
30. Alroy J *et al.* 2001 Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proc. Natl Acad. Sci. USA* **98**, 6261–6266. (doi:10.1073/pnas.111144698)
31. Foote M. 2003 Origination and extinction through the Phanerozoic: a new approach. *J. Geol.* **111**, 125–148. (doi:10.1086/345841)
32. Liow LH, Nichols JD. 2010 Estimating rates and probabilities of origination and extinction using taxonomic occurrence data: capture–mark–recapture (CMR) approaches. In *Short courses in paleontology: quantitative paleobiology* (eds G Hunt, J Alroy), pp. 81–94. Bethesda, MD: Paleontological Society.
33. Busk G. 1852 *Catalogue of marine polyzoa in the collection of the British museum. Part I. Cheilostomata. British museum (natural history)*. London, UK: Department of Zoology.
34. Bock P. 2018 Recent and fossil bryozoa. http://www.bryozoa.net.
35. Peters SE, Husson JM, Czaplewski J. 2018 Macrostrat: a platform for geological data integration and deep-time earth crust research. *Geochemistry, Geophysics, Geosystems* **19**, 1393–1409. (doi:10.1029/2018GC007467)
36. Rohde RA. 2005 GeoWhen. http://www.stratigraphy.org/bak/geowhen/index.html
37. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. 2014 The Stanford CoreNLP natural language processing toolkit. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics: system demonstrations* (eds K Bontcheva, J Zhu), pp. 55–60. Stroudsburg, PA: Association for Computational Linguistics. (doi:10.3115/v1/P14-5010)
38. Strubell E, Verga P, Belanger D, McCallum A. 2017 Fast and accurate entity recognition with iterated dilated convolutions. In *Proc. Conf. Empirical Methods in Natural Language Processing* (eds M Palmer, R Hwa, S Riedel), pp. 2670–2680. Stroudsburg, PA: Association for Computational Linguistics. (doi:10.18653/v1/D17-1283)
39. Chang AX, Manning CD. 2014 TokensRegex: defining cascaded regular expressions over tokens. Stanford University Computer Science Technical Reports 2014–02.
40. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. 2017 Snorkel: rapid training data creation with weak supervision. *Proc. Very Large Databases Endow.* **11**, 269–282. (doi:10.14778/3157794.3157797)
41. Di Martino E, Taylor PD. 2014 Miocene Bryozoa from East Kalimantan, Indonesia. Part I: cyclostomata and 'Anascan' Cheilostomata. *Scr. Geol.* **146**, 17–126.
42. Winston JE. 1999 *Describing species: practical taxonomic procedure for biologists*. New York, NY: Columbia University Press.
43. Chen D, Manning CD. 2014 A fast and accurate dependency parser using neural networks. In *Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)* (eds D Chen, C Manning), pp. 740–750. Stroudsburg, PA: Association for Computational Linguistics. (doi:10.3115/v1/D14-1082)
44. Bunescu RC, Mooney RJ. 2005 A shortest path dependency kernel for relation extraction. In *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing* (eds R Mooney, C Brew, C Lee-Feng, K Kirchoff), pp. 724–731. Stroudsburg, PA: Association for Computational Linguistics.
45. Hochreiter S, Schmidhuber J. 1997 Long short-term memory. *Neural Comput.* **9**, 1735–1780. (doi:10.1162/neco.1997.9.8.1735)
46. Chollet F *et al.* 2015 Keras. See https://keras.io.
47. Powers DMW. 2011 Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *J. Mach. Learn. Technol.* **2**, 37–63. (doi:10.9735/2229-3981)
48. Bassler RS. 1953 Part G, Bryozoa. In *Treatise on invertebrate paleontology* (ed. RC Moore), pp. 1–253. Lawrence, KS: University of Kansas Press.
49. Liow LH, Finarelli JA. 2014 A dynamic global equilibrium in carnivoran diversification over 20 million years. *Proc. R. Soc. B* **281**, 20132312. (doi:10.1098/rspb.2013.2312)
50. Liow LH, Reitan T, Harnik PG. 2015 Ecological interactions on macroevolutionary time scales: clams and brachiopods are more than ships that pass in the night. *Ecol. Lett.* **18**, 1030–1039. (doi:10.1111/ele.12485)
51. Adrain JM, Westrop SR. 2000 An empirical assessment of taxic paleobiology. *Science* **289**, 110–112. (doi:10.1126/science.289.5476.110)
52. Alroy J. 2010 The shifting balance of diversity among major marine animal groups. *Science* **329**, 1191–1194. (doi:10.1126/science.1189910)
53. Starrfelt J, Liow LH. 2016 How many dinosaur species were there? Fossil bias and true richness estimated using a Poisson sampling model. *Phil. Trans. R. Soc. B* **371**, 20150219. (doi:10.1098/rstb.2015.0219)
54. Rohatgi A. 2011 WebPlotDigitizer. https://automeris.io/WebPlotDigitizer
55. McKinney FK, Taylor PD. 2001 Bryozoan generic extinctions and originations during the last one hundred million years. *Palaeontol. Electron.* **4**, 1–26.
56. Clark K, Manning CD. 2016 Improving coreference resolution by learning entity-level distributed representations. In *Proc. 54th Annual Meeting of the Association for Computational Linguistics* (eds K Erk, NA Smith), pp. 643–653. Stroudsburg, PA: Association for Computational Linguistics. (doi:10.18653/v1/P16-1105)
57. Govindaraju V, Zhang C, Ré C. 2013 Understanding tables in context using standard NLP toolkits. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics* (eds H Schuetze, P Fung, M Poesio), pp. 658–664. Stroudsburg, PA: Association for Computational Linguistics.
58. Khusro S, Latif A, Ullah I. 2015 On methods and tools of table detection, extraction and annotation in PDF documents. *J. Inf. Sci.* **41**, 41–57. (doi:10.1177/0165551514551903)

59. Liu F, Weng F, Jiang X. 2012 A broad-coverage normalization system for social media language. In *Proc. 50th Annual Meeting of the Association for Computational Linguistics* (eds H li, L Chin-Yew, GG Lee, JC Park), pp. 1035–1044. Stroudsburg, PA: Association for Computational Linguistics.

60. Chrupała G. 2014 Normalizing tweets with edit scripts and recurrent neural embeddings. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics* (eds K Toutanova, H Wu), pp. 680–686. Stroudsburg, PA: Association for Computational Linguistics. (doi:10.3115/v1/P14-2111)

61. Gray JE (ed.) 1848 *List of the specimens of British animals in the collection of the British museum: part 1. Centroniae or radiated animals*. London, UK: Trustees of the British Museum.

62. Ernst A, Nakrem HA. 2011 Stenolaemate bryozoans from the Mjøsa Formation (Late Ordovician, Katian) of Helgøya (Mjøsa), southern Norway. *Nor. Geol. Tidsskr.* **91**, 163–180.

63. Hall J. 1852 *Palaeontology of New York vol. 2; containing descriptions of the organic remains of the Lower Middle Division of the New York System*. Albany, NY: C. Van Benthuysen.

64. Sepkoski JJ. 1993 10 years in the library - new data confirm paleontolgoical patterns. *Paleobiology* **19**, 43–51. (doi:10.1017/S0094837300012306)

65. Agnarsson I, Kuntner M. 2007 Taxonomy in a changing world: seeking solutions for a science in crisis. *Syst. Biol.* **56**, 531–539. (doi:10.1080/10635150701424546)

66. Paknia O, Rajaei H, Koch A. 2015 Lack of well-maintained natural history collections and taxonomists in megadiverse developing countries hampers global biodiversity exploration. *Organ. Divers. Evol.* **15**, 619–629. (doi:10.1007/s13127-015-0202-1)