Past and Future uses of Text Mining in Ecology & Evolution

Authors
Maxwell J. Farrell[1], Liam Brierley[2], Anna Willoughby[3,4], Andrew Yates[5], Nicole Mideo[1]

1. Department of Ecology & Evolutionary Biology, University of Toronto, Canada
2. Department of Health Data Science, University of Liverpool, UK
3. Odum School of Ecology, University of Georgia, USA
4. Center for the Ecology of Infectious Diseases, University of Georgia, USA
5. University of Amsterdam, Netherlands

April 5th 2022

Abstract

Ecology and evolutionary biology, like other scientific fields, are experiencing an exponential growth of academic manuscripts. As domain knowledge accumulates, scientists will need new computational approaches for identifying relevant literature to read and include in formal literature reviews and meta-analyses. Importantly, these approaches can also facilitate automated, large-scale data synthesis tasks and build structured databases from the information in the texts of primary journal articles, books, grey literature, and websites. The increasing availability of digital text, computational resources, and machine-learning based language models have led to a revolution in text analysis and Natural Language Processing (NLP) in recent years. NLP has been widely adopted across the biomedical sciences, but is rarely used in ecology and evolutionary biology. Applying computational tools from text mining and NLP will increase the efficiency of data synthesis, improve the reproducibility of literature reviews, formalise analyses of research biases and knowledge gaps, and promote data-driven discovery of patterns across ecology and evolutionary biology. Here we present recent use cases from ecology and evolution, and discuss future applications, limitations, and ethical issues.

Why use text mining?

The volume of scientific literature is growing exponentially [1], with over three million peer-reviewed academic articles published each year [2]. In a sample of 33 ecology journals alone, over 80,000 articles have been published since 1980 [3]. Reading this overwhelming amount of material is an insurmountable task, making manual literature syntheses and compilation of literature-based datasets increasingly difficult. As bodies of literature continue to grow, highly cited papers are more likely to be cited compared to recent work, which can result in slowing of scientific progress as transformative ideas are less likely to permeate and make substantive impact [4]. Adopting computational approaches for analysis of scientific texts allows researchers to rapidly and systematically identify relevant publications and synthesize larger amounts of literature compared to manual approaches. Beyond literature syntheses, computational tools can be used to efficiently extract information from texts and update existing literature-based databases, ultimately increasing the value of published research.

When humans read, we interpret information in text through the meaning of words and grammatical contexts. To a computer, human language is complex and difficult to convert to structured formats, such as tabular or relational databases commonly used in scientific research. Therefore, raw text is commonly referred to as “unstructured”. To convert unstructured data in scientific texts to a format ready for statistical analysis, we can apply a diverse set of computational approaches. These tools broadly fall under the umbrella of “text mining”, but often come from Natural Language Processing (NLP), a field that focuses on computational interpretation of human language, blending theory and approaches from linguistics, computer science, statistics, and artificial intelligence. NLP comprises an extremely broad set of computational methods that allow us to gather, sort, translate, and understand written documents.

Tools for mining scientific texts have seen wide-scale adoption in other fields, such as biomedical sciences, where models have been developed to recommend relevant literature, and extract data for further analysis. Exciting examples include the construction of large scale databases of protein-protein interactions [5], drug-drug interactions [6], gene-disease relations [7], chemical-disease relations [8], and interfaces to extract information using structured searches [9]. Applications of NLP in ecology & evolution are relatively rare compared to biomedical sciences (Fig. 1). The disparity in onset and magnitude of adoption suggests that ecology and evolution researchers could look to biomedical studies for inspiration on applying classical and cutting edge NLP approaches in their projects.

NLP itself is a rapidly growing field with many approaches applicable to ecology and evolution. In recent years, ecologists and evolutionary biologists have begun to develop similar domain-specific approaches, but their applications have largely been restricted to the analysis of publishing trends and related metrics. Given the growing and diverse types of literature, the importance of research syntheses, and increasing computational literacy in the field, ecology and evolutionary biology are prime candidates for the application of more advanced text mining and NLP approaches.

Using NLP to create literature-based databases holds particular value for comparative studies and biodiversity syntheses as these studies can be greatly accelerated by improving the reproducibility and efficiency of data integration [10]. Further, the aggregation of key biodiversity data enables analyses that would not otherwise be possible [11]. While peer-reviewed literature in journals represents the most common source of scientific texts, application of NLP to other texts, such as preprints [12], could highlight emergent and rapidly-changing science including the COVID-19 pandemic [13,14]. Considerable ecological knowledge is also stored in older books and texts associated with archival samples and natural history collections [15], but recent advances in document scanning, digitization, and Optical Character Recognition (e.g., from printed or handwritten texts) mean NLP approaches are now feasible and promising [16,17]. This technological advancement parallels the invention of new sensors and machine learning tools for image analysis in wildlife conservation [18,19]. Similarly, there exist vast amounts of text published alongside online genetic sequence databases such as GenBank or the Gene Expression Omnibus [20]. With increasing digitization efforts and availability of associated texts, adoption of text mining in ecology and evolution could greatly expand metadata and maximise the utility of these ever-growing resources.

Beyond supporting the efficient creation and expansion of literature-derived databases, using scripted and archived computational processes for text analysis can dramatically improve transparency, and help the reproducibility in all phases of research, from identifying relevant papers, analysing research trends, constructing and expanding datasets, and automated translation of text into data ready for statistical analysis. Here we outline current and future applications of text mining in ecology and evolutionary biology, and discuss current barriers to implementation.

Recent applications in ecology & evolution

Detecting trends and topics

The most common uses of text analysis in ecology and evolution have been under the umbrella of bibliometrics: quantitative research that studies trends in subject matter, authorship, and impact of publications. For example, Anderson et al. (2021) [21] analysed over 130,000 articles to explore the increasing diversity of ecological hypotheses and theories published over the past 80 years. Similar studies of publishing trends have explored ecological topics in high impact journals [22], showed the emergence of conservation biology as a separate discipline from ecology [23], analysed the growth of interdisciplinarity in biodiversity science [24], tracked shifting popularity of topics within industrial ecology [25] and fish ecology [26], identified research themes in disease ecology [27], and pinpointed critical research gaps in conservation science [28] and pollination ecology [29]. Outside of academic articles, text mining can reveal important trends for environmental management and biodiversity conservation [30]. In conservation science, analysis of online texts and social media posts led to the development of *conservation culturomics*, a field that evaluates public interest in nature [31], tracks opinions on conservation topics [32], and quantifies people's experiences in nature [33] based on an increasingly diverse set of data sources [34].

Beyond tracking trends, text analysis can be used to gather evidence supporting the success of conservation actions and develop more culturally-relevant policies.

Evidence synthesis and literature reviews

The growth of scientific literature is making evidence synthesis an increasingly difficult task, leading to an ever widening “synthesis gap” [35]. For both narrative and systematic reviews, text mining is projected to become a necessary tool to circumvent literature overload [36]. Text analysis can be implemented at multiple phases of a review, from identifying search terms using keyword co-occurrence networks [37], to applying predictive approaches to screen studies for inclusion [35]. Abstract screening using text mining and machine learning can be a precise and efficient alternative to the common practice of screening abstracts with two reviewers [38], which may help limit individual biases by providing a consensus annotation, but is time consuming and can be error-prone. The future of systematic reviewing will likely necessitate the interaction of humans and machine learning algorithms to tackle the rapid growth in publications [39]. Overall, implementing computational processes can dramatically expand literature assessments to include more diverse texts, increase the efficiency of reviews and literature syntheses, and allow rapid reproducibility and updating as new literature is published [36]. These tools need to be properly calibrated and validated to ensure accuracy compared to manual search and screening [35,40,41].

Expanding literature-based datasets

Large-scale studies in ecology are often based on data compiled from previously published research and typically involve significant manual investment for literature searching, acquisition, screening, data extraction, and harmonisation of entities such as species names, place names, measurement units, experimental designs, and terminology with inconsistent definitions [11]. As such, these studies require substantial effort to update as new papers are published. In NLP, the sub-field of information retrieval develops search algorithms and models that suggest articles of potential interest. In a recent ecological application, Cornford et al. (2021) [42] train machine learning models to classify literature as relevant to the PREDICTS database [43], a literature-based database of biodiversity responses to human impacts. Their best models could distinguish relevant from non-relevant articles with over 90% accuracy based only on title and abstract text, significantly improving the speed and ease with which new articles can be screened for database inclusion. A similar machine learning approach was used by Roll, Correia, and Berger-Tal (2018) to identify articles using the term ‘reintroduction’ in a conservation context (release of organisms into their historical native habitat), rather than a non-ecological context. Outside of search engines, a number of machine learning models for text classification have been developed in recent years [45], but are rarely used in ecology and evolutionary biology [46]. The ability to continually flag and integrate relevant publications will help transition from static ecological datasets to living ones, and help promote more efficient, timely, and impactful science.

Extraction and integration of primary biodiversity data

Integrating data from across the life sciences is currently a major challenge, but will likely foster the interdisciplinary research needed to address pressing global issues [47]. With NLP approaches, unstructured texts can be more efficiently transformed into structured data commonly analysed in ecological and evolutionary studies. With dictionaries containing terms of interest (e.g., species names, traits, keywords describing an ecological interaction), the frequency of term co-occurrences can be used to discover associations [48]. For example, by quantifying the co-occurrence frequencies of ant species names and terms describing ant-plant mutualisms, Kaur et al. [49] were able to identify ant species associated with mutualistic behaviours, and used the compiled dataset to study the evolution of plant mutualisms. Similar approaches have been used to infer inter-species associations via descriptions from the Encyclopedia of Life [50], and NCBI and PubMed [51,52]. Ecologists have also used text from Twitter [53] and news sources to gather species-linked data that can infer population trends, geographic ranges, or even monetary values, that support innovative systems to monitor and respond to conservation concerns [54].

These studies used dictionaries to identify relevant terms, but to go beyond lists of words, terms can be linked to other datasets using ontologies. In linguistics, an ontology refers to a set of terms and their relationships, forming a network of concepts in a domain [55]. Ontologies capture expert knowledge and allow users to translate concepts across databases, disambiguate terms with different disciplinary meanings, or collapse terms into larger concepts (much like a taxonomy allows collapsing species into genera, families, orders, etc…). Ontologies have proven useful in biomedicine [56,57] and for harmonising data across diverse texts to study important problems within environmental science, bacterial evolution, and comparative anatomy [58–63].

Ecology and evolution are rife with ambiguously defined terminology (e.g. the definition of "virulence" depends on if the pathogen infects a plant or animal host, and often differs between theory and empirical papers [64]), which slows research progress and limits the ability to synthesise across studies [65,66]. Creating platforms with consistent naming conventions and connected concepts will facilitate data harmonisation, sharing and annotation, and aid collaborative research projects already common in biodiversity science [67]. There exist a number of related ontologies describing ecological observations [68], biological collections [69], phenotypes [70,71], and biodiversity science [72]. Recent efforts have aimed to generate consensus definitions for ecological traits with ontologies [73]. These act as resources for describing, accessing, and manipulating phenotypic data by making phenotypic data more manipulable by computers [74], efficiently extracting phenotypic data from taxonomic descriptions [75], structuring species information [76], and harmonising traits with taxa [77]. Developing diverse vocabularies, definitions, and relationships among concepts is crucial for dealing with the heterogeneous nature of information in ecology and evolution, and these initiatives will lay the groundwork for more automated text analyses in the future.

Future uses of text mining and Natural Language Processing in ecology & evolution

Given the current limited use of NLP approaches in ecology and evolution, we suggest that their adoption will have the greatest impact on the construction of large scale comparative databases. We highlight three tasks that are likely to be extremely useful: document classification, tagging domain-specific entities in text, and building structured databases through relation extraction (Figure 2). Each of these tools can be generalised to future research projects, or linked together to build a workflow from raw texts to a structured database ready for analysis. In general, model performance will differ based on the specific task, goals of the larger project, and to what degree metrics such as precision or recall should be optimised. For example, a computational approach may not return all articles identified in a manual search, but may still be desirable if it identifies a larger number of relevant texts to include, or offers the ability to much more rapidly analyse a larger set of documents. Below we assume that some source texts (corpus) have already been identified, either through targeted literature searches, or choice of an existing body of literature. We do not discuss article search strategies, as detailed guides exist [78], but note that this is an important consideration when gathering a corpus and designing a text mining project.

Document classification

The success of document classification by Cornford et al. [42] demonstrates the potential for document-level predictive models to aid the updating of large-scale comparative databases. As a general template, if databases derived from published articles can be linked with abstracts or full texts, classifiers can be trained to predict whether subsequently published articles are likely to contain relevant data. Training a classifier requires examples of both positive and negative cases (e.g. relevant and irrelevant articles). Databases that report discarded literature are great resources of positive and negative examples. However, because existing databases rarely document these, “irrelevant” papers may be identified by sampling papers in the discipline, such as the use of general ecology papers by Cornford et al. [42]. These irrelevant articles are similar to the use of “background” or “pseudo-absence” data in species distribution models [79] in that they may contain undocumented positives (i.e. relevant articles), but the assumption is that the majority will be irrelevant and provide a useful contrast to those in existing databases.

The choice of negative examples for training should reflect future search strategies, whether it be searching through all ecology papers, or a more specific set. If the source database clearly outlines their strategy for literature inclusion (i.e. search terms, targeted journals, publication dates), it may be possible to compile more targeted sets of negative examples for training. Future development of document classifiers should explore the influence of these different approaches for generating negative training data on accuracy, and validate these predictive models on articles which have been expert-validated rather than assumed to be irrelevant (Fig. 2). In addition to periodic updating, using predictive models to expand existing datasets will lay the foundation for systems that can automatically alert researchers of relevant papers as they are published, and automatically extract data from these papers.

Identifying entities specific to ecology & evolution

Once relevant texts are identified, the next task is extraction of relevant terms. If researchers know exactly what they are looking for and terms of interest are completely known and can be listed, simple methods such as text matching can be used to identify them. However, given the diversity of specialised terms in ecology, this is unlikely to be the case. When relevant terms are not known, or texts are expected to include never-before-seen terms, Named Entity Recognition (NER) will be extremely useful. NER involves identifying real-world objects ("named entities") based on the context of their surrounding text, such as people, locations, organisations, etc. In biomedical text analysis, specialised NER tools are built to identify mentions of diseases, genes, proteins, cell types, and chemicals [80]. NER tools designed for ecology and evolutionary biology are currently rare, but would greatly improve literature exploration and information extraction. Contemporary NER tools are often created by adapting deep learning based language models [81]. Therefore, given suitable training data, NER models can be trained to recognize and disambiguate ecology-specific entities (see Fig. 2). For example, the recently developed TaxoNERD [77] is a deep-learning based model that recognizes scientific and common species names, and can normalise names to match NCBI or GBIF. One current challenge to developing deep learning-based ecological language models from scratch is the lack of domain specific "gold standard" training data. However, the authors of TaxoNERD overcome this by starting with a pre-trained biomedical language model and updating it for an ecological task. This successful example of transfer learning demonstrates the potential of large deep-learning based models to generalise to novel tasks and reduce the amount of labelled training data needed to build a useful tool explicitly for ecology & evolutionary biology. Moving forward, the development of NLP tools for ecology & evolution could be greatly supported by hubs of open access training data, such as those created for image analysis in biology and conservation [82].

Once named entities are recognized, a text analysis pipeline can take many different paths. To better understand context, researchers may cross-reference terms with ontologies to connect concepts or collapse terms into higher groups. For example, scispaCy v2.5.0 supports entity linking to biomedical ontologies including the Unified Medical Language System (UMLS) [83] and the Medical Subject Headings (MeSH terms) [84], which in turn allow them to be connected to a diverse array of databases. These may be used to group organs into larger anatomical systems, or categorise proteins into enzymes, hormones, or antibodies. While approaches have been developed to identify taxonomic, morphological, and habitat entities [62,85], merge existing ontologies [86,87], and create standards for publishing of biodiversity information [88], these initiatives remain disconnected, and have not yet been integrated with contemporary NLP software.

Relation extraction & creation of structured datasets

Once entities are recognized, and disambiguated or linked to an ontology, multiple approaches can be used to identify relationships among these entities (for examples, see Fig. 2 & Table 1). One approach is analysis of term co-occurrences, as used by Kaur et al. [49] to identify ant-plant mutualisms. Alternatively, the structure of the text itself can be used to identify the relationships, through a task referred to as relation extraction. Relation extraction can be done by incorporating linguistic information, such as semantic relationships between entities, or through training of a deep-learning based language model if one is available. For example, identifying protein-protein interactions in text has progressed from using a dictionary of protein names and co-occurrences, to adding information about parts of speech (e.g. verbs, nouns, adjectives), to supervised and deep learning approaches that incorporate vector representations of articles as predictors [89]. Relation extraction can also be used to identify relations between different classes of entities, such as disease-gene interactions [7]. Relation extraction is often a complex task, which can be daunting for researchers new to text mining. However, given the diversity and value of relational data in ecology and evolution, we suggest that relation extraction will be an increasingly important means of generating structured, analysis-ready data in the future. This offers exciting new frontiers for ecologists and evolutionary biologists to collaborate with computational linguists and computer scientists.

Current barriers to adoption and pathways forward

Despite the promise of text mining to revolutionise literature synthesis and database creation, several technical and social barriers currently limit widespread adoption in ecology and evolution. These include a lack of knowledge of existing tools, best practices, and shared vocabularies needed for collaboration with computational linguists [35]**.** Further, there are inequalities in access to software, data, and academic literature [46,90,91]. To use text mining and NLP in ecology and evolution to their full potential, we need to promote awareness of these methods, improve access to scientific literature and article-level metadata, facilitate cross-disciplinary collaborations, create domain-specific software, and develop an ecosystem of scientific language tools that work across all the world's languages, not just English. Recent successful applications of NLP approaches in ecology and conservation biology have involved close collaborations between biologists and computer scientists [54,92] highlighting the importance of cross-disciplinary research. However, as general tools and frameworks exist, their adoption in ecology and evolution is now limited by access to texts, development of applications specific to ecology, and the dissemination and uptake of these tools.

For primary literature, abstracts are among the most readily accessible documents and can be sufficient for document classification and database creation [42,49,93]. However, abstracts may not be available for more historical papers [21], and analyses of manuscript full texts are likely to outperform the use of abstracts only, as shown for relation extraction [93]. Unlike abstracts, access to full academic texts is limited by institutional subscriptions [90], with only half of publishers releasing manuscripts in a machine readable format [94]. Access to paywalled articles and copyright issues will limit the reproducibility of studies using text mining, and re-publishing or hosting source texts as supplementary materials may be illegal. Projects such as the PMC Open Access Subset offers bulk download of 100,000s of articles in machine-readable format [95], and The General Index [96], an open access database of text sequences and keywords extracted from 107 million journal articles, offers researchers the ability to perform specialised searches and analyse thematic trends in scientific literature without barriers imposed by paywalls or institutional access. While such databases can greatly improve interaction with published literature, their success relies on unrestricted bulk access to primary texts. Interfaces such as application programming interfaces (APIs) can facilitate scripted retrieval of texts, but usually involve arbitrary rate-limitation which makes large-scale analyses difficult and hampers literature-based research [97]. Thus, scientific advances in synthesising studies in ecology & evolution are limited by business decisions and publisher-imposed restrictions that create artificial scarcity [98]. In turn, when analysing large volumes of papers, researchers should take care to cite primary sources appropriately. However, the mainstreaming of text mining has resulted in a need for new bibliometric and citation infrastructure to facilitate transparent and permanent linking of large citation lists, and allow proper acknowledgement of individual studies that underlie large-scale literature surveys. Overall, the scale and reproducibility of text mining studies will be hindered until scientific articles are considered a public good and made open and freely accessible.

Parallel to variation in access to scientific publications, the dominance of English in science has led to data from non-English publications being omitted from ecological syntheses [99]. There also exist systematic inequalities in the representation and performance of NLP technologies across languages [91,100]: largely because of the historical dominance of English as the *lingua franca* of scientific publishing, current scientific language models are designed only for English texts [101,102]. As training data and models for previously overlooked languages continue to grow [103], the future looks promising for expansion of NLP approaches to non-English scientific texts. This could promote broader inclusion in science by facilitating translation of publications across languages, easing barriers for researchers to publish in their chosen languages, and allowing broader inclusion of non-English scientific texts in synthetic research.

Conclusion

The application of text mining and Natural Language models to domain-specific text in ecology and evolutionary biology shows great promise for summarising historical research and current gaps in knowledge, efficiently identifying pertinent literature, constructing structured databases from unstructured texts, and developing real-time biodiversity surveillance for issues such as emerging diseases and conservation threats. “We urge early career scientists and established researchers alike to explore and apply these tools in their own research, foster interdisciplinary collaborations, build open access corpora, contribute their expertise to developing open-source software and expert-created training data, and develop tools that are designed specifically for processing texts in ecology and evolution.”

Acknowledgements

We thank Ross Mounce, Nicolas Le Guillarme, and Ricardo Correia, the Mideo Lab, and two anonymous reviewers for constructive feedback that improved the manuscript. MJF was supported by the University of Toronto EEB Fellowship, an NSERC PDF, and an RBC Postdoctoral Fellowship. LB was supported by an MRC Skills Development Fellowship award (MR/T027355/1). NM was supported by an NSERC Discovery Grant. ARW was supported by the National Science Foundation under Grant No. DGE-1545433.

Figures & Tables

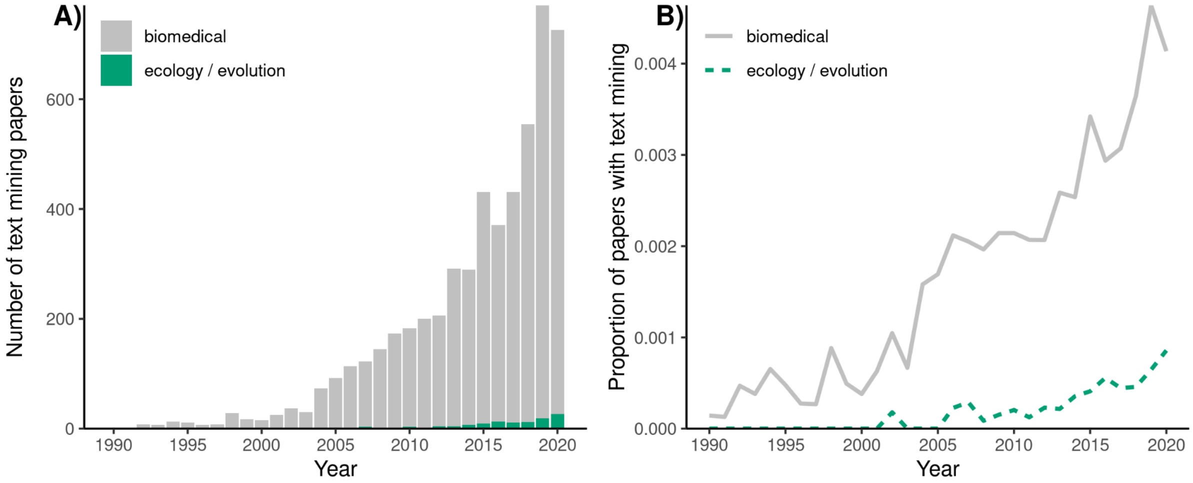

Figure 1. Publication trends indicating an earlier adoption, and greater A) absolute number and B) proportion of papers involving text mining in biomedical publications compared to ecology and evolutionary biology. Data were from two Web of Science (WOS) searches: one with "*medic*" and the other with "ecology" OR "evolutionary biology" OR "biodiversity" in the Topic field, plus "text mining" OR "Natural Language Processing" OR "NLP" in All Fields for each search. A total of 5,262 biomedical papers and 120 ecology / evolutionary biology papers employing text mining or NLP were identified out of a total 2,355,632 biomedical and 354,798 ecology / evolution papers. Searches were conducted on September 10th 2021 via the University of Toronto subscription. Note that variation in WOS search results varies due to institutional subscriptions [90]. Search results were subset to the years 1990-2020 inclusive. Data and R code to reproduce the figure, and .bib files with citation information for the returned articles can be accessed at https://github.com/maxfarrell/textmining_trends

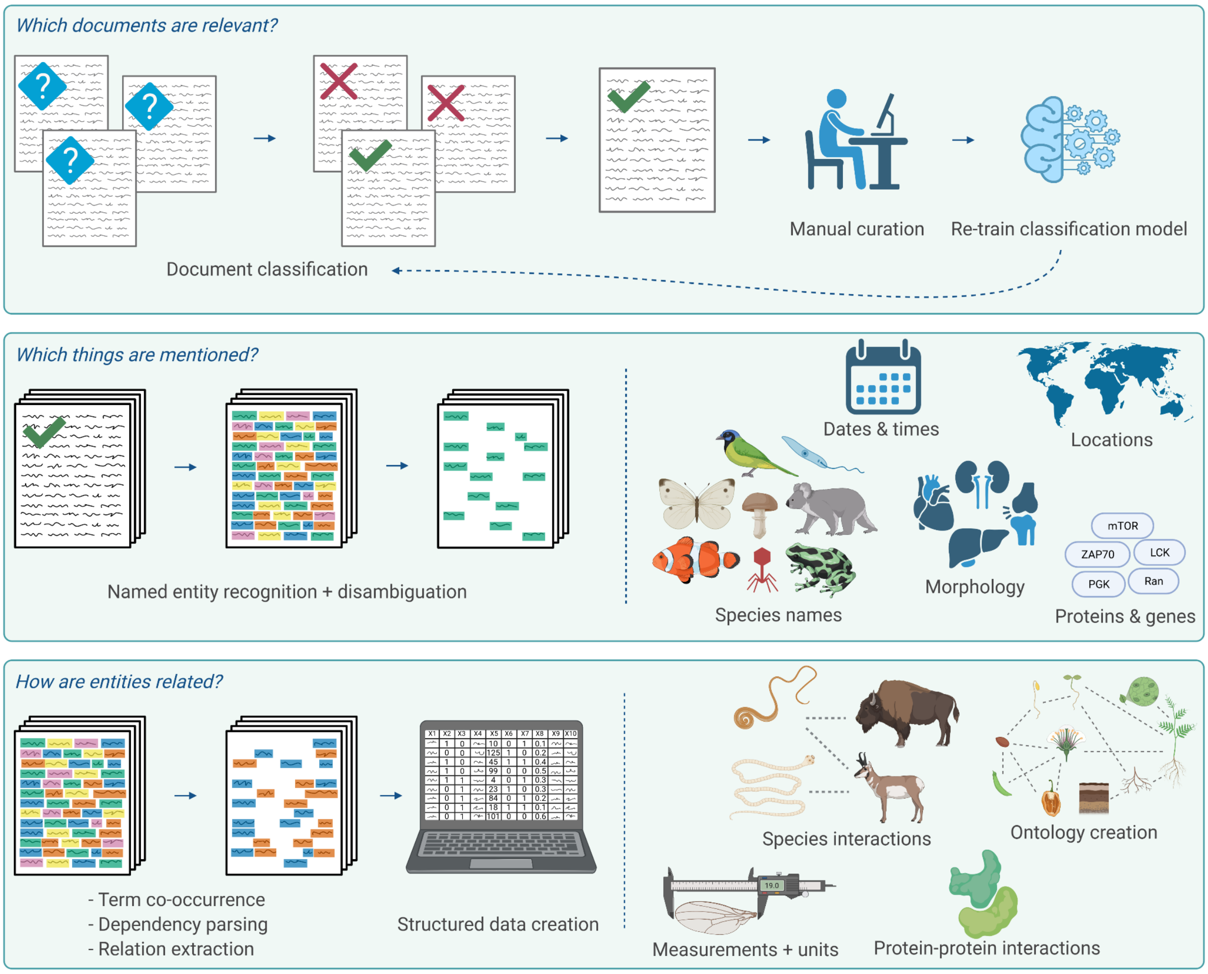

Figure 2. Potential applications of Natural Language models in ecology & evolution. The simplest application is training and applying a document classifier to predict relevant documents (top row). Given a training set of relevant and non-relevant documents (may come from existing databases, a manually-curated training set, or documents tagged by a set of rules), the relevance of new documents may be predicted and prioritized for manual screening and curation, or downstream information extraction. Manual screening may be used to validate predictive models, or re-train and fine-tune the original classifier. Once a set of relevant documents is identified, the subjects of the documents can be explored through named entity recognition (NER; middle row). Named entities can be identified by comparing text strings to a dictionary. If a complete set of entities is not known or available, a machine learning based NER tool can be used to predict entities and identify never-before-seen terms. Given a training set, NER can be used to identify terms in a text (for example, species, genes, proteins, locations, morphological structures) and tag their locations in a text. Once components of a document are tagged (parts of speech, named entities, numbers), relationships among them can be identified to create structured datasets for analysis (bottom row). Relationships may be inferred through term co-occurrence frequencies, sentence structures (dependency parsing), or through machine learning-based models that predict the nature of the relationship. Relational data can take a variety of forms including species interactions, biological measurements and their associated units, or networks of different relationship types (ontologies). Figure created with BioRender.com.

Example of relation	***Example text***
Measurements and units	“The average length of human gestation **is** 280 days”
Model-specific parameters	“R_0 was **estimated** to be 1.13”
Species interactions	“*Anoplocephala manubriata* **parasitizes** Asian elephants”
Protein-protein interactions	“Pleiotropic drug resistance 1p (Pdr1p) **regulates** Pdr5p”
Habitat associations	“*Ribes mandschuricum* is **found in** shady areas”
Species occurrences	“Cercopia moths were **collected from** sites throughout Massachusetts ”
Linnaean taxonomy / common names / synonyms	“*Boops boops*, **commonly called** the bogue, is a species of seabream native to the eastern Atlantic”
Anthropogenic impacts	“*Inversodicraea botswana* is **threatened by** sewage discharge”

Table 1. Table of common relationship types in ecology and evolution, and example texts.

References

1. Bornmann L, Mutz R. 2015 Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.* **66**, 2215–2222. (doi:https://doi.org/10.1002/asi.23329)
2. Johnson R, Watkinson A, Mabe M. 2018 The STM Report: An overview of scientific and scholarly publishing.
3. McCallen E, Knott J, Nunez‐Mir G, Taylor B, Jo I, Fei S. 2019 Trends in ecology: shifts in ecological research themes over the past four decades. *Front. Ecol. Environ.* **17**, 109–116. (doi:https://doi.org/10.1002/fee.1993)
4. Chu JSG, Evans JA. 2021 Slowed canonical progress in large fields of science. *Proc. Natl. Acad. Sci.* **118**. (doi:10.1073/pnas.2021636118)
5. Papanikolaou N, Pavlopoulos GA, Theodosiou T, Iliopoulos I. 2015 Protein-protein interaction predictions using text mining methods. *Methods San Diego Calif* **74**, 47–53. (doi:10.1016/j.ymeth.2014.10.026)
6. Wu H-Y, Chiang C-W, Li L. 2014 Text Mining for Drug–Drug Interaction. In *Biomedical Literature Mining* (eds VD Kumar, HJ Tipney), pp. 47–75. New York, NY: Springer. (doi:10.1007/978-1-4939-0709-0_4)
7. Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. 2015 Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics* **16**, 55. (doi:10.1186/s12859-015-0472-9)
8. Wei C-H, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, Wiegers TC, Lu Z. 2016 Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database* **2016**. (doi:10.1093/database/baw032)
9. Taub-Tabib H, Shlain M, Sadde S, Lahav D, Eyal M, Cohen Y, Goldberg Y. 2020 Interactive Extractive Search over Biomedical Corpora. *ArXiv200604148 Cs*
10. Poisot T, Bruneau A, Gonzalez A, Gravel D, Peres-Neto P. 2019 Ecological Data Should Not Be So Hard to Find and Reuse. *Trends Ecol. Evol.* **34**, 494–496. (doi:10.1016/j.tree.2019.04.005)
11. Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. 2021 Data integration enables global biodiversity synthesis. *Proc. Natl. Acad. Sci.* **118**. (doi:10.1073/pnas.2018093118)
12. Nicholson DN, Rubinetti V, Hu D, Thielk M, Hunter LE, Greene CS. 2021 Linguistic Analysis of the bioRxiv Preprint Landscape. *bioRxiv* , 2021.03.04.433874. (doi:10.1101/2021.03.04.433874)
13. Brierley L, Nanni F, Polka JK, Dey G, Pálfy M, Fraser N, Coates JA. 2021 Preprints in motion: tracking changes between preprint posting and journal publication during a pandemic. *bioRxiv* (doi:10.1101/2021.02.20.432090)
14. Chen Q, Allot A, Lu Z. 2021 LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.* **49**, D1534–D1540. (doi:10.1093/nar/gkaa952)
15. Hedrick BP *et al.* 2020 Digitization and the Future of Natural History Collections. *BioScience* **70**, 243–251. (doi:10.1093/biosci/biz163)
16. Rice A, Glick L, Abadi S, Einhorn M, Kopelman NM, Salman-Minkov A, Mayzel J, Chay O, Mayrose I. 2015 The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. *New Phytol.* **206**, 19–26. (doi:10.1111/nph.13191)
17. Owen D, Groom Q, Hardisty A, Leegwater T, Livermore L, Walsum M van, Wijkamp N, Spasić I. 2020 Towards a scientific workflow featuring Natural Language Processing for the digitisation of natural history collections. *Res. Ideas Outcomes* **6**, e58030. (doi:10.3897/rio.6.e58030)
18. Tuia D *et al.* 2022 Perspectives in machine learning for wildlife conservation. *Nat. Commun.* **13**, 792. (doi:10.1038/s41467-022-27980-y)

19. Lamba A, Cassey P, Segaran RR, Koh LP. 2019 Deep learning for environmental conservation. *Curr. Biol.* **29**, R977–R982. (doi:10.1016/j.cub.2019.08.016)
20. Wang Z, Lachmann A, Ma'ayan A. 2019 Mining data and metadata from the gene expression omnibus. *Biophys. Rev.* **11**, 103–110. (doi:10.1007/s12551-018-0490-8)
21. Anderson SC *et al.* 2021 Trends in ecology and conservation over eight decades. *Front. Ecol. Environ.* **19**, 274–282. (doi:10.1002/fee.2320)
22. Knott J, LaRue E, Ward S, McCallen E, Ordonez K, Wagner F, Jo I, Elliott J, Fei S. 2019 A roadmap for exploring the thematic content of ecology journals. *Ecosphere* **10**, e02801. (doi:https://doi.org/10.1002/ecs2.2801)
23. Hintzen RE, Papadopoulou M, Mounce R, Banks‐Leite C, Holt RD, Mills M, Knight AT, Leroi AM, Rosindell J. 2020 Relationship between conservation biology and ecology shown through machine reading of 32,000 articles. *Conserv. Biol.* **34**, 721–732. (doi:10.1111/cobi.13435)
24. Craven D, Winter M, Hotzel K, Gaikwad J, Eisenhauer N, Hohmuth M, König-Ries B, Wirth C. 2019 Evolution of interdisciplinarity in biodiversity science. *Ecol. Evol.* **9**, 6744–6755. (doi:10.1002/ece3.5244)
25. Dayeen FR, Sharma AS, Derrible S. 2020 A text mining analysis of the climate change literature in industrial ecology. *J. Ind. Ecol.* **24**, 276–284. (doi:https://doi.org/10.1111/jiec.12998)
26. Luiz OJ, Olden JD, Kennard MJ, Crook DA, Douglas MM, Saunders TM, King AJ. 2019 Trait-based ecology of fishes: A quantitative assessment of literature trends and knowledge gaps using topic modelling. *Fish Fish.* **20**, 1100–1110. (doi:10.1111/faf.12399)
27. Han BA, Ostfeld RS. 2019 Topic modeling of major research themes in disease ecology of mammals. *J. Mammal.* **100**, 1008–1018. (doi:10.1093/jmammal/gyy174)
28. Westgate MJ, Barton PS, Pierson JC, Lindenmayer DB. 2015 Text analysis tools for identification of emerging topics and research gaps in conservation science. *Conserv. Biol.* **29**, 1606–1614. (doi:10.1111/cobi.12605)
29. Millard JW, Freeman R, Newbold T. 2020 Text-analysis reveals taxonomic and geographic disparities in animal pollination literature. *Ecography* **43**, 44–59. (doi:https://doi.org/10.1111/ecog.04532)
30. Kang A, Ren L, Hua C, Song H, Dong M, Fang Z, Zhu M. 2021 Environmental management strategy in response to COVID-19 in China: Based on text mining of government open information. *Sci. Total Environ.* **769**, 145158. (doi:10.1016/j.scitotenv.2021.145158)
31. Ladle RJ, Correia RA, Do Y, Joo G-J, Malhado AC, Proulx R, Roberge J-M, Jepson P. 2016 Conservation culturomics. *Front. Ecol. Environ.* **14**, 269–275. (doi:10.1002/fee.1260)
32. Fink C, Hausmann A, Di Minin E. 2020 Online sentiment towards iconic species. *Biol. Conserv.* **241**, 108289. (doi:10.1016/j.biocon.2019.108289)
33. Wartmann FM, Koblet O, Purves RS. 2021 Assessing experienced tranquillity through natural language processing and landscape ecology measures. *Landsc. Ecol.* (doi:10.1007/s10980-020-01181-8)
34. Correia RA *et al.* 2021 Digital data sources and methods for conservation culturomics. *Conserv. Biol.* **35**, 398–411. (doi:10.1111/cobi.13706)
35. Westgate MJ, Haddaway NR, Cheng SH, McIntosh EJ, Marshall C, Lindenmayer DB. 2018 Software support for environmental evidence synthesis. *Nat. Ecol. Evol.* **2**, 588–590. (doi:10.1038/s41559-018-0502-x)
36. Nakagawa S, Samarasinghe G, Haddaway NR, Westgate MJ, O'Dea RE, Noble DWA, Lagisz M. 2019 Research Weaving: Visualizing the Future of Research Synthesis. *Trends Ecol. Evol.* **34**, 224–238. (doi:10.1016/j.tree.2018.11.007)
37. Grames EM, Stillman AN, Tingley MW, Elphick CS. 2019 An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods Ecol. Evol.* **10**, 1645–1654. (doi:10.1111/2041-210X.13268)
38. Pham B *et al.* 2021 Text mining to support abstract screening for knowledge syntheses:

a semi-automated workflow. *Syst. Rev.* **10**, 156. (doi:10.1186/s13643-021-01700-x)

39. van de Schoot R *et al.* 2021 An open source machine learning framework for efficient and transparent systematic reviews. *Nat. Mach. Intell.* **3**, 125–133. (doi:10.1038/s42256-020-00287-7)
40. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. 2015 Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst. Rev.* **4**, 5. (doi:10.1186/2046-4053-4-5)
41. Olorisade BK, de Quincey E, Brereton P, Andras P. 2016 A critical analysis of studies that address the use of text mining for citation screening in systematic reviews. In *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*, pp. 1–11. New York, NY, USA: Association for Computing Machinery. (doi:10.1145/2915970.2915982)
42. Cornford R, Deinet S, Palma AD, Hill SLL, McRae L, Pettit B, Marconi V, Purvis A, Freeman R. 2021 Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets. *Glob. Ecol. Biogeogr.* **30**, 339–347. (doi:https://doi.org/10.1111/geb.13219)
43. Hudson LN *et al.* 2014 The PREDICTS database: a global database of how local terrestrial biodiversity responds to human impacts. *Ecol. Evol.* **4**, 4701–4735. (doi:https://doi.org/10.1002/ece3.1303)
44. Roll U, Correia RA, Berger‐Tal O. 2018 Using machine learning to disentangle homonyms in large text corpora. *Conserv. Biol.* **32**, 716–724. (doi:https://doi.org/10.1111/cobi.13044)
45. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. 2021 Deep Learning Based Text Classification: A Comprehensive Review. *ArXiv200403705 Cs Stat*
46. Nunez‐Mir GC, Iannone BV, Pijanowski BC, Kong N, Fei S. 2016 Automated content analysis: addressing the big literature challenge in ecology and evolution. *Methods Ecol. Evol.* **7**, 1262–1272. (doi:10.1111/2041-210X.12602)
47. Thessen AE, Bogdan P, Patterson DJ, Casey TM, Hinojo-Hinojo C, Lange O de, Haendel MA. 2021 From Reductionism to Reintegration: Solving society's most pressing problems requires building bridges between data types across the life sciences. *PLOS Biol.* **19**, e3001129. (doi:10.1371/journal.pbio.3001129)
48. Muñoz G, Kissling WD, van Loon EE. 2019 Biodiversity Observations Miner: A web application to unlock primary biodiversity data from published literature. *Biodivers. Data J.* **7**, e28737. (doi:10.3897/BDJ.7.e28737)
49. Kaur KM, Malé P-JG, Spence E, Gomez C, Frederickson ME. 2019 Using text-mined trait data to test for cooperate-and-radiate co-evolution between ants and plants. *PLOS Comput. Biol.* **15**, e1007323. (doi:10.1371/journal.pcbi.1007323)
50. Thessen AE, Parr CS. 2014 Knowledge Extraction and Semantic Annotation of Text from the Encyclopedia of Life. *PLoS ONE* **9**, e89550. (doi:10.1371/journal.pone.0089550)
51. McIntyre KM, Setzkorn C, Wardeh M, Hepworth PJ, Radford AD, Baylis M. 2014 Using open-access taxonomic and spatial information to create a comprehensive database for the study of Mammalian and avian livestock and pet infections. *Prev. Vet. Med.* **116**, 325–335. (doi:10.1016/j.prevetmed.2013.07.002)
52. Wardeh M, Risley C, McIntyre MK, Setzkorn C, Baylis M. 2015 Database of host-pathogen and related species interactions, and their global distribution. *Sci. Data* **2**, 150049. (doi:10.1038/sdata.2015.49)
53. Hart AG, Carpenter WS, Hlustik-Smith E, Reed M, Goodenough AE. 2018 Testing the potential of Twitter mining methods for data acquisition: Evaluating novel opportunities for ecological research in multiple taxa. *Methods Ecol. Evol.* **9**, 2194–2205. (doi:10.1111/2041-210X.13063)
54. Kulkarni R, Di Minin E. 2021 Automated retrieval of information on threatened species from online sources using machine learning. *Methods Ecol. Evol.* **12**, 1226–1239. (doi:10.1111/2041-210X.13608)

55. Schalley AC. 2019 Ontologies and ontological methods in linguistics. *Lang. Linguist. Compass* **13**, e12356. (doi:10.1111/lnc3.12356)
56. Bodenreider O. 2008 Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb. Med. Inform.* , 67–79.
57. Alexopoulou D, Andreopoulos B, Dietze H, Doms A, Gandon F, Hakenberg J, Khelif K, Schroeder M, Wächter T. 2009 Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. *BMC Bioinformatics* **10**, 28. (doi:10.1186/1471-2105-10-28)
58. Smith CL, Eppig JT. 2009 The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1**, 390–399. (doi:10.1002/wsbm.44)
59. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. 2012 Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5. (doi:10.1186/gb-2012-13-1-r5)
60. Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE, the ENVO Consortium. 2013 The environment ontology: contextualising biological and biomedical entities. *J. Biomed. Semant.* **4**, 43. (doi:10.1186/2041-1480-4-43)
61. Frey U, Cox M. 2015 Building a diagnostic ontology of social-ecological systems. *Int. J. Commons* **9**, 595–618. (doi:10.18352/ijc.505)
62. Batista-Navarro R, Laporte M-A, Regan M, Ulate W, Weiland C. 2018 Extending the Environment Ontology with Text-mined Habitat Mentions. *ICEI 2018 10th Int. Conf. Ecol. Inform.- Transl. Ecol. Data Knowl. Decis. Rapidly Chang. World*. (doi:10.22032/dbt.37907)
63. Mendelsohn E *et al.* 2021 A global repository of novel antimicrobial emergence events. *F1000Research* **9**, 1320. (doi:10.12688/f1000research.26870.2)
64. Alizon S, Hurford A, Mideo N, Van Baalen M. 2009 Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future: Virulence evolution and trade-off hypothesis. *J. Evol. Biol.* **22**, 245–259. (doi:10.1111/j.1420-9101.2008.01658.x)
65. Hodges KE. 2008 Defining the problem: terminology and progress in ecology. *Front. Ecol. Environ.* **6**, 35–42. (doi:10.1890/060108)
66. Trombley CA, Cottenie K. 2019 Quantifying the Scientific Cost of Ambiguous Terminology in Community Ecology. *Philos. Top.* **47**, 203–218.
67. Nadrowski K *et al.* 2013 Harmonizing, annotating and sharing data in biodiversity–ecosystem functioning research. *Methods Ecol. Evol.* **4**, 201–205. (doi:10.1111/2041-210x.12009)
68. Madin J, Bowers S, Schildhauer M, Krivov S, Pennington D, Villa F. 2007 An ontology for describing and synthesizing ecological observation data. *Ecol. Inform.* **2**, 279–296. (doi:10.1016/j.ecoinf.2007.05.004)
69. Walls RL *et al.* 2014 Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLoS ONE* **9**, e89606. (doi:10.1371/journal.pone.0089606)
70. Dahdul WM *et al.* 2010 Evolutionary Characters, Phenotypes and Ontologies: Curating Data from the Systematic Biology Literature. *PLOS ONE* **5**, e10708. (doi:10.1371/journal.pone.0010708)
71. Dececchi TA, Balhoff JP, Lapp H, Mabee PM. 2015 Toward Synthesizing Our Knowledge of Morphology: Using Ontologies and Machine Reasoning to Extract Presence/Absence Evolutionary Phenotypes across Studies. *Syst. Biol.* **64**, 936–952. (doi:10.1093/sysbio/syv031)
72. Penev L, Dimitrova M, Senderov V, Zhelezov G, Georgiev T, Stoev P, Simov K. 2019 OpenBiodiv: A Knowledge Graph for Literature-Extracted Linked Open Data in Biodiversity Science. *Publications* **7**, 38. (doi:10.3390/publications7020038)
73. Schneider FD *et al.* 2019 Towards an ecological trait-data standard. *Methods Ecol. Evol.* **10**, 2006–2019. (doi:10.1111/2041-210X.13288)
74. Deans AR *et al.* 2015 Finding Our Way through Phenotypes. *PLOS Biol.* **13**, e1002033.

(doi:10.1371/journal.pbio.1002033)

75. Endara L, Burleigh G, Cooper L, Jaiswal P, Laporte M-A. 2018 A Natural Language Processing Pipeline to extract phenotypic data from formal taxonomic descriptions with a focus on flagellate plants.
76. Vattakaven T, Rajagopal P, Dhandapani B, Grard P, Le Bourgeois T. 2018 Traits: Structuring Species Information for Discoverability, Navigation and Identification. In *Multimedia Tools and Applications for Environmental & Biodiversity Informatics* (eds A Joly, S Vrochidis, K Karatzas, A Karppinen, P Bonnet), pp. 93–112. Cham: Springer International Publishing. (doi:10.1007/978-3-319-76445-0_6)
77. Gallagher RV *et al.* 2020 Open Science principles for accelerating trait-based science across the Tree of Life. *Nat. Ecol. Evol.* **4**, 294–303. (doi:10.1038/s41559-020-1109-6)
78. Foo YZ, O'Dea RE, Koricheva J, Nakagawa S, Lagisz M. 2021 A practical guide to question formation, systematic searching and study screening for literature reviews in ecology and evolution. *Methods Ecol. Evol.* **n/a**. (doi:10.1111/2041-210X.13654)
79. Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. 2012 Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.* **3**, 327–338. (doi:10.1111/j.2041-210X.2011.00172.x)
80. Neumann M, King D, Beltagy I, Ammar W. 2019 ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proc. 18th BioNLP Workshop Shar. Task* , 319–327. (doi:10.18653/v1/W19-5034)
81. Li J, Sun A, Han J, Li C. 2020 A Survey on Deep Learning for Named Entity Recognition. *ArXiv181209449 Cs*
82. In press. LILA BC (Labeled Image Library of Alexandria: Biology and Conservation). *LILA BC*. See https://lila.science/ (accessed on 5 April 2022).
83. Bodenreider O. 2004 The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270. (doi:10.1093/nar/gkh061)
84. Rogers FB. 1963 Medical subject headings. *Bull. Med. Libr. Assoc.* **51**, 114–116.
85. Thessen AE, Cui H, Mozzherin D. 2012 Applications of Natural Language Processing in Biodiversity Science. *Adv. Bioinforma.* **2012**, 391574. (doi:10.1155/2012/391574)
86. Thessen AE *et al.* 2015 Emerging semantics to link phenotype and environment. *PeerJ* **3**, e1470. (doi:10.7717/peerj.1470)
87. Karam N, Khiat A, Algergawy A, Sattler M, Weiland C, Schmidt M. 2020 Matching biodiversity and ecology ontologies: challenges and evaluation results. *Knowl. Eng. Rev.* **35**. (doi:10.1017/S0269888920000132)
88. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D. 2012 Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLOS ONE* **7**, e29715. (doi:10.1371/journal.pone.0029715)
89. Nair S. 2017 A Biomedical Information Extraction Primer for NLP Researchers. *ArXiv170505437 Cs*
90. Dallas T, Gehman A-L, Farrell MJ. 2018 Variable Bibliographic Database Access Could Limit Reproducibility. *BioScience* **68**, 552–553. (doi:10.1093/biosci/biy074)
91. Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M. 2020 The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293. Online: Association for Computational Linguistics. (doi:10.18653/v1/2020.acl-main.560)
92. Le Guillarme N, Thuiller W. 2022 TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods Ecol. Evol.* **13**, 625–641. (doi:10.1111/2041-210X.13778)
93. Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S. 2018 A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLOS Comput. Biol.* **14**, e1005962. (doi:10.1371/journal.pcbi.1005962)
94. Admin. In press. The State of Journal Production and Access 2020: Report on survey of society and university publishers | STM Publishing News.
95. In press. PMC Open Access Subset - PMC. See

https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ (accessed on 29 March 2022).

96. Public Resource. 2021 The General Index.
97. Bloudoff-Indelicato M. 2015 Text-mining block prompts online response. *Nature* **527**, 413–413. (doi:10.1038/527413f)
98. Mounce R. 2015 Dark Research: information content in many modern research papers is not easily discoverable online. (doi:10.7287/peerj.preprints.773v1)
99. Angulo E *et al.* 2021 Non-English languages enrich scientific knowledge: The example of economic costs of biological invasions. *Sci. Total Environ.* , 144441. (doi:10.1016/j.scitotenv.2020.144441)
100. Blasi D, Anastasopoulos A, Neubig G. 2021 Systematic Inequalities in Language Technology Performance across the World's Languages. *ArXiv211006733 Cs*
101. Beltagy I, Lo K, Cohan A. 2019 *SciBERT: A Pretrained Language Model for Scientific Text*.
102. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. 2020 BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240.
103. Orife I *et al.* 2020 Masakhane -- Machine Translation For Africa. *ArXiv200311529 Cs*